High-quality genome assembly of *Metaphire vulgaris*

Feng Jin[1], Zhaoli Zhou[2], Qi Guo[1], Zhenwen Liang[1], Ruoyu Yang[1], Jibao Jiang[3], Yanlin He[1], Qi Zhao[3] and Qiang Zhao[2]

[1] College of Rehabilitation Sciences, Shanghai University of Medicine and Health Sciences, Shanghai, China
[2] Shanghai Key Laboratory of Molecular Imaging, Shanghai University of Medicine and Health Sciences, Shanghai, China
[3] Department of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Earthworms enrich the soil and protect the health of their ecological environment. Previous studies on these invertebrates determined their protein content, hormone secretions, medicinal value, and ecological habits, but their whole genomic sequence remains incomplete. We performed whole genome sequencing of *Metaphire vulgaris* (Chen, 1930), which belongs to the genus *Metaphire* of the family Megascolecidae. The genome assembly was 729 Mb, with a N50 contig size of 4.2 Mb. In total, 559 contigs were anchored to 41 chromosomes according to the results of Hi-C (High-throughput Chromosome Conformation Capture) technology, which was confirmed by karyological analysis. A comparison of the genomic sequences and genes indicated that there was a whole-genome duplication in *M. vulgaris* followed by several chromosome fusion events. Hox genes and lumbrokinase genes were identified as partial clusters surrounding the genome. Our high-quality genome assembly of *M. vulgaris* will provide valuable information for gene function and evolutionary studies in earthworms.

Subjects Evolutionary Studies, Genomics
Keywords Genome assembly, *Metaphire vulgaris*, Whole genome sequencing, Hi-C, Genome duplication, Hox, Lumbrokinase

Submitted 22 June 2020
Accepted 15 October 2020
Published 12 November 2020

Corresponding authors
Qi Zhao, zhaoq@sjtu.edu.cn
Qiang Zhao, zqiang_99@yahoo.com

Academic editor
Timothy Driscoll

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.10313**

INTRODUCTION

Earthworms are terrestrial invertebrates belonging to Oligochaeta in the phylum Annelida. There are more than 3,000 species of earthworm in the world, with more than 600 species found in China alone (*Csuzdi, 2012*; *Jiang & Qiu, 2018*). Earthworms burrow in the soil, decompose organic matter, and create ideal conditions for the growth and reproduction of soil microorganisms. They are particularly important for soil enrichment and protecting the health of their surrounding ecological environment. Many countries use earthworms to process domestic and organic wastes, and purify sewage. Earthworms are also used in traditional Chinese medicine (TCM) to treat a variety of diseases, and can be used as a high-protein feed.

A number of earthworm studies have focused on their protein content, hormone secretions, medicinal value, and ecological habits but only a few studies have investigated earthworm genomics. *Zwarycz et al. (2015)* performed whole genome sequencing on *Eisenia*

fetida (Savigny, 1826), an earthworm from the family Lumbricidae. And the annotation study were conducted by *Paul et al. (2018)*, although the quality of the assembled genome sequences was not of a sufficient quality for downstream analyses (N50 = 1.85 Kb). *Bhambri et al. (2018)* sequenced a different strain of *E. fetida* and conducted genome-wide analyses and transcriptome studies. This study also failed to produce a genome assembly sufficient for accurate gene annotation (N50 = 9.31 Kb).

Metaphire vulgaris (Chen, 1930) is called ''Hu dilong'' in Traditional Chinese medicine (*Chinese Pharmacopoeia Commission, 2015*) and belongs to the genus *Metaphire* of the family Megascolecidae. It is commonly found in many Chinese provinces, including in Jiangsu, Shanghai, Zhejiang, and Guizhou. It is 120–215 mm long and 5–8 mm wide with body segments 90–124 mm in length. We conducted whole genome sequencing of *Metaphire vulgaris* by combining the single-molecule long sequences and the second-generation high-throughput short sequences to produce a 729 Mb high-quality sequence assembly of this species. We built a chromosome-level assembly with 41 complete chromosomes using Hi-C technology. We also performed precise genome annotations, comparative genome analysis, and phylogenetic studies of Oligochaetes and other related species in Annelida using our high-quality genome assembly and the transcriptome data from multiple tissue samples.

MATERIALS AND METHODS

Genome sequencing and assembly

One clitellata of *Metaphire vulgaris* (Chen, 1930) grown in Jiangsu, China was prepared for genome and transcriptome sequencing. It was about three month in age, with the length and width of 17 cm and 0.7 cm, respectively Genomic DNA was extracted from the head muscle of the earthworm using the QIAGEN® Genomic DNA Extraction Kit (Cat#13323, Qiagen) according to the manufacturer's instructions.

We sequenced the whole genome of *M. vulgaris* using PromethION single molecule platform (Oxford Nanopore) and Illumina NovaSeq sequencing platform. A DNA library was constructed following the standard Oxford Nanopore protocol and was sequenced on the PromethION platform (Oxford Nanopore Technologies, ONT, UK). A Paired-End (PE) library was simultaneously constructed according to the manufacturer's instructions for genomic DNA sequencing (Illumina, San Diego, CA, USA). The insert size was approximately 400bp and was sequenced on an Illumina NovaSeq system (read length 150 bp).

Primers (5′-GGTCAACAAATCATAAAGATATTGG-3′ and 5′-TAAACTTCAGGGTG ACCAAAAAATCA-3′) were used to amplify the mitochondrial cytochrome c oxidase subunit I (COX1) gene from the extracted DNA. The PCR products were sequenced on the ABI 3730xl DNA Analyzer. And the sequences were compared by BLASTN (*Altschul et al., 1990*) with default settings.

We adopted a combined strategy of filtering, assembling, and polishing with multiple software pipelines to obtain a high-quality genome assembly of *M. vulgaris*. Canu version 1.8 (*Koren et al., 2017*) was used with default parameters to filter and correct the raw

reads from the Nanopore high-noise single-molecule sequencing. SMARTdenovo (*Istace et al., 2017*) was used to assemble the contigs of the *M. vulgaris* genome (-c 1 -k 21). The ONT reads were re-aligned to the assembled contigs using racon v1.0.0 (default settings; https://github.com/lbcb-sci/racon) and minimap2 (v2.1, -x map-ont; *Li, 2018*) to reduce the assembly errors. The ONT contigs were polished three times by the PE reads, which were produced from whole-genome shotgun dataset using BWA (v0.7.17-r1188, default settings; *Li & Durbin, 2009*) and Pilon (v1.22, –changes –vcf –diploid –mindepth 10; *Walker et al., 2014*) to remove minor errors (SNP and indels). And BUSCO (Benchmarking Universal Single Copy Orthologs; *Simão et al., 2015*; v3.0.2, -m geno -l metazoa_odb9) was conducted to evaluate the assembled contigs by searching for 978 metazoa-conserved genes.

The Hi-C library was constructed following the method of *Wang et al. (2015)*, with the DNA extracted from head muscle tissue of *M. vulgaris*. The 150 bp paired-end reads were sequenced using the Illumina NovaSeq system. Contigs were clustered and sorted into chromosomes by LACHESIS based on the Hi-C data (*Burton et al., 2013*; CLUSTER MIN RE SITES = 100; CLUSTER MAX LINK DENSITY = 2.5; CLUSTER NONINFORMATIVE RATIO = 1.4; ORDER MIN N RES IN TRUNK = 60; ORDER MIN N RES IN SHREDS = 60). Finally, placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were manually adjusted. Briefly, a heat map was drawn by the bin matrix file from LACHESIS's result. Then we retrieved the scatter points of the collinear signal that obviously did not conform to the correct positional relationship of the chromosomes, and revised the bin matrix file by adjusting the contig order or cutting off the contig to generate a new interactive heat map which was consistent with the chromosome spatial position signal. Purge Haplotigs (v1.1.1; Roach et al., 2018) was used to evaluate the genome assembly (default parameters), along with minimap2 (-ax map-ont) and SAMtools (v1.9, view -hF 256; *Li et al., 2009*). The Illumina paired-end data were mapped to assembled contigs using Bowtie2 (v2.2.6, -I 50 -X 1000; *Langmead & Salzberg, 2012*), and pile upped with SAMtools (mpileup -f).

Karyological analysis and genome size estimation

One individual of *Metaphire vulgaris* (17 cm/0.7 cm in length/width, 3 months old) was prepared for the karyological analysis. 0.2 ml colchicine (1mg/ml) was injected into mature subjects of *M. vulgaris*. The testes and sperm sacs of the earthworm were isolated and triturated 24 h after the injection. The sample was dyed using 0.1ug/ml DAPI for 5–10 min and was observed under a fluorescence microscope (ZEISS Axio Imager2). Ikaros software (https://metasystems-international.com/cn/products/ikaros/) was used to analyze the karyotype.

The k-mer analysis software Kmerfreq_AR (SOAPec_v2.01 package https://sourceforge.net/projects/soapdenovo2/files/ErrorCorrection/SOAPec_v2.01.tar.gz/download) was adopted to estimate the genome size of *M. vulgaris* at k-mer 17.

RNA preparation and sequencing

From the same individual used in genome sequencing, total RNA of six tissues (heart, ventral nerve cord, gonad, epidermis, intestine, and tail) was prepared. RNA degradation

and contamination was monitored on 1% agarose gels. RNA concentration was measured using Qubit® RNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies,CA, USA). RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). A total amount of 1 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using TruSeq RNA Library Preparation Kit (Illumina, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. The library preparations were sequenced on an Illumina Novaseq platform and 150 bp paired-end reads were generated. Fastp (version 0.12.6; *Chen et al., 2018*) with default parameters was applied to filter out low quality reads.

Genome annotation

RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html, version 1.0.5) was used to build the custom repeat library from the genome assembly sequence of *M. vulgaris*. The homologous repeat elements in the genome of *M. vulgaris* were identified and classified using RepeatMasker (http://www.repeatmasker.org/, version 3.3.0).

In order to build the preliminary gene models on the repeat-masked genome sequence. Augustus (*Stanke et al., 2006*; v3.2.1, –species=caenorhabditis) and SNAP (*Korf, 2004*; version 2006-07-28, C.elegans.hmm) were applied to predict *de novo* genes using gene model parameters trained by *Caenorhabditis elegans*. With the splice junctions identified by STAR (*Dobin et al., 2013*; STAR-2.6.1c, –outSAMtype BAM SortedByCoordinate –outFilterType BySJout –outFilterMultimapNmax 20 –alignSJoverhangMin 8 –alignSJDBoverhangMin 1 –outFilterMismatchNmax 999 –outFilterMismatchNoverLmax 0.04 –alignIntronMin 20 –alignIntronMax 20000 –alignMatesGapMax 20000 –chimSegmentMin 20), GeneMark-ET (*Lomsadze, Burns & Borodovsky, 2014*; v4.46, gmes_petap.pl –ET) was also used to perform unsupervised training with RNA-Seq data from six tissues (heart, ventral nerve cord, gonad, epidermis, intestine, and tail) and subsequently generates *ab initio* gene predictions. In the mean time, the filtered RNA-seq reads generated from six tissues were assembled through two approach: (1) RNA-seq reads were aligned to the genome assembly by HISAT2 (*Kim et al., 2019*; v2.0.5, –dta), and then imported to the genome-guided assembler StringTie (*Pertea et al., 2015*; v1.3.0, default settings); (2) RNA-seq reads were imported to the *de novo* assembler Trinity (*Grabherr et al., 2011*; v2.1.1, –normalize_reads –SS_lib_type FR). The two sets of assembled transcripts were reassembled based on the overlapping alignments by PASA (*Campbell et al., 2006*; v2.0.1, –ALIGNERS gmap -I 25000 -C -R). And the protein sequences of three close related species (*Capitella teleta*, *Helobdella robusta*, *Caenorhabditis elegans*, from EnsemblMetazoa (http://metazoa.ensembl.org/index.html) were also aligned to the genome assembly by Exonerate (*Slater & Birney, 2005*; v2.2.0, -m protein2genome –percent 50 –querytype protein –targettype dna) to find homologus genes. Finally, the predicted gene structures were integrated into consensus gene structures using EVidenceModeler (EVM; *Haas et al., 2008*; v1.1.1, –segmentSize 5000000 –overlapSize 100000, –weights PROTEIN 5; TRANSCRIPT gmap 5/assembler 10; ABINITIO_PREDICTION 5). Genes

with expression evidences or protein homologues were regarded as high quality (HQ) genes.

The functional classification of Gene Ontology (GO; http://www.geneontology.org/) of the genes was performed by the InterProScan program (*Zdobnov & Apweiler, 2001*). RNA-seq reads generated from the six tissues were mapped to the coding sequence of genes by hisat2 v2-2.0.5 (*Kim et al., 2019*) using default parameters. Normalized read counts based on the gene annotation were calculated using R package DESeq2 (*Love, Huber & Anders, 2014*).

Comparative genomics analysis

Protein-coding genes and CDS of *Caenorhabditis elegans*, *Lottia gigantea*, *Capitella teleta* and *Helobdella robusta* were downloaded from the EnsemblMetazoa database (https://metazoa.ensembl.org/info/data/ftp/index.html, release-45). Protein-coding genes and CDS of *Drosophila melanogaster*, *Danio rerio*, *Xenopus tropicalis*, *Gallus gallus*, *Homo sapiens*, and *Mus musculus* were downloaded from Ensembl database (http://www.ensembl.org/info/data/index.html, release-98). Only the longest transcript was selected for the genes with alternative splice variants. OrthoFinder (*Emms & Kelly, 2015*) was used to identify orthologs in these 12 species. The species evolution tree was constructed using Bayesian method based on single-copy gene families identified by OrthoFinder. The supergene sequences were subjected to phylogenetic analyses by mrbayes-3.2.7 (*Ronquist et al., 2012*) software with the parameter (mcmc ngen = 100000, samplefreq = 10), and *A. japonicus* was set as the outgroup. Then the first 25% samples from the cold chain (relburnin=yes and burninfrac = 0.25) were discarded. The Ks-based (Ks: synonymous substitution rate) ortholog age distributions were determined based on one-to-one orthologs between *M. vulgaris* and the other three species of Lophotrochozoa using default settings. The Ks estimation was calculated using KaKs_Calculator1.2 (*Wang et al., 2009*).

Intraspecific synteny analysis was performed for all gene models of *M. vulgaris* using MCscan (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)) with default parameters. Ks of the paralog gene pairs in synteny blocks was also estimated by KaKs_Calculator1.2. Circos software (*Krzywinski et al., 2009*) was used for synteny visualization within the genome of *M. vulgaris*.

Hox genes and lumbrokinase genes

Hox genes were identified using the homology search. The homeodomain sequences downloaded from the homeobox database (http://homeodb.zoo.ox.ac.uk/) were aligned using clustalw2 (*Larkin et al., 2007*). These sequences were used to construct a homeobox HMM using hmmbuild from the HMMER v3.1b suite (*Eddy, 2011*). Predicted proteins in the *M. vulgaris* genome were scanned using hmmsearch from HMMER v3.1b suite. The candidate genes belonging to the HOXL subclass of ANTP class were further filtered based on manual curation and molecular phylogeny. The subclass of these Hox genes was determined using the homeodomain regions of filtered Hox genes in *M. vulgaris*. The homeobox genes from spiralian genomes (*Simakov et al., 2013*) were used to construct

the phylogenetic tree using the Neighbor-Joining method in MEGA7 (*Kumar, Stecher & Tamura, 2016*). Heat maps of Hox gene expression were drawn with the pheatmap package in R (https://cran.r-project.org/web/packages/pheatmap/index.html).

Amino acid sequences of 22 lumbrokinase gene families obtained from NCBI (AAL28118, AAN28692, AAN78282, AAP04532, AAP92795, AAR13225, AAR13226, AAR13227, AAT74899, AAT74900, AAW27919, ABA43718, ABQ23217, ABW04903, ABW04904, ABW04905, ABW04906, AIC77168, AKQ13274, ARD24433, ATP16189, QBA57435) were aligned to the gene models of *M. vulgaris* by BLASTP with e-value 1×10^{-10}. Aligned hits with greater than 50% identity and 50% coverage were considered homologs of lumbrokinase, and the results were summarized manually.

RESULTS

Genome sequencing and de novo assembly

The DNA fragment identified by PCR amplification and sequencing of the mitochondrial DNA (*Zhang et al., 2016*) was found to be nearly identical (99.7%) to the published *M. vulgaris* mitochondrial sequence, confirming the identity of *M. vulgaris*. Karyological analysis allowed us to verify as many as 41 pairs of chromosomes for *M. vulgaris*. One chromosome in particular was shown to be much larger than the other chromosomes (Fig. S1).

Approximately 37 gigabase (Gb) single molecule reads (over 50-fold genome coverage) and 45 Gb short reads were generated and used to assemble the whole genome of *M. vulgaris*. The estimated genome size of *M. vulgaris* was determined to be about 0.65 Gb based on analysis of 17-mer sequences on short reads. The *k*-mer distribution showed that the genome is highly heterozygous (Fig. S2). A high-quality genome assembly of *M. vulgaris* was obtained using the combined strategy of filtering, assembling, and polishing with multiple software pipelines. A total of 559 contigs were generated with a total length of 729 Mb and an N50 size of 4.2 Mb (Table 1). The largest contig reached the length of 16.6 Mb. Chromosome conformation capture (Hi-C) and short read sequencing were used to assemble the contigs into chromosomes, resulting in 104 Gb raw data (∼150-fold).

The assembled *M. vulgaris* contigs were clustered and sorted according to the Hi-C data, with the assembly errors corrected. We obtained 41 major groups corresponding to 41 chromosomes (MV001∼MV041, Fig. S3), representing 95.71% of all the contigs. The largest chromosome was as long as 50.2 Mb, with the others ranging from 11.0 to 30.9 Mb (Table S1), which was consistent with the karyological analysis. Purge Haplotigs showed that the estimated haplotigs and artefacts were only 24,929,850 bp (3.42% of the whole assembly 728,570,957 bp) and 442,973 bp (0.06%), respectively. And the haplotigs and artefacts were all belonged to the small and unlocated scaffolds other than the 41 chromosomes. Read coverage statistics indicated that approximately 93.5% of raw reads could be aligned with the final assembly, covering 99.6% of the genome. The BUSCO results indicated that 94.3% (922) of genes were completely captured (81.2%/794 were complete and single-copy BUSCOs, 13.1%/128 were complete and duplicated BUSCOs), 1.2% (12) were fragmented, and only 4.5% (44) were missing from the assembly.

Table 1 ***Metaphire vulgaris*** **genome statistics and gene predictions.**

Assembly feature	
Assembled sequences	728,570,957 bp
N50 contig length	4,202,844 bp
N90 contig length	649,861 bp
Longest contig	16,632,089 bp
Number of contigs	559
N50 scaffold length	16,345,198 bp
N90 scaffold length	12,220,767 bp
Longest scaffold	50,191,389 bp
Number of scaffold	280
Repeat sequences	
SINEs	2,444,842 bp (0.34%)
LINEs	73,457,768 bp (10.8%)
LTR elements	8,821,633 bp (1.21%)
DNA elements	59,603,311 bp (8.18%)
Total repeats	334,237,560 bp (45.88%)
Gene annotation	
Gene models (high confidence)	43,842
Gene models (low confidence)	6,397
Average gene length	6,154.5 bp
Average CDS length	1,185.3 bp

Notes.
[a]The average gene and CDS length were estimated based on the gene models with high confidence (HQ gene models).

Genome annotation

The *M. vulgaris*-specific repetitive sequences were *de novo* identified, which was followed by the whole genome screening. The overall repeat content of *M. vulgaris* was approximately 45.88%, of which approximately 22.0% and 17.8% were LINEs and DNA elements, respectively. The proportion of repetitive elements in *M. vulgaris* was much greater than that was indicated in a previous study of *Hirudinaria manillensis* (19.52%) and *Helobdella robusta* (17.33%) (*Guan et al., 2020*), the two species of leech also belonging to the Annelida. The contrast in repetitive element content may be the reason for the larger genome size of *M. vulgaris*.

A combined strategy of ab initio predictions, RNA-seq supported evidence, and homologous searching was used to obtain a reliable protein-coding gene set. RNA-seq reads with an average of 8.6 Gb raw data per tissue were generated from six tissues (heart, ventral nerve cord, gonad, epidermis, intestine, and tail) and used to generate a set of high quality (HQ) gene models. 50,239 gene models were created from the EvidenceModeler (EVM) pipeline; 43,842 of them were HQ genes with expression evidences or protein homologues, which were also assessed by BUSCO (Table S2). The average gene length was 6,154.5 bp among the HQ genes, the average coding sequence (CDS) length was 1,185.3 bp, and the average exons per gene was 5.7. The gene length of *M. vulgaris* was much larger than the Asian buffalo leech but the CDS lengths were similar (*Guan et al., 2020*)

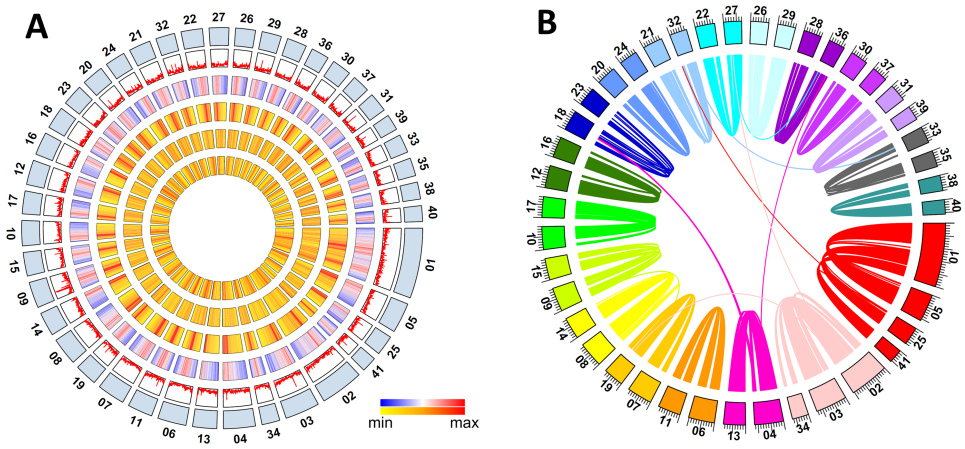

Figure 1 ***Metaphire vulgaris* genome features.** (A) Characteristics of the 41 chromosomes of *M. vulgaris*. From the outer to the inner circles are: The chromosomes; gene density; density of DNA elements; density of LINE; density of SINE; density of LTR (densities shown as percent nucleotides per 100 Kb). (B) Syntenic blocks within the *M. vulgaris* genome (6,453 gene pairs, 298 blocks which consisted of four continuous genes at least).

Full-size DOI: 10.7717/peerj.10313/fig-1

(Table S3). The genome features, including the distribution of genes and repeats along the chromosomes, are shown in Fig. 1A.

Evolutionary and comparative genomic analysis

Intraspecific synteny analysis was performed with all the gene models of *M. vulgaris* using self pair-to-pair alignment. Approximately 25.7% of the protein coding genes had synteny blocks (6,453 gene pairs, 298 blocks with at least four continuous genes) within the assembled genome. Intra-genomic gene comparison showed that co-linearity was distributed along almost the entire 41 chromosomes (Fig. S4), indicating the whole genome duplication occurred during the evolution of the *M. vulgaris* genome (Fig. 1B). Chromosome fusion events could also be inferred, including the event in which three chromosomes merged into chromosome number one (MV001), and two chromosomes merged into chromosome number two (MV002), creating the unusually large size of the two largest chromosomes (50.2 Mb and 30.9Mb, respectively). Gene synteny analysis between *M. vulgaris* and other species (*H. robusta*, *C. teleta*, *L. gigantea*, *C. elegans*) was also conducted. However, few instances of gene synteny could be identified in all genomes, revealing the different genome structure and genome evolution of *M. vulgaris*.

A total of 310 singlecopy ortholog gene families were selected from *M. vulgaris* and eleven other published animals, including *Drosophila melanogaster*, *Caenorhabditis elegans*, *Lottia gigantea*, *Capitella teleta*, *Helobdella robusta*, *Apostichopus japonicus*, *Danio rerio*, *Xenopus tropicalis*, *Gallus gallus*, *Homo sapiens*, and *Mus musculus*. *M. vulgaris* and *H. robusta* (leech) are in the same branch of the phylogenetic tree with *C. teleta* in the nearest branch (Fig. 2A). All three species are Annelida but *C. teleta* is considered to be more primordial than the other two (*Simakov et al., 2013*), which is supported by the whole genome data. Common gene families were identified from the four Lophotrochozoa genomes (Fig. 2B), which

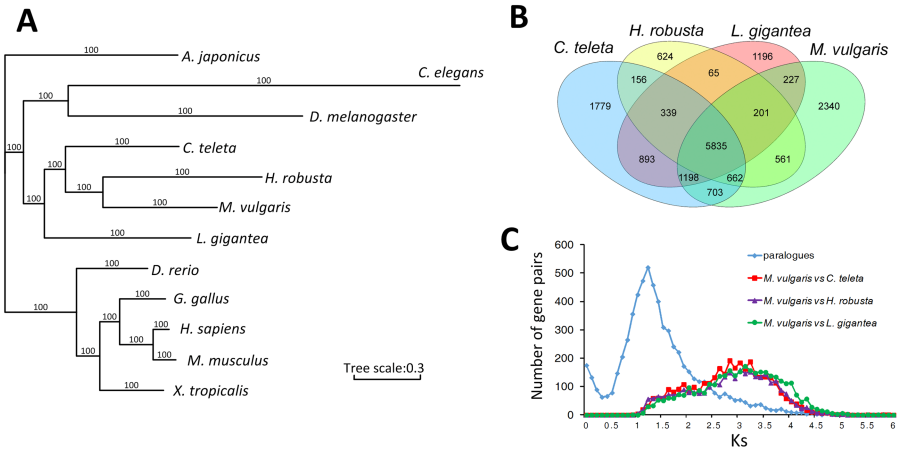

Figure 2 **Comparison of homologue genes between *M. vulagris* and other species.** (A) Phylogenic tree shows the relationships among metazoans by whole genome evidence of ortholog single copy gene families (*A. japonicus* was set as the outgroup); (B) Venn plot of the common identified gene families among the four Lophotrochozoa Genomes; (C) the Ks distributions of ortholog genes between *M.vulgaris* and other three Lophotrochozoa species, along with the Ks of paralog genes within *M.vulgaris*.

Full-size DOI: 10.7717/peerj.10313/fig-2

showed that *M. vulgaris* had 2,340 specific genes. A one-to-one comparison of the ortholog genes of *M. vulgaris* and the other three Lophotrochozoa species determined that the divergence of *M. vulgaris* was similar, while the Ks distribution of paralog genes within *M. vulgaris* showed a much earlier peak (Fig. 2C).

Hox genes and lumbrokinase genes

A total of 343 homeobox genes were retrieved from *M. vulgaris*, which is consistent with previous studies (*Zwarycz et al., 2015*). 41 Hox genes, which belong to the HOXL subclass of ANTP class, were identified by a homeodomain search (Fig. S5). Among them, 28 genes were clustered in 9 chromosomes. The Hox gene clusters of *M. vulgaris* and their paralog groups in other five genomes (*D. melanogaster*, *C. elegans*, *L. gigantean*, *C. teleta*, *H. robusta*) were analyzed together due to their relatively good genome assemblies. There were more Hox gene clusters in *M. vulgaris* than in other species but the order of the paralog genes was similar (Fig. 3A, Table S4) and the distribution pattern was similar to that of *H. robusta*, which is also in Annelida. The expression levels of the Hox genes in *M. vulgaris* were evaluated in the heart, ventral nerve cord, and tail indicating that *Hox1(Lab)/Hox2(Pb)/Hox3(Zen)* were mostly expressed in the heart, *Hox4(Dfd)/Hox5(Scr)/Hox6(Lox5)/Hox7(Antp)/Ubx(Lox4)/Lox2* were mostly expressed in the ventral nerve cord, and *Post2/Post1* were expressed in the tail (Fig. 3B). This showed tissue-specific expression within the genes of the Hox paralog groups.

By homologue search against the *M. vulgaris* genes, twenty lumbrokinase-like genes were found within the whole genome and most of them were tandemly arranged (Fig. S6, Table S5). They were distributed on six chromosomes (MV001, MV005, MV010, MV017, MV020 and MV026) and one unanchored contig (MV154).

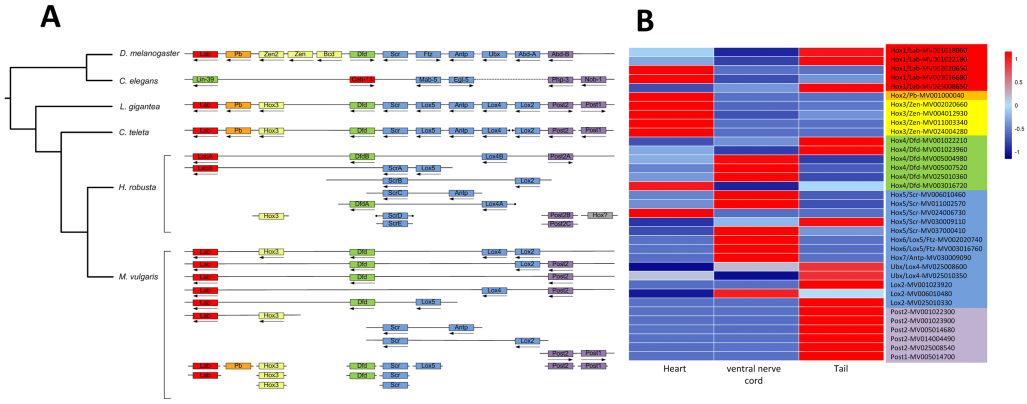

Figure 3 Hox gene complex and their expression in *M. vulgaris*. (A) The clusters of Hox genes in *M. vulgaris* and other five species. All the six species have been sequenced and assembly with good quality. The arrows, dots, and colors were defined as those from *Simakov et al. (2013)*. Red, orange, yellow, green, blue, purple were assigned to paralog groups (Hox1, Hox2, Hox3, Hox4, central class and posterior class). Arrows indicated the direction of transcription. The ends of the scaffolds were marked by black dots; (B) the expression of the Hox genes in heart, ventral nerve cord and tail of *M. vulgaris*. The colors of the gene name on the right side were according to the Hox paralog groups described above.

Full-size DOI: 10.7717/peerj.10313/fig-3

DISCUSSION

The karyotype information for earthworms in Lumbricidae was published many years ago (*Gregory & Hebert, 2002*), revealing that the C-values ranged from 0.43–1.20 and chromosome numbers $2n = 22$–190. However, there have been no recent studies on the chromosomes of any species in Megascolecidae and there has been no whole genome sequencing for the species in the *Metaphire* genus, with the exception of the mitochondrial genome (*Zhang et al., 2016*). We performed karyological analysis of this earthworm to verify the number of its chromosomes to build a better genome assembly of the newly sequenced species, *M. vulgaris*. 41 pairs of chromosomes were identified, which corresponded to the number of clusters obtained from Hi-C data. Intraspecific synteny analysis showed a whole genome duplication of the *M. vulgaris*, followed by several chromosome fusion events, revealing the evolutionary route from an ancient tetraploid to a modern diploid species. The genome duplication (tetraploidization) may have led to the bigger body size of *M. vulgaris* versus other common earthworms.

Previous studies reported on two genome assemblies of earthworms (*Zwarycz et al., 2015*; *Bhambri et al., 2018*), both on *Eisenia fetida*, from the Lumbricidae family. Neither study produced a high-quality chromosome level genome assembly. The N50 scaffold lengths were only 9.31 Kb and 1.85 Kb, respectively, which was not long enough to identify an intact gene model. We generated high-quality genome and transcriptome datasets for *M. vulgaris* and created the first chromosome-level genome assembly of the oligochaete species, with almost all of the contigs clustered into 41 groups or chromosomes. The N50 contig length of the *M. vulgaris* genome assembly was as long as 4.2 Mb, which was hundreds of times larger than those of *Eisenia fetida*. This assembly was suitable for gene modeling and

synteny analysis. The subsequent results of gene annotation, gene expression, and genome structure will provide valuable information for gene function and evolutionary studies.

Hox genes are a subset of homeobox genes and are a type of animal gene that specifically regulates biological structures. The Hox genes tend to be clustered in the genome and show a collinear correspondence between gene order and the body levels where these genes are expressed during development (*Duboule, 2007*). In our findings, Hox genes partial clustered around the *M. vulgaris* genome and showed tissue-specific expression. Lumbrokinase is a six-enzyme protein with fibrinolytic activity, which was first isolated from the crude of *Lumbricus rubellus* (Hoffmeister, 1843; *Mihara et al., 1991*). Fibrinolytic enzymes have been purified and characterized in *Eisenia foetida*, (*Hrzenjak et al., 1998*; *Li, Zhao & He, 2003*; *Wang et al., 2003*). And lumbrokinase is a medicinally valuable enzyme used to dissolve thrombi, reduce blood viscosity, and inhibit platelet aggregation (*Wang et al., 2013*). The identification of lumbrokinase-like genes of *M. vulgaris* were performed in the whole genome level, which indicated that they might tend to be clustered or tandemly distributed in earthworm genomes. The high-quality genome assembly and annotation of *M. vulgaris* will lay the foundation for the study of functional genes in earthworms, such as lumbrokinase and drilodefensin (*Liebeke et al., 2015*).

CONCLUSIONS

We sequenced, assembled, annotated, and analyzed the genome of the *M. vulgaris*, which belongs to the genus *Metaphire* of the family Megascolecidae. The assembled sequence consisted of 559 contigs, with a length of 729 Mb, and an N50 contig size of 4.2 Mb. The contigs were anchored onto 41 chromosomes according to the Hi-C result, which was verified by karyological analysis. Whole-genome duplication and chromosome fusion events have been observed within the *M. vulgaris* genome. Hox genes and lumbrokinase genes were identified at the whole genome level and both of them were distributed as partial clusters. The high-quality genome assembly of *M. vulgaris* may be a valuable resource for genetic studies, gene cloning, and phylogenetic analysis in earthworms.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was supported by the National Natural Science Foundation of China (Grant No. 81503640, 41771279 and 81830052); Construction project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400); Shanghai Municipal Education Commission (Class II Plateau Disciplinary Construction Program of Medical Technology of SUMHS, 2018-2020). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 81503640, 41771279, 81830052.

Construction project of Shanghai Key Laboratory of Molecular Imaging: 18DZ2260400. Shanghai Municipal Education Commission (Class II Plateau Disciplinary Construction Program of Medical Technology of SUMHS, 2018–2020).

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Feng Jin and Qiang Zhao conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Zhaoli Zhou analyzed the data, prepared figures and/or tables, and approved the final draft.
- Qi Guo conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Zhenwen Liang, Ruoyu Yang and Yanlin He performed the experiments, prepared figures and/or tables, and approved the final draft.
- Jibao Jiang performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Qi Zhao conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Field Study Permissions

The following information was supplied relating to field study approvals (i.e., approving body and any reference numbers):

The earthworms were collected in a typical farmland habitat with the permission of the farmer Genqing Wang from Yilong Farm (Jiangsu, China).

Data Availability

The following information was supplied regarding data availability:

Raw-data and the genome assembly are available at GenBank: PRJNA656665. This includes the Illumina whole-genome shotgun sequence (SRR12458316), Oxford Nanopore sequence (SRR12436614) and Hi-C raw-data (SRR12458315) of *M. vulgaris*; and the RNA-seq data for the six tissues (SRR12458313, SRR12458314, SRR12458317–SRR12458320).

Raw data are also available at Bioproject: PRJCA002730.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10313#supplemental-information.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215(3)**:403–410 DOI 10.1016/S0022-2836(05)80360-2.

Bhambri A, Dhaunta N, Patel SS, Hardikar M, Bhatt A, Srikakulam N, Shridhar S, Vellarikka S, Pandey R, Jayarajan R, Verma A, Kumar V, Gautam P, Khanna Y, Khan JA, Fromm B, Peterson KJ, Scaria V, Sivasubbu S, Pillai B. 2018. Large scale changes in the transcriptome of *Eisenia fetida* during regeneration. *PLOS ONE* **13(9)**:e0204234 DOI 10.1371/journal.pone.0204234.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* **31(12)**:1119–1125 DOI 10.1038/nbt.2727.

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. 2006. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics* **7**:327 DOI 10.1186/1471-2164-7-327.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34(17)**:i884–i890 DOI 10.1093/bioinformatics/bty560.

Chinese Pharmacopoeia Commission. 2015. *Pharmacopoeia of the People's republic of China 2015 (Chinese)*. Vol. 1. Beijing: China Medical Science Press, 122–123.

Csuzdi C. 2012. Earthworm species, a searchable database. *Opuscula Zoologica Budapest* **43(1)**:97–99.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29(1)**:15–21 DOI 10.1093/bioinformatics/bts635.

Duboule D. 2007. The rise and fall of hox gene clusters. *Development* **134(14)**:2549–2560 DOI 10.1242/dev.001065.

Eddy SR. 2011. Accelerated profile HMM searches. *PLOS Computational Biology* **7(10)**:e1002195 DOI 10.1371/journal.pcbi.1002195.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16(1)**:157 DOI 10.1186/s13059-015-0721-2.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29(7)**:644–652 DOI 10.1038/nbt.1883.

Gregory TR, Hebert PDN. 2002. Genome size estimates for some oligochaete annelids. *Canadian Journal of Zoology* **80**:1485–1489 DOI 10.1139/Z02-145.

Guan DL, Yang J, Liu YK, Li Y, Mi D, Ma LB, Wang ZZ, Xu SQ, Qiu Q. 2020. Draft genome of the Asian buffalo leech *Hirudinaria manillensis*. *Frontiers in Genetics 16* **10**:1321 DOI 10.3389/fgene.2019.01321.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9(1)**:R7 DOI 10.1186/gb-2008-9-1-r7.

Hrzenjak T, Popović M, Bozić T, Grdisa M, Kobrehel D, Tiska-Rudman L. 1998. Fibrinolytic and anticoagulative activities from the earthworm *Eisenia foetida*.

Comparative Biochemistry and Physiology B-Biochemistry and Molecular Biology **119**(**4**):825–832 DOI 10.1016/s0305-0491(98)00060-1.

Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, Caradec C, Davidas S, Cruaud C, Liti G, Lemainque A, Engelen S, Wincker P, Schacherer J, Aury JM. 2017. De novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6**(**2**):1–13 DOI 10.1093/gigascience/giw018.

Jiang JB, Qiu JP. 2018. Origin and evolution of earthworms belonging to the family Megascolecidae in China. *Biodiversity Science* **26**(**10**):1074–1082 DOI 10.17520/biods.2018105.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**(**8**):907–915 DOI 10.1038/s41587-019-0201-4.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**(**5**):722–736 DOI 10.1101/gr.215087.116.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**:59 DOI 10.1186/1471-2105-5-59.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* **19**(**9**):1639–1645 DOI 10.1101/gr.092759.109.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33**(**7**):1870–1874 DOI 10.1093/molbev/msw054.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(**4**):357–359 DOI 10.1038/nmeth.1923.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X Version 2.0. *Bioinformatics* **23**(**21**):2947–2948 DOI 10.1093/bioinformatics/btm404.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(**18**):3094–3100 DOI 10.1093/bioinformatics/bty191.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(**14**):1754–1760 DOI 10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(**16**):2078–2079 DOI 10.1093/bioinformatics/btp352.

Li L, Zhao J, He RQ. 2003. Isolation and some characterizations of a glycosylated fibrinolytic enzyme of earthworm *Eisenia foetida*. *Protein and Peptide Letters* **10**(**2**):183–190 DOI 10.2174/0929866033479095.

Liebeke M, Strittmatter N, Fearn S, Morgan J, Kille P, Fuchser J, Wallis D, Palchykov V, Robertson J, Lahive E, Spurgeon D, McPhail D, Takats Z, Bundy J. 2015. Unique

metabolites protect earthworms against plant polyphenols. *Nature Communications* **6**:7869 DOI 10.1038/ncomms8869.

Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* **42(15)**:e119 DOI 10.1093/nar/gku557.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15(12)**:550 DOI 10.1186/s13059-014-0550-8.

Mihara H, Sumi H, Yoneta T, Mizumoto H, Ikeda R, Seiki M, Maruyama M. 1991. A novel fibrinolytic enzyme extracted from the earthworm *Lumbricus rubellus*. *Japanese Journal of Physiology* **41(3)**:461–472 DOI 10.2170/jjphysiol.41.461.

Paul S, Arumugaperumal A, Rathy R, Ponesakki V, Arunachalam P, Sivasubramaniam S. 2018. Data on genome annotation and analysis of earthworm *Eisenia fetida*. *Data Brief* **20**:525–534 DOI 10.1016/j.dib.2018.08.067.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33(3)**:290–295 DOI 10.1038/nbt.3122.

Ronquist F, Teslenko M, Van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61(3)**:539–542 DOI 10.1093/sysbio/sys029.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31(19)**:3210–3212 DOI 10.1093/bioinformatics/btv351.

Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH, Larsson T, Lv J, Arendt D, Savage R, Osoegawa K, De Jong P, Grimwood J, Chapman JA, Shapiro H, Aerts A, Otillar RP, Terry AY, Boore JL, Grigoriev IV, Lindberg DR, Seaver EC, Weisblat DA, Putnam NH, Rokhsar DS. 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493(7433)**:526–531 DOI 10.1038/nature11696.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**:31 DOI 10.1186/1471-2105-6-31.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**:62 DOI 10.1186/1471-2105-7-62.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9(11)**:e112963 DOI 10.1371/journal.pone.0112963.

Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, Lanz C, Weigel D. 2015. Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Research* **25(2)**:246–256 DOI 10.1101/gr.170332.113.

Wang DP, Wan HL, Zhang S, Yu J. 2009. Gamma-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. *Biology Direct* **4**:20 DOI 10.1186/1745-6150-4-20.

Wang F, Wang C, Li M, Cui L, Zhang J, Chang W. 2003. Purification, characterization and crystallization of a group of earthworm fibrinolytic enzymes from *Eisenia foetida*. *Biotechnology Letters* **25**(**13**):1105–1109 DOI 10.1023/a:1024196232252.

Wang KY, Tull L, Cooper E, Wang N, Liu D. 2013. Recombinant protein production of earthworm lumbrokinase for potential antithrombotic application. *Evidence-based Complementary and Alternative Medicine* **2013**:783971 DOI 10.1155/2013/783971.

Zdobnov EM, Apweiler R. 2001. InterProScan–an integration platform for the signature -recognition methods in interpro. *Bioinformatics* **17**(**9**):847–848 DOI 10.1093/bioinformatics/17.9.847.

Zhang L, Jiang J, Dong Y, Qiu J. 2016. Complete mitochondrial genome of a pheretimoid earthworm *Metaphire Vulgaris* (Oligochaeta: Megascolecidae). *Mitochondrial DNA Part A: DNA Mapping, Sequencing, and Analysis* **27**(**1**):297–298 DOI 10.3109/19401736.2014.892085.

Zwarycz AS, Nossa CW, Putnam NH, Ryan JF. 2015. Timing and scope of genomic expansion within Annelida: evidence from Homeoboxes in the Genome of the Earthworm *Eisenia fetida*. *Genome Biology and Evolution* **8**(**1**):271–281 DOI 10.1093/gbe/evv243.