The Israeli Journal of Aquaculture - Bamidgeh, IJA_71.2020.1105393, 16 pages
CCBY-NC-ND-4.0 • https://doi.org/10.46989/001c.21465

De novo transcriptome analysis of the sandworm (*Sipunculus nudus*) and identification of differentially expressed genes associated with body size

Xiaohui Cai[1#], Jing Fang[1#], Yuna Zhou[2], Honglin Chen[1,3]
Xinzhong Wu[1*], Yinhui Peng[1*]

[1] Guangxi Key Laboratory of Beibu Gulf Marine Biodiversity Conservation, Beibu Gulf University, Qinzhou, 536011, China

[2] Beihai Fishery Technology Extension Station, Beihai, 536000, China

[3] Guangxi University College of Life Science and Technology, Nanning, 530000, China

Key Words: Sandworm (*Sipunculus nudus*); Transcriptome; Body size; Body wall muscle

Abstract

The sandworm (*Sipunculus nudus*) is an aquatic species of economic importance because of its high nutritional and medicinal value. Under the same culture conditions, substantial individual growth variation is often found in populations of sandworms. However, the genetic mechanisms of individual growth variation are poorly understood. In this study, the transcriptome of the body wall muscle of the sandworm at different growth rates was analyzed by Illumina sequencing and bioinformatics analysis. A total of 185 181 unigenes were obtained after processing raw reads and about 96,824 (47.72%) of them were annotated. Among the annotated transcripts, 418 differentially expressed genes were identified, of which 207 were upregulated and 211 were downregulated in large worms relative to small worms. We identified several genes that had a possible association with individual growth variation. These results will provide insight into the growth mechanism of sandworm, and will further assist in the selective breeding of improved strains of this species.

* Corresponding author. e-mail: wuxzqinzhou@163.com; pyinhui@163.com.

Introduction

The sandworm *Sipunculus nudus* belongs to the phylum Sipuncula and has a global distribution in subtidal zones or the seabed of temperate or tropical sea along the coast of the Indian Ocean, the Western Pacific, and the Atlantic Ocean. The production of sandworms can reach around 20,000 tons each year in China at a price of approximately 14 USA dollars kg^{-1} (fresh weight) (Li et al., 2017). Because of their high nutritional and medicinal value, sandworms have become an important part of the fisheries economy. For example, as an edible marine organism, it has been popular at the dinner table for its delicious taste in southeast China for many years and has been used as functional seafood because of its diverse nutritional and functional components consisting of free amino acids, fatty acids, polysaccharides, mineral elements, and so on (Ge et al., 2016). Furthermore, sandworms have long been used as a traditional Chinese medicine for the treatment of several diseases such as carbuncles, tuberculosis, and nocturia, regulating the functions of the stomach and the spleen as well as in treating disabilities caused by various pathogens and aging (Su et al., 2016). However, with the over-exploitation of natural resources and the destruction of habitats by human activities, the wild population and germplasm resources of the sandworm has declined sharply. Therefore, the breeding of sandworms is an imminent necessity.

Since the breakthrough of the artificial seedling and intermediate culture technology of sandworms in 2010 by the Guangxi Institute of Oceanography, an artificial aquaculture of sandworms has been successfully realized in China (Peng et al., 2015). The average yield of sandworms from artificial seedling farms has reached 140.98–331.49 g/m^2 in Qiaogang Town, Fucheng Town, Xichang Town, and Beihai City, Guangxi Province. During the same period, the daily quota of sandworms dug by fishermen was 2–5 kg/person, which is more than five times higher than the 0.5–1.0 kg/person 5 years ago. This indicated that the natural resource of sandworms has been restored with the extension of the whole artificial culture technology (Yang et al., 2013). At present, sandworms are cultured along the beaches without a supplemental artificial diet, and adult worms can bury themselves into sandy substrates at a depth of 20–50 cm depending on the nutrients from the surface sediment, which include microalgae and other organic matter (Li et al., 2015). Under the same culture conditions, a substantial individual growth variation is often found in populations of sandworms. It is a common phenomenon that differences in growth rate are exhibited by different individuals under similar environmental conditions, which is usually represented by the coefficient of variation (CV) of growth. Previous studies have shown that the CV within most fish species is 20±35% and over 50% in invertebrate sea cucumber (*Apostichopus japonicus*) (Gao et al., 2017). The individual growth variation has already seriously affected cultural production and has caused economic losses to farmers. Previous studies have shown that both genetic and environmental differentiation have an impact on the individual variation, with genetic differentiation having the greater impact (Liang et al., 2010). A total of 168 differentially expressed genes (DEGs) were identified between fast growing *Pinctada fucata martensii* and a slow growing group by transcriptome analysis (Hao et al., 2019). The pathway enrichment analysis indicated that DEGs were involved in extracellular matrix (ECM)-receptor interaction, pentose phosphate pathway, and aromatic compound degradation. The fast growing individuals exhibited better digestion, anabolic ability, and osmotic regulation than the slow growing group in response to environmental stress. The sea cucumber (*Apostichopus japonicus*) transcriptome with individual growth variation was also characterized and the candidate growth-related genes and potential growth molecules were identified (Gao et al., 2017).

In the present study, the body wall transcriptome of sandworm with different growth rates was analyzed by Illumina sequencing and bioinformatics. This is the first report characterizing the sandworm transcriptome with individual growth variation, aiming to identify candidate growth-related genes and investigate potential growth molecules. The results will support further investigations of the growth mechanism in sandworm and will improve the current understanding of individual growth variation.

Materials and Methods

Animal materials and sample preparation. Parent wild sandworms were collected from the Fucheng sea area of the Beibu Gulf. According to the unbalanced nested design and the principle of one male and four females, 21 paternal and half-sib families were constructed and the larvae were cultured under the same conditions, placed in a sand pond and fed once a day. At 180 days post-hatch (dph), the sandworms from the same family were sorted by size, and fast and slow growing individuals were sampled. The body weight of each sample was measured using an electronic balance (0.01 g accuracy). To identify genes that potentially influence body size, 20 individuals with extreme phenotypes were selected in this study. Worms with relatively large body size (mean body weight 2.03±0.05 g) were placed in one group and worms with relatively small body size (mean body weight 0.25±0.03g) were placed in another group. The sandworms were euthanized with prilocaine hydrochloride (0.25–0.5%) for 5–10 min (Podolak-Machowska et al., 2014). The body wall muscle tissues were immediately frozen in liquid nitrogen and stored at –80°C until further use. The animal protocol was approved by the Institutional Animal Care and Use Committee of Qinzhou University (Qinzhou, Guangxi, China).

RNA extraction, library construction, and transcriptome sequencing. All samples were flash-frozen in dry ice and transported to Novogene Bioinformatics Institute (Beijing, China) for RNA extraction, library construction, and transcriptome sequencing. In brief, total RNA was isolated with TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. RNA degradation and contamination were monitored with 1% agarose gels. RNA purity was assessed using a NanoPhotometer spectrophotometer (Implen, Westlake Village, CA, USA). RNA concentration was tested using a Qubit RNA Assay Kit (Invitrogen). RNA integrity was checked in an Agilent Bioanalyzer 2100 system (Agilent Technologies, Palo Alto, CA, USA). A total amount of 6 μg RNA from 10 sandworms in each group was pooled to construct the cDNA libraries. Poly (A) mRNA was purified from the total RNA using oligo (dT) magnetic beads. Equal amounts of high quality mRNA samples were obtained from each group for cDNA library preparation using a NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA) and purified using Agencount AMPure XP beads (Beckman Coulter, Krefeld, Germany) according to the manufacturer's recommendations. The concentration of the cDNA library was determined with an Agilent Bioanalyzer 2100 system. Transcriptome sequencing of the library preparations was carried out on an Illumina HiSeq 2000 platform by Novogene (Beijing, China) to obtain paired-end reads.

Data processing, assembly, and functional annotation. Raw reads generated by Illumina HiSeq 2000 were cleaned by removing the adaptor containing sequences, clean reads were obtained after removing any ambiguous bases >0.1% reads and low quality reads (1/2 reads with *q*-value ≤20). De novo transcriptome assembly was carried out with a short-read assembling program-Trinity. To avoid redundant annotations, only one unigene (the longest transcript) was selected and compared with the Swiss-Prot database using BLASTx with an E-value threshold of $1e^{-5}$. The annotation of gene function was based on the following databases: Nr, Nt, Pfam, KOG/COG, Swiss-Prot, KEGG, and gene ontology (GO). GO analysis was performed by the topGO R package based on the Kolmogorov-Smirnov test through searching the above databases. The GO terms assigned biological process, molecular function, and cellular component to the query sequences. KEGG analysis was performed using the KOBAS software to test the statistical enrichment of DEGs in KEGG pathways.

Differentially expressed genes between the small/large-worm libraries. The calculation of unigene expression levels and the identification of unigenes that were differentially expressed between the libraries were performed by DEGseq based on TMM normalized counts. Gene expression differences were considered to be significant if the *q*-value <0.005 and the absolute value of log2FoldChange >1. The DEGs were annotated by BLASTx alignment with an E-value threshold of $1e^{-5}$. Subsequently, GO and KEGG databases were searched for enriched biological processes or known pathways in the genes which were differentially expressed.

Real-time PCR assays. To further validate the confidence of transcriptome sequencing, six DEGs were selected and analyzed via real-time PCR. Primers (listed in **Table S1**) were designed based on sequences from the RNA-Seq unigene sequences by Primer 5.0. The elongation factor 1-alpha (EF1-ɑ1) gene was used as an internal reference. Total RNA was extracted from body wall muscle tissues of six sandworms (three biological replicates per sample). First-strand cDNA was synthesized from 0.5 μg total RNA with TransScript One-Step gDNA Removal and cDNA Synthesis SuperMix (Transgen Biotech, Beijing, China) according to the manufacturer's protocol. Then, qPCR was conducted using TransStart Top Green qPCR SuperMix (Transgen Biotech, Beijing, China) in 20 μL reactions, containing 10 μL TransStart Top Green qPCR SuperMix (2×), 0.4 μL forward primer (10 μM), 0.4 μL reverse primer (10 μM), 0.4 μL Passive Reference Dye (50×), 1 μL cDNA, and 7.8 μL ddH_2O. The PCR amplification procedure was carried out at 94°C for 30 s, followed by 40 cycles of 94°C for 5 s and 60°C for 30 s; Melt curve analysis of the amplification product was performed over a range of 70–95°C at the end of each PCR reaction aiming to confirm single product generation. Samples were run in triplicate on the QuantStudio™ 6 Flex Real-Time PCR System (ThermoFisher Scientific Inc., Waltham, MA, USA). The relative expression fold changes of six genes in small versus large body wall muscles were analyzed using the $2^{-\Delta\Delta Ct}$ method. All quantitative data were presented as the means ± standard deviation (SD). Statistical analysis was performed using SPSS statistics 17.0 software. A p-value of less than 0.05 was considered to be significant.

Results

Sequence analysis and assembly. To understand the molecular mechanism of individual growth variations in sandworms, cDNA libraries were constructed from groups of worms that were either significantly large or small at 180 dph ($p<0.01$). RNA-Seq produced 62,408,178 (large worms) and 75,134,818 (small worms) clean reads with a Q20 of 98.07% and 97.99%, encompassing 9.36 Gb and 11.27 Gb sequencing data, respectively (**Table 1**). The high-quality clean reads were de novo assembled and generated 185,181 unigenes with a mean length of 681 bp and an N50 length of 1,296 bp. The unigene lengths ranged from 201 to 35,022 bp (**Table 1**). In detail, 129,400 (69.88%) unigenes were 200–499 bp in length, 26,802 (14.47%) unigenes were 500–999 bp in length, 15,484 (8.36%) were 1,000–1,999 bp, and 13,495 (7.29%) unigenes were >2,000 bp in length (**Figure 1**).

Table 1 Summary for the Illumina sequencing, de novo assembly and annotations of sandworm transcriptome.

	Sequencing and preprocessing	
samples	*BI_m*	*SI_m*
No of Clean Reads	62 408 178	75 134 818
No of clean Bases	9.36G	11.27G
Q20(%)	98.07	97.99
	Assembly (De novo)	
No of UniGenes	185 181	
Average length (bp)	681	
Min-Max length (bp)	201-35 022	
N50 (bp)	1 296	
	Annotations	
	Number	Percentage (%)
Nr	2 643	1.42
Nt	9 503	5.13
Ko	11 125	6
Swissport	3 926	2.12
Pfam	26 021	14.05
Go	26 021	14.05
Kog	17 585	9.49
All annotated	96 824	52.28

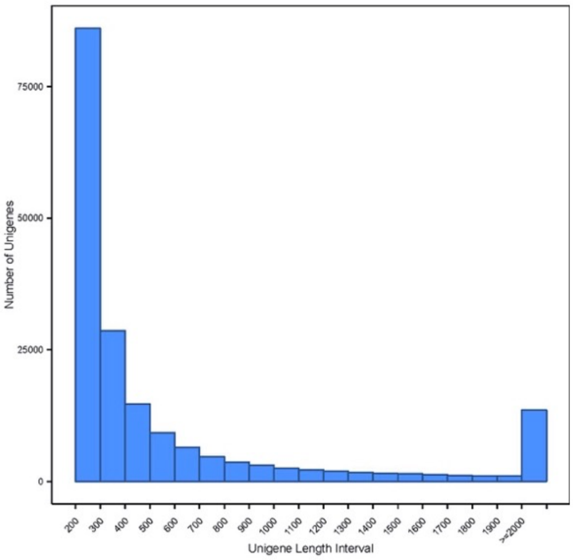

Figure 1 Length distribution of all unigenes of sandworm.

Functional annotation and classification. Results of functional annotation showed that 96,824 (52.28%) of the 185,181 unigenes were annotated against the Nt, Nr, KO, GO, KOG, Pfam, and Swiss-Prot databases, among which Pfam (14.05%) and GO (14.05%) contained the most homologies (**Table 1**), and 276 unigenes are annotated in all five databases, including Nr, Nt, KOG, GO, and Pfam as shown in the Venn diagram (**Figure 2a**). A total of 2,643 unigenes were annotated in the NCBI protein database. Of all annotated unigenes, 45% had homology with the aligned proteins (E-value<1E^{-45}) (**Figure 2b**). Moreover, 65.80% of the annotated unigenes displayed more than 60.00% similarity (**Figure 2c**). These results confirmed that the transcriptome data were successfully annotated although there is no reference genome for the sandworm. The remaining 88,357 (47.72%) unigenes had no BLAST hits in these databases, indicating that they might contain novel genes with unknown functions.

Based on the BLASTx similarity analysis, the unigenes matched sequences from a range of helminths, mollusks, insects, and other species (**Figure 2d**). In detail, the highest number of hits were with *Capitella teleta* (20.1%), followed by *Crassostrea gigas* (14.9%), *Branchiostoma floridae* (5.5%), *Lottia gigantean* (5.2%), and *Drosophila melanogaster* (5.2%). This result suggested that sandworm was closely related to *C. teleta*, but not highly homologous with this species in the Nr protein database.

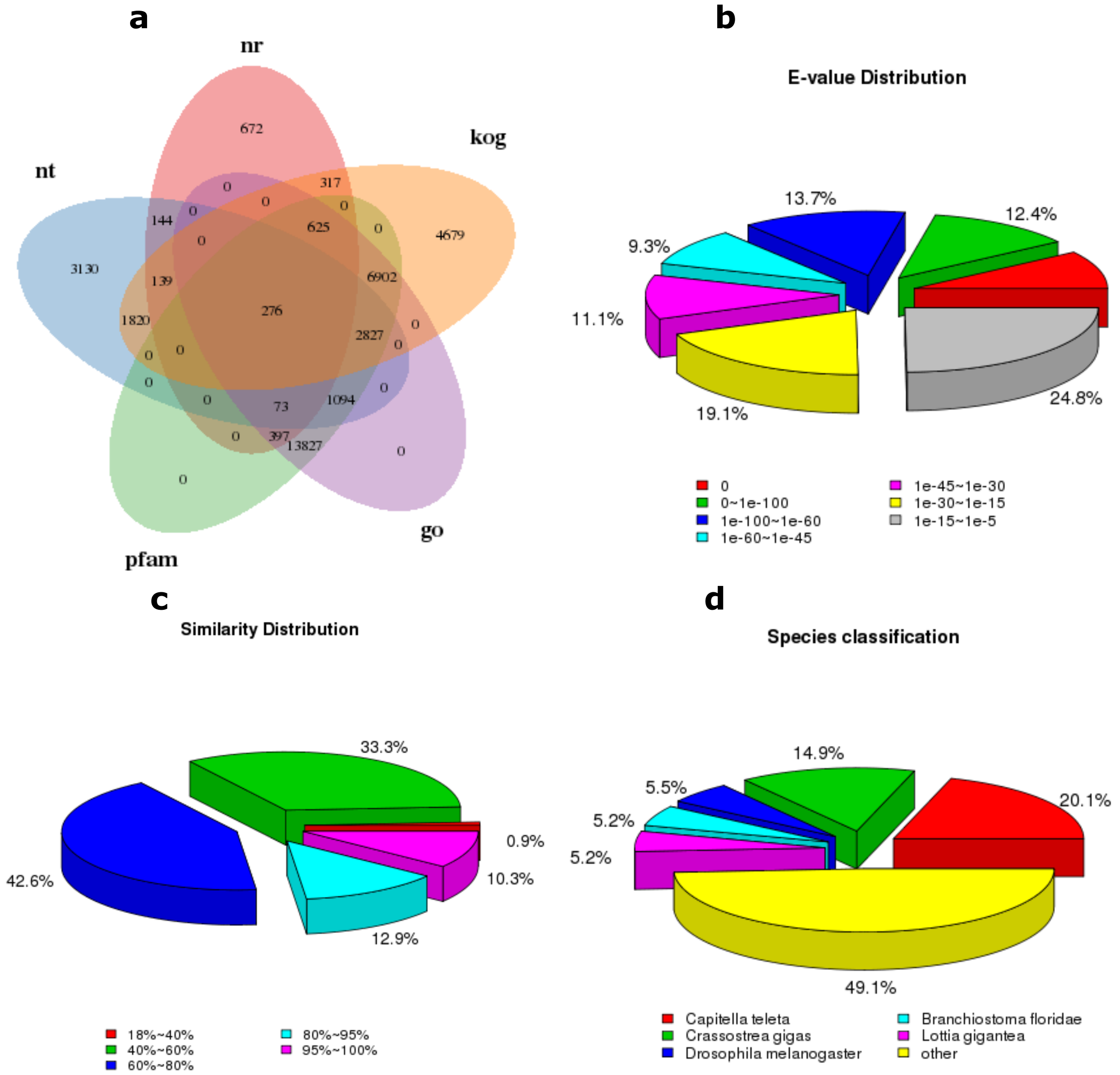

Figure 2 Function annotation of all unigenes of sandworm a: Venn graph b: E-value distribution c: Similarity distribution d: Species classification

After Blast2GO analysis, a total of 26,021 annotated unigenes in the GO database were assigned to 54 functional groups (**Figure 3a**). Of those, 60,885 (46.94%) transcripts comprised the largest category, biological process, followed by cellular component (41,245; 31.80%), and molecular function (27,570; 21.26%). According to KEGG pathway analysis, 12,633 unigenes were assigned to five main categories, which consisted of 32 different pathways (**Figure 3b**). Among the five main categories, metabolic pathways were the most abundant group with 3,674 (29.08%) genes, followed by organismal systems (3,554; 28.13%), environmental information processing (1,974; 15.63%), cellular processes (1,740; 13.77%), and genetic information processing (1,691; 13.39%).

Identification of differentially expressed genes. Through a comparative analysis of all unigenes between the large and small-size worms, a total of 418 DEGs were identified with the criteria of *q*-value <0.005 and the absolute value of log2FoldChange >1. Of these DEGs, 207 were upregulated and 211 were downregulated in the large relative to small worm groups (**Figure 4**).

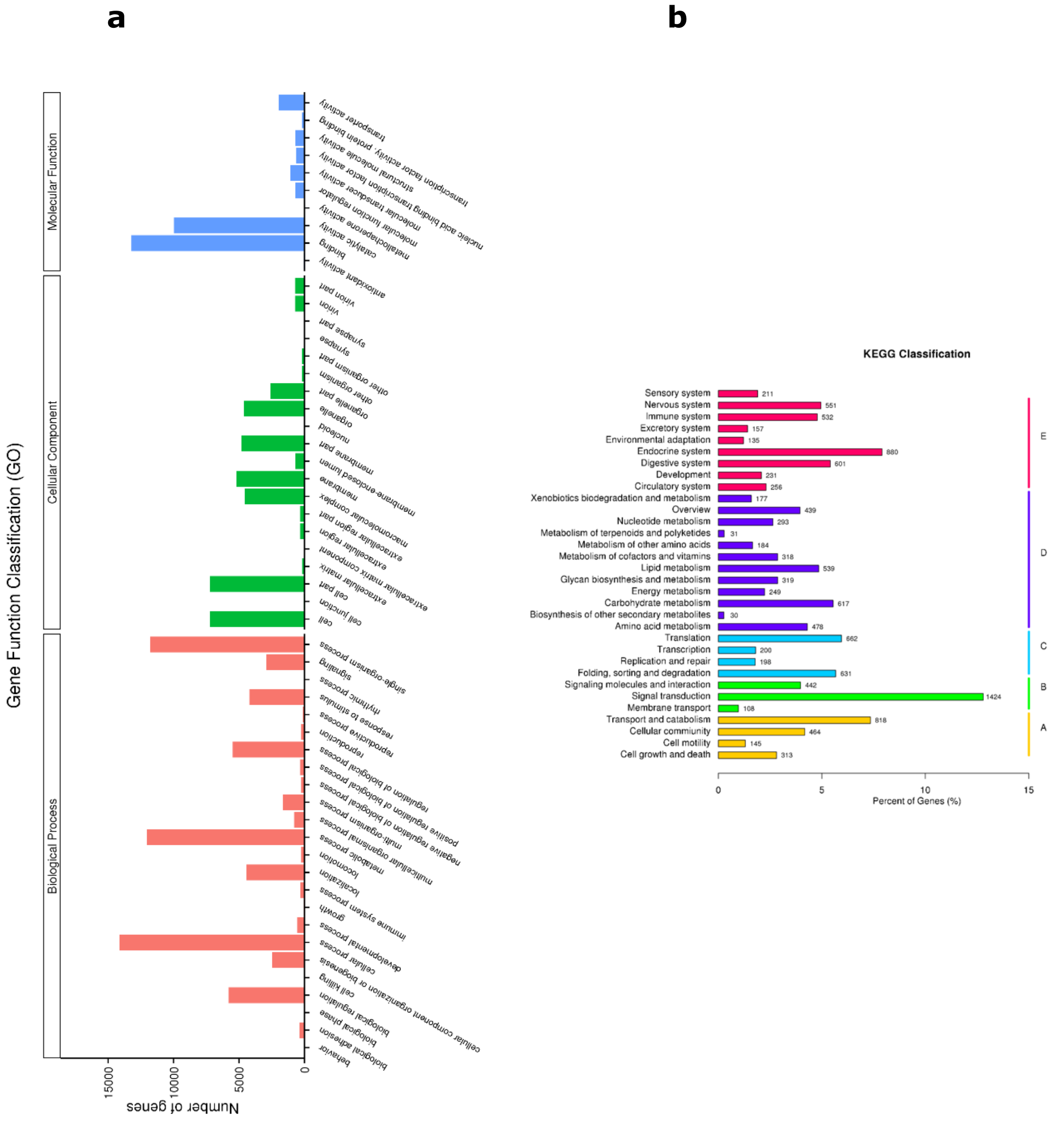

Figure 3 Gene Ontology (GO) and KEGG pathways functional classification of annotated unigenes. a: GO functional classification, a total of 26 021 unigenes with significant similarity in the GO database were assigned to three main categories: cellular component, molecular function and biological progress; b: KEGG pathway assignment based on five main categories: cellular processes (A), environmental information processing (B), genetic information processing (C), metabolism (D) and organismal systems (E).

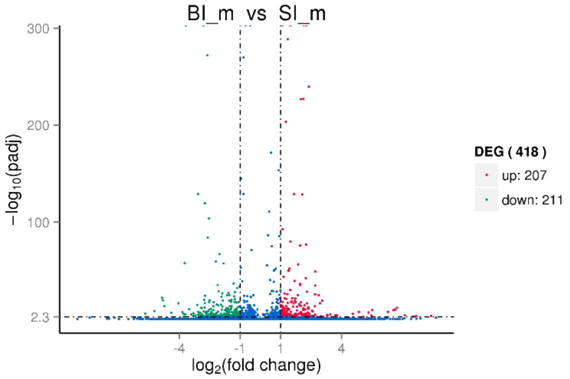

Figure 4 The volcano plot of differentially expressed genes between BI_m and SI_m sandworm. The X-axis represents fold change between BI_m and SI_m group, the Y-axis indicates significance of differential expression. The blue pots mean no significantly change unigenes (*p*>0.05, false discovery rate (FDR) *q*>0.05), while the red pots and the green pots mean up- and down-regulated unigenes (*p* <0.05, FDR *q*<0.05), respectively. .

To better understand the function of these DEGs, GO term enrichment was conducted and the DEGs were annotated with one or more GO terms, including 5,877 subclasses of biological process, 1,548 subclasses of cellular components, and 2,565 subclasses of molecular functions. In the biological process classification, the major categories represented were metabolic process (GO: 0008152), cellular process (GO:0009987), single organism process (GO:0044699), organic substance metabolic process (GO:0071704), and primary metabolic process (GO:0044238). Several unigenes were also involved in categories of cellular metabolic process, single organism cellular process, macromolecule metabolic process, and nitrogen compound metabolic process. Regarding the cellular components, the major categories represented were cell (GO:0005623), cell part (GO:0044464), intracellular (GO:0005622), membrane (GO:0016020), and intracellular part (GO:0044424). In the molecular function classification, binding (GO:0005488) was the most strongly represented GO term.

GO enrichment analysis of the 418 DEGs was performed in the three categories of cellular component, molecular function, and biological process. Among the three categories, the most DEGs were enriched in the cellular component category, which includes 31 transcripts in extracellular regions (GO: 0005576), 10 transcripts in the ECM (GO: 0031012), and eight transcripts in proteinaceous ECM (GO: 0005578). Regarding the biological process category, the most DEGs were assigned to the alpha-amino acid biosynthetic process term (GO:1901607, 10 transcripts). However, there were no significant enrichment results in the GO classification (Corrected p-Value>0.05) (**Table 2**).

Table 2 Gene Ontology (GO) classification of differentially expressed genes between BI_m and SI_m sandworm.

GO accession	*Description*	*Term type*	*p-Value*	*Corrected p-Value*	*DEG item*	*DEG list*
GO:0005576	extracellular region	cellular_component	1.4595e-05	0.072741	31	194
GO:0031012	extracellular matrix	cellular_component	6.9272e-05	0.17263	10	194
GO:0005578	proteinaceous extracellular matrix	cellular_component	0.00015822	0.19147	8	194
GO:1901607	alpha-amino acid biosynthetic process	biological_process	0.00018266	0.19147	10	194

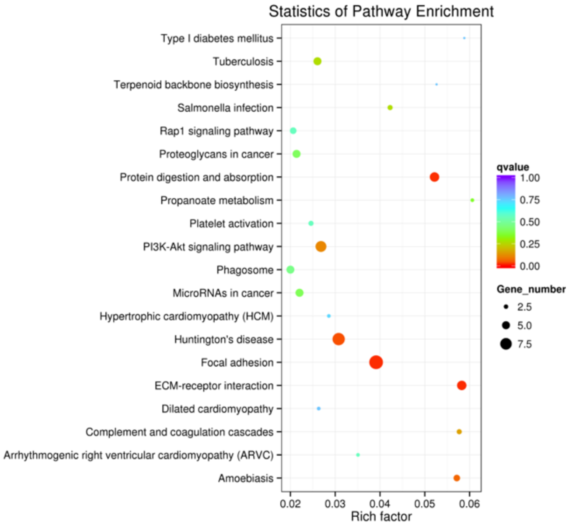

Figure 5 Scatter diagram of pathway enrichment for DEGs. In this scatter diagram, the top 20 pathways were listed, and rich factor is the ratio of DEGs in this pathway to all the genes in this pathway. The X-axis corresponds to rich factor of pathway, and the Y-axis represents different pathway. The magnitude of the pots displays gene number ranged from 10 to 20, and q-value is described by the color classification.

DEGs were further annotated by KEGG and classified into 107 subclasses. The top 20 of the most enriched KEGG pathways are shown in **Figure 5**, and four significant pathways ($p<0.05$), which included focal adhesion, ECM-receptor interaction, protein digestion and absorption, and Huntington's disease, were obtained, and 17 DEGs were associated with these four pathways (**Table 3**). However, the other changed pathways with no significant difference ($p>0.05$) included tuberculosis, salmonella infection, terpenoid backbone biosynthesis, type I diabetes mellitus, Rap1 signaling pathway, proteoglycans in cancer, propanoate metabolism, platelet activation, PI3K-Akt signaling pathway, phagosome, MicroRNAs in cancer, hypertrophic cardiomyopathy, dilated cardiomyopathy, complement and coagulation cascades, arrhythmogenic right ventricular cardiomyopathy, and amoebiasis.

Table 3 Pathway related to growth and development.

Pathways ID	*Name*	*Gene count*	*P value*	*Corrected P-Value*	*Genes*
ko04510	Focal adhesion	8	0.000148133	0.014084	COL5, COL4A, TN, COL6A, COL1AS, TSP1, ACTB_G1, FLNA
ko04512	ECM-receptor interaction	6	0.000260814	0.014084	COL5, COL4A, TN, COL6A, COL1AS, TSP1
ko04974	Protein digestion and absorption	6	0.000457069	0.016454	COL5, COL1AS, COL22A, COL4A, COL6A, COL12A
ko05016	Huntington's disease	7	0.001562708	0.042193	DNAHb1, DNAH2, DNAH5, DNAH6, DNAH7, DNAH12, atpB

Table 4 Genes of interest related to growth and muscle development in sandworm.

Unigene ID	*Gene Annotation*	*log2FC（FG/SG）*	*P value**	*FDR*
Dorso-ventral axis formation				
c104551_g1	neurogenic locus notch homolog protein 1	2.6901	2.32E-29	2.73E-26
c103370_g2	neurogenic locus notch homolog protein 2-like	2.2969	5.23E-32	6.80E29
TGF-beta signaling pathway				
c168063_g1	thrombospondin-1	2.685	9.98E-17	5.77E-14
c108880_g2	thrombospondin-4	1.1612	5.62E-07	0.0001048
c103595_g1	transforming growth factor beta	-1.5765	1.10E-07	2.27E-05
Osteoclast differentiation				
c95875_g1	four and a half LIM domains protein 2	1.5157	2.75E-21	2.14E-18
Cytoskeleton and myofibril component genes				
c86206_g2	skeletal organic matrix protein 7-like	4.8514	1.39E-10	4.37E-08
c93345_g1	skeletal organic matrix protein 7-like	3.7043	8.66E-06	0.0012592
c106970_g1	myosin-9-like	2.462	1.05E-05	0.0014977
c109026_g1	myosin-2 heavy chain	1.7988	3.44E-12	1.29E-09
c95633_g1	myosin-2 heavy chain	1.079	0	0
c110377_g1	myosin-2 heavy chain	1.0017	4.21E-18	2.68E-15
c44349_g1	actin beta/gamma 1	1.4789	0	0
c102346_g1	matrilin-2	2.6972	2.08E-38	3.13E-35
c107330_g1	matrilin-2	2.3585	4.26E-14	1.95E-11
c93702_g1	filamin	1.4798	4.67E-84	1.36E-80
c109617_g1	filamin	1.2506	3.14E-208	1.83E-204
c99782_g1	fibrillin-1-like	2.5402	4.14E-12	1.54E-09
c102657_g1	fibrillin-3-like isoform X3	2.3284	2.66E-06	0.0004306
c109653_g1	fibrillin-2-like isoform X1	2.3538	1.39E-15	7.23E-13
c89216_g1	fibrinolytic protein	2.1302	0	0
c99427_g1	tenascin W	1.1564	1.75E-06	0.0002972
c107224_g1	ankyrin	1.3962	9.69E-08	2.03E-05
c106085_g1	PDZ and LIM domain protein 5	1.344	2.91E-20	2.08E-17
c108817_g1	Intermediate filament tail domain-containing protein 1	1.146	6.95E-06	0.00104
c92606_g1	collagen, type I, alpha	2.1189	1.55E-231	1.04E-227
c94794_g5	collagen, type I, alpha	1.229	4.90E-07	9.31E-05
c96986_g2	collagen, type I, alpha chain-like	1.1231	3.29E-15	1.66E-12
c96789_g2	collagen, type I, alpha chain isoform X1	1.0982	2.06E-06	0.0003443
c96789_g3	collagen, type I, alpha	1.0535	3.87E-05	0.0049192
c96986_g3	collagen, type I, alpha chain isoform X1	1.05	1.60E-12	6.36E-10
c96944_g2	collagen type II，alpha-1 chain	2.3846	3.56E-244	2.59E-240
c94794_g6	collagen type II，alpha-1 chain isoform X1	1.2164	8.66E-14	3.84E-11

c96789_g1	collagen, type II, alpha-1 chain-like isoform X4	1.0807	2.02E-17	1.23E-14
c96986_g1	collagen, type II, alpha-1 chain prec□rsor	1.0184	5.17E-09	1.33E-06
c107642_g1	collagen, type IV, alpha	1.2988	1.21E-05	0.0016964
c94794_g4	collagen, type V, alpha	1.2732	5.57E-17	3.31E-14
c103119_g1	collagen, type VI, alpha-3	1.5455	5.91E-22	4.82E-19
c93327_g1	collagen, type VI, alpha	-1.8728	2.24E-16	1.25E-13
c94794_g3	collagen，type VII，alpha-1 chain isoform X3	1.1925	5.01E-25	4.80E-22
c102936_g1	collagen，type VII，alpha-1 chain	1.6576	3.52E-133	1.46E-129
c102346_g1	collagen, type XII, alpha	2.6972	2.08E-38	3.13E-35
c85556_g1	collagen, type XII, alpha	-2.5968	3.22E-88	9.67E-8
c97371_g1	collagen, type XIII	2.0647	6.44E-133	2.55E-129
c87691_g3	collagen, type XXII, alpha	1.4624	3.15E-07	6.17E-05
Heat shock protein				
c100263_g1	heat shock protein beta-1	-1.0522	1.31E-10	4.15E-08
Glycolysis / Gluconeogenesis				
c97925_g1	phosphoglycerate kinase	1.1989	2.06E-10	6.34E-08
c104780_g1	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	1.0596	1.12E-09	3.15E-07
HIF-1 signaling pathway				
c94892_g1	angiopoietin 4	1.6246	4.82E-06	0.0007421
FoxO signaling pathway				
c109496_g1	F-box protein 25/32	-2.108	4.68E-39	7.16E-36

Analysis of genes related to growth. Based on gene annotation using the Nr and Nt database, we identified 69 DEGs related to growth changes in the transcriptome of sandworms (**Table 4**). These DEGs included growth regulatory factors and their receptors or activators such as transforming growth factor-beta (TGF-β), thrombospondin-1 (TSP-1), thrombospondin-4 (TSP-4), neurogenic locus notch homolog protein 1 (Notch1), neurogenic locus notch homolog protein 2-like (Notch2), and tenascin-W (TN-W), cytoskeleton and myofibril component genes such as skeletal organic matrix protein, myosin-2 heavy chain (MYH2), actin beta/gamma 1 (ACTB/G-1), matrilin-2 (Matn2), filamin (FLN), fibrillin-1-like (FBN-1), fibrillin-2-like isoform X1 (FBN-2), fibrillin-3-like isoform X3 (FBN-3), and several other growth-related genes such as collagens I/II/IV/V/VI/VII/XII/XIII/XXII (COL), heat shock protein beta-1 (HSPB1), angiopoietin 4 (ANGPT4), F-box protein 25/32 (FBOX 25/32), phosphoglycerate kinase (PGK), and 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (dPGM).

Validation of RNA-Seq results of RT-PCR. To validate the RNA-Seq results, six DEGs (Unigens c88887_g1, Unigens c93702_g1, Unigens c74179_g1, Unigens c100263_g1, Unigens c89314_g4, and Unigens c86739_g1) in the sandworms were selected for an RT-PCR assay. As shown in **Figure 6**, Unigens c88887_g1 (annotated as calmodulin, CaM), Unigens c93702_g1 (annotated as filamin, FLNA) and Unigens c74179_g1 (annotated as hypotaurocyamine kinase, HTK) were upregulated, whereas Unigens c100263_g1 (annotated as heat shock protein beta-1, HSPB1), Unigens c89314_g4 (annotated as bZIP Maf transcription factor, MAF), and Unigens c86739_g1 (annotated as Fumarate reductase subunit D, FRD) were downregulated in the large size body wall muscle tissues at 180 dph, which showed that the expression profiles of RT-PCR were consistent with the results of the RNA-Seq analysis, indicating the reliability of RNA-Seq data in the present study.

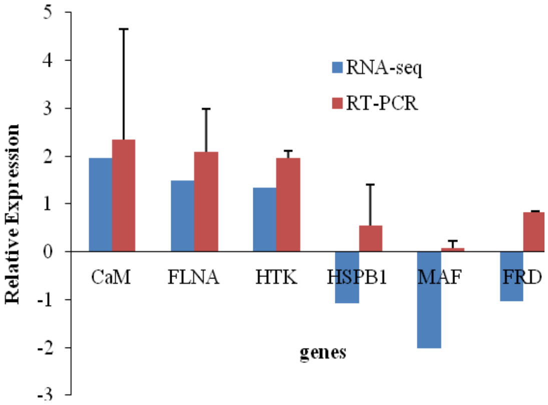

Figure 6 Comparison of genes expression levels between RNA-Seq and qRT-PCR results in sandworm body wall. The gene expression values were normalized to the EF1-α1 gene. Log2FC indicates the log2 fold change.

Discussion

S. nudus, an unsegmented worm-like animal, is composed of two sections, namely the trunk (main body wall) and an introvert (extending and contracting neck-like "feeler"), which contains internal organs and coelomic fluid (Ge et al., 2016; Liu and Qiu 2016). In practice, the latter two parts of sandworms are removed and the body wall is used in medicines and food. Therefore, as it is a farmed species, the body size is closely related to the profits of enterprises and farmers. However, studies on the breeding of sandworm mainly focus on germplasm resources (Zhong et al., 2018). To improve the growth properties of sandworms, our laboratory took the lead in the selection of growth-related traits of the sandworm by using the half-sib family method. In hybrid or cultured populations of sandworms, the phenomenon of individual growth variation is often found. To understand the molecular mechanism, the body wall muscle transcriptome between large and small-size worms at 180 dph was performed by Illumina sequencing.

After processing raw reads, a total of 185,181 unigenes were obtained in this study. Among these, about 88,357 (47.72%) unigenes had no BLAST hits indicating that they might contain novel genes with unknown functions. The same phenomenon also exists in the transcriptome annotation of other species, for instance, in sea cucumber *Apostichopus japonicus* and crab *Portunus trituberculatus* transcriptome up to 76.26% and 64.77% of the unigenes, respectively, failed to match known protein databases (Gao et al., 2017). This is mainly because the genome sequences of these species have not been obtained yet. Among the annotated transcripts, 418 DEGs were obtained from comparative transcriptome analysis. The results of the transcriptome reliability verification showed that the expression pattern of six DEGs obtained by RT-PCR was consistent with that of the transcriptome, indicating the data of the RNA-Seq is reliable in the present study and can be used for further analysis.

Annotation and comparative analysis of DEGs in sandworms. The body wall muscle of the sandworm is mainly composed of cuticle, epithelium, circular muscle, longitudinal muscle, and coelomic membrane. Muscle development is a complicated physiological process, which needs a large number of genes involved in cell regeneration, differentiation, and migration. In this study, a total of 418 DEGs were identified between small-size and large size worms at 180 dph (Corrected *p*-value<0.005), of which 207 genes were upregulated and 211 genes were downregulated in the large-size relative to small-size groups (**Figure 4**). Based on KEGG classifications, unigenes were mainly enriched in the cell junction pathway and related pathways, such as "Focal adhesion" and "ECM-receptor interaction". Focal adhesion is a large, dynamic protein complex that provides structural connections between the cytoskeleton and ECM, and is part of the signaling transduction process that controls growth (Sastry and Burridge 2000). ECM consists of a complex mixture of structural and functional macromolecules including glycosaminoglycans (GAGs) and fibrous proteins (collagen, elastin, fibronectin, and lamin) and plays a significant role in regulating many cellular behaviors, such as cell shape, adhesion, migration, proliferation, polarity, differentiation, and apoptosis (Wang et al., 2011). This is consistent with the results of the sea cucumber with variation in individual growth, which found that "Focal adhesion" and "ECM-receptor interaction" are both enriched KEGG pathways (Gao et al., 2017). Our results suggested that cell junction and related pathways may play an essential role in the process of sandworm individual growth variation. Furthermore, our study showed that protein adsorption and digestion were also the most significant pathways between small-size and large-size sandworms at 180 dph. This could be useful information for preparing suitable artificial cultivation of sandworms on the beach due to limited research on the nutritional requirements of this species.

It is noteworthy that there were six types of collagen, COLV, COLIVA, COLVIA, COLIAS, COLXXIIA, and COLXIIA, that were identified in the above three pathways (**Table 3**). Collagens are ECM molecules used by cells for structural integrity and a variety of functions (Gordon and Hahn 2010), such as tissue regeneration, metabolism, and the ability to interact with other substrates and cells. As is well known, COLI is the most abundant structural protein in most tissues and organs. It can provide mechanical support for organisms, maintain the integrity of organs and tissues, and guarantee their normal function. Recent studies have shown that COLI also plays an important role in cell differentiation and development, as a signal molecule to determine cell shape and behavior. For example, COLI can interact specifically with reactive astrocytes and induce astrocytic scar formation (Hara et al., 2017), which indicates that COLI may be a key factor in regulating the growth of sandworms. Furthermore, COLIVA, COLV, COLXIA, COLXIIA, and COLXXIIA were also enriched in "Focal adhesion", "ECM-receptor interaction", and "protein adsorption and digestion" pathways. It was reported that COLIVA, COLVIA, and COLXIIA have a crucial part in the function of muscle since mutations cause myopathy and muscular dystrophy (Hicks et al. 2013). COLXXIIA is found in the basement membrane of the myotendinous junction and associated with cartilage microfibrils (Koch et al., 2004). However, COLV exhibits different functions, such as cell adhesion and ECM repair process (Godoy et al., 2015), and repressed the attachment, spread, and growth of smooth muscle due to it being hidden between COLI and COLIII and together comprise heterotypic fibers (Sakata et al., 1992).

Differential expression of growth-related genes. The GH/IGF axis, known to regulate vertebrate body growth and cellular proliferation, is lacking in invertebrates. However, growth-related genes are considered to be the primary reason for genetic variation in animal growth, which are also abundant in invertebrates (Gao et al., 2017). In invertebrates and vertebrates, the genes that affect body size commonly exert their effect by altering the production of growth factors, or by altering the cellular response to growth regulators. In this study, a total of 49 unigenes were found to be related to the growth and development of sandworms, of which 44 unigenes were upregulated in the large-size group. Among these unigenes, 12 DEGs annotated with dorso-ventral axis formation, TGF-beta signaling pathway, osteoclast differentiation, glycolysis/gluconeogenesis, HIF-1 signaling pathway, VEGF signaling pathway, and FoxO signaling pathway were identified, including thrombospondin-1 (TSP-1), thrombospondin-4 (TSP-4), transforming growth factor-β1 (TGF-β1), angiopoietin 4 (ANGPT4), four and a half LIM domains protein 2 (FHL2), and heat shock protein beta-1 (HSPB1). Among these, TSP-4 and TGF-β1 were annotated in the TGF-β signaling pathway. TGF-β1, a pleiotropic cytokine, exerts potent and diverse effects on many different cell types and is involved in a wide variety of biological processes such as embryonic development, cell growth and differentiation, cell proliferation and survival, fibrosis, and regulation of the immune and inflammatory response (Dobaczewski et al., 2011). Although TGF-β1 is secreted as a latent complex and unable to associate with its receptors, TSP-1 can specifically bind to the sequence LSKL in the TGF-β latency-associated peptide and alter the conformation of TGF-β by making it accessible to its receptor to activate TGF-β signals (Frangogiannis et al., 2005). Moreover, angiopoietins are also an important family of growth factors that are closely related to angiogenesis and have four distinct types: ANGPT1, ANGPT2, ANGPT4, and ANGPT4. In this study, we found that ANGPT4 was upregulated 62 times in the large size groups. ANGPT4 was found as a ligand for tyrosine kinase with immunoglobulin and epidermal growth factor homology domains 2 (TIE2) to regulate later stages of vascular development, stabilization, and maturation. In addition, we noticed that the expression of FHL2 annotated to the protein synthesis pathway was upregulated by 1.5 times in this study. It is reported that FHL2, a LIM-only protein with four and a half LIM domains, is expressed in various tissues, where it may interact with several proteins to control cell proliferation and differentiation (Johannessen et al., 2006). Moreover, FHL2 interacts with titin and acts as an adaptor molecule that links the metabolic enzymes MM-creatine kinase, adenylate cyclase, and phosphofructokinase with titin, thereby helping to recruit metabolic enzymes needed for energy provision during muscle contraction.

It is worth noting that up to 38 DEGs involved in the regulation of cytoskeleton and myofibril components were identified in this transcriptome (**Table 4**), such as myosin, actin, matrilin, filamin, fibrillin, ankyrin, PDZ and LIM domain protein 5, tenascin, and different types of collagen. The role of these proteins in muscle growth has been confirmed by many studies. For example, two homodimeric myosins with different metabolic activities were characterized in the body wall of the earthworm *Lumbricus terrestris*, and the expression patterns of different types of myosin affect the growth of the body wall. As a multiadhesion adaptor protein, Matn2 can interact with other ECM proteins and promote neuromuscular junction formation and skeletal muscle regeneration (Korpos et al., 2015). The cytoskeletal protein FLNA can bind to actin through the N-terminal domain and thereby orthogonally cross-link actin filaments (Iwamoto et al., 2018). Fibrillins are the primary components of microfibrils in the ECM of many elastic and non-elastic connective tissues. The fibrillin family contains fibrillin-1 (FBN-1), fibrillin-2 (FBN-2), and fibrillin-3 (FBN-3), which are upregulated about 2.54, 2.35, and 2.32 times, respectively, in the large size group. FBN-1 and FBN-2 are known to exist in muscle and FBN-3 has been found in perimysium (Purslow 2017). Recent studies showed that FBN-2 mutants exhibit defects in tracheal smooth muscle cell alignment and polarity (Yin et al., 2019). Moreover, tenascin genes with Ca^{2+} binding epidermal growth factor (EGF) domain are upregulated in the large size groups. The Ca^{2+} binding EGF-like domain belongs to a subset of EGF-like domains, and Ca^{2+} plays a role in binding to EGF-like domains from the connective tissue protein fibrillin-1, which causes pleiotropic proliferative and developmental effects. The tenascin family with EGF repeats is also associated with ECM and acts both as integrin ligands and modifiers of fibronectin-integrin interactions to regulate cell adhesion, migration, proliferation, and differentiation (Adams et al., 2015). So far, six members of the tenascin family have been identified in various species from zebrafish to mammals: tenascin-C, tenascin-R, tenascin-X, tenascin-Y, and tenascin-W (Chiovaro et al. 2015). In this study, tenascin-W genes of the sandworm were upregulated in the large-size group. Studies showed that tenascin-W in the chicken and mouse was expressed at sites of osteogenesis and in subsets of smooth and skeletal muscle during development (Meloty-Kapella et al., 2006), which indicated that tenascin-W may affect muscle growth in the sandworm.

In conclusions transcriptome sequencing was performed using Illumina sequencing platform on material taken from body wall muscle tissues of sandworms from fast- and slow-growing groups. The analysis of the transcriptome data showed that fast-growing individuals exhibited high levels of cell junction, protein adsorption and digestion abilities. Furthermore, several growth-related genes (such as angiopoietin 4, myosin, actin, matrilin-2, filamin, fibrillin, thrombospondin 1, thrombospondin-4, and transforming growth factor beta) were identified in this study. These results will provide insight into the growth mechanism in sandworms, and be of further help in supporting the selective breeding of improved strains of sandworm.

Acknowledgements

We gratefully acknowledge Yan Jiang from Guangxi Institute of Oceanography regarding the breeding of sandworm and Tasev Ivaylo Valentinov from Beibu Gulf University for the grammar modification of this paper. We also express our gratitude to the reviewers and academic editor for their helpful suggestions and comments for improving our manuscript. This work was supported by Guangxi Natural Science Foundation [Grant No. 2018GXNSFAA281208 and 2016JJD130059], Science and Technology Program of Guangxi Province [Grant No. AB18221112], the Middle-aged and Young Teachers' Basic Ability Promotion Project of Guangxi [Grant No. 2019KY0436], Guangxi Key Laboratory of Beibu Gulf Marine Biodiversity Conservation [Grant No. 2018ZB11 and 2019ZB07], and High-Level Talents Scientific Research Start-Up Fund Project of Beibu Gulf University [Grant No. 2019KYQD17].

References

Adams J.C., Chiquet-Ehrismann R. and Tucker R.P, 2015. The evolution of tenascins and fibronectin. *Cell Adhes. Migr.,* 9: 22-33. 10.4161/19336918.2014.970030

Chiovaro F., Chiquet-Ehrismann R. and Chiquet M., 2015. Transcriptional regulation of tenascin genes. *Cell Adhes. Migr.,* 9: 34-47. 10.1080/19336918.2015.1008333

Dobaczewski M., Chen W. and Frangogiannis N.G., 2011. Transforming growth factor (TGF)-β signaling in cardiac remodeling. *J. Mol. Cell Cardiol.,* 51: 600-606. 10.1016/j.yjmcc.2010.10.033

Frangogiannis N.G., Ren G., Dewald O., Zymek P., Haudek S., Koerting A., Winkelmann K., Michael L.H., Lawler J. and Entman M.L., 2005. The critical role of endogenous Thrombospondin-1 in preventing expansion of healing myocardial infarcts. *Circulation,* 111: 2935-2942. 10.1161/CIRCULATIONAHA.104.510354

Gao L., He C.B., Bao X.B., Tian M.L. and Ma Z., 2017. Transcriptome analysis of the sea cucumber (*Apostichopus japonicus*) with variation in individual growth. *PloS One,* 12: e0181471. 10.1371/journal.pone.0181471

Ge Y.H., Tang Y.P., Guo S., Liu X., Zhu Z.H., Zhang L.L., Liu P., Ding S.X., Lin X.Z., Lin R.R. and Duan J.A., 2016. Simultaneous quantitation of free amino acids, nucleosides and nucleobases in *Sipunculus nudus* by ultra-high performance liquid chromatography with triple quadrupole mass spectrometry. *Molecules,* 21: 408-427. 10.3390/molecules21040408

Godoy C.A., Teodoro W.R., Velosa A.P.P., Garippo A.L., Eher E.M., Parra E.R., Sotto M.N. and Capelozzi V.L., 2015. Unusual remodeling of the hyalinization band in vulval lichen sclerosus by type V collagen and ECM 1 protein. *Clinics,* 70: 356-362. 10.6061/clinics/2015(05)09

Gordon M.K. and Hahn R.A., 2010. Collagens. *Cell Tissue Res.,* 339: 247-257.

Hao R.J., Du X.D., Yang C.Y., Deng Y.W., Zheng Z. and Wang Q.H., 2019. Integrated application of transcriptomics and metabolomics provides insights into unsynchronized growth in pearl oyster *Pinctada fucata martensii*. *Sci. Total Environ.,* 666: 46-56. 10.1016/j.scitotenv.2019.02.221

Hara M., Kobayakawa K., Ohkawa Y., Kumamaru H., Yokota K., Saito T., Kijima K., Yoshizaki S., Harimaya K., Nakashima Y. and Okada S., 2017. Interaction of reactive astrocytes with type I collagen induces astrocytic scar formation through the integrin N-cadherin pathway after spinal cord injury. *Nat. Med.,* 23: 818-828. https://doi.org/10.1038/nm.4354

Hicks D., Farsani G.T., Laval S., Collins J., Sarkozy A., Martoni E., Shah A., Zou Y.Q., Koch M., Bönnemann C., Roberts M., Lochmüller H., Bushby K. and Straub V., 2013. Mutations in the collagen XII gene define a new form of extracellular matrix-related myopathy. *Hum. Mol. Genet.,* 23: 2353-2363. 10.1093/hmg/ddt637

Iwamoto D.V., Huehn A., Simon B., Huet-Calderwood C., Baldassarre M., Sindelar C.V. and Calderwood D.A., 2018. Structural basis of the filamin A actin-binding domain interaction with F-actin. *Nat. Struct. Mol. Biol.,* 25: 918-927. 10.1038/s41594-018-0128-3

Johannessen M., Møller S., Hansen T., Moens U. and Van Ghelue M., 2006. The multifunctional roles of the four-and-a-half-LIM only protein FHL2. *Cell Mol. Life Sci.,* 63:

268-284. 10.1007/s00018-005-5438-z

Koch M., Schulze J., Hansen U., Ashwodt T., Keene D.R., Brunken W.J., Burgeson R.E., Bruckner P. and Bruckner-Tuderman L., 2004. A novel marker of tissue junctions, collagen XXII. *J. Biol. Chem.,* 279: 22514-22521. 10.1074/jbc.M400536200

Korpos É., Deák F. and Kiss I., 2015. Matrilin-2, an extracellular adaptor protein, is needed for the regeneration of muscle, nerve and other tissues. *Neural. Regen. Res.,* 10: 866-869. 10.4103/1673-5374.158332

Liang M., Dong S., Gao Q., Wang F. and Tian X., 2010. Individual variation in growth in sea cucumber *Apostichopus japonicus* (Selenck) housed individually. *J. Ocean U. China,* 9, 291-296. CNKI:SUN:QDHB.0.2010-03-015

Li J., Xie X., Zhu C., Guo Y. and Chen S., 2017. Edible peanut worm (*Sipunculus nudus*) in the Beibu Gulf: resource, aquaculture, ecological impact and counterplan. *J. Ocean U. China,* 16: 823-830. 10.1007/s11802-017-3310-z

Li J., Zhu C., Guo Y., Xie X., Huang G. and Chen S., 2015. Experimental study of bioturbation by *Sipunculus nudus* in a polyculture system. *Aquaculture,* 437: 175-181. 10.1016/j.aquaculture.2014.12.002

Liu Y. and Qiu C., 2016. Calculated taste activity values and umami equivalences explain why dried Sha-chong (*Sipunculus nudus*) is a valuable condiment. *J. Aquat. Food Prod. T.,* 25: 177-184. 10.1080/10498850.2013.839591

Meloty-Kapella C.V., Degen M., Chiquet-Ehrismann R. and Tucker R.P., 2006. Avian tenascin-W: expression in smooth muscle and bone, and effects on calvarial cell spreading and adhesion in vitro. *Dev. Dynam.,* 235: 1532-1542. 10.1002/dvdy.20731

Peng Y.H., Huang G.Q. and Liu X.J., 2015. Advances in Germplasm research and artificial culture of *Sipunculus nudus*. *J. Guangxi Acad. Sci.,* 31: 9-15.(in Chinese)

Podolak-Machowska A., Kostecka J., Librowski T., Santocki M., Bigaj J. and Plytycz B., 2014. Effects of anesthetic compounds on responses of earthworms to electrostimulation. *Folia biol.,* 62: 155-162. 10.3409/fb62_2.155

Purslow P. P., 2017. The structure and growth of muscle. *In Lawrie´s Meat Science. Woodhead Publishing,* 49-97. 10.1533/9781845691615.41

Sakata N., Jimi S., Takebayashi S. and Marques M.A., 1992. Type V collagen represses the attachment, spread, and growth of porcine vascular smooth muscle cells in vitro. *Exp. Mol. Pathol.,* 56: 20-36. 10.1016/0014-4800(92)90020-C

Sastry S.K. and Burridge K., 2000. Focal adhesions: a nexus for intracellular signaling and cytoskeletal dynamics. *Exp. Cell Res.,* 261: 25-36. 10.1006/excr.2000.5043

Su J., Jiang L.L., Wu J.N., Liu Z.Y. and Wu Y.P., 2016. Anti-tumor and anti-virus activity of polysaccharides extracted from *Sipunculus nudus* (SNP) on Hepg2.2.15. *Int. J. Biol. Macromol.,* 87: 597-602. 10.1016/j.ijbiomac.2016.03.022

Wang X.C., Gong Y., Wang D.J., Xie Q., Zheng M.Z., Zhou Y., Li Q., Yang Z., Tang H.L., Li Y.M., Hu R.M., Stamper B.D., Park S.S., Beyer R.P., Bammler T.K., Farin F.M., Mecham B. and Cunningham M.L., 2011. Differential expression of extracellular matrix-mediated pathways in single-suture craniosynostosis. *PloS One,* 6: e26557. 10.1371/journal.pone.0026557

Yang J.L., Zou J., Jiang Y. and Zhang Q., 2013, Demonstration and extension of artificial culture technology *Sipunculus nudus*. *J. Aquacult.,*19-21. (in Chinese)

Yin W.G., Kim H.T., Wang S.P., Gunawan F., Li R., Buettner C., Sengle G., Sinner D., Offernamms S. and Stainier D.Y., 2019. Fibrillin-2 is a key mediator of smooth muscle extracellular matrix homeostasis during mouse tracheal tubulogenesis. *Eur. Respir. J.,* 53: 1800840. 10.1183/13993003.00840-2018

Zhong S., Zhao Y., Zhang Q. and Chen X., 2018. The complete mitochondrial genome of the cryptic species in peanut worm *Sipunculus nudus* (Sipuncula, Sipunculidae) from Beibu Bay. *Mitochondrial DNA Part B,* 3(2): 484-485. 10.1080/23802359.2018.1463830