Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

De-novo characterization of the soft-shelled turtle *Pelodiscus sinensis* transcriptome using Illumina RNA-Seq technology*#

Wei WANG[1], Cai-yan LI[1], Chu-tian GE[1], Lei LEI[2], You-ling GAO[1], Guo-ying QIAN†‡[1]

([1]*Zhejiang Provincial Top Key Discipline of Modern Microbiology and Application, College of Biological and Environmental Sciences, Zhejiang Wanli University, Ningbo 315100, China*)

([2]*Henan Provincial Bureau of Animal Husbandry, Zhengzhou 450008, China*)

†E-mail: qiangy@zwu.edu.cn

Received Aug. 16, 2012; Revision accepted Nov. 9, 2012; Crosschecked Dec. 6, 2012

Abstract: The soft-shelled turtle *Pelodiscus sinensis* is a high-profile turtle species because of its nutritional and medicinal value in Asian countries. However, little is known about the genes that are involved in formation of their nutritional quality traits, especially the molecular mechanisms responsible for unsaturated fatty acid and collagen biosynthesis. In the present study, the transcriptomes from six tissues from *Pelodiscus sinensis* were sequenced using an Illumina paired-end sequencing platform. We obtained more than 47 million sequencing reads and 73 954 unigenes with an average size of 754 bp by de-novo assembly. In total, 55.19% of the unigenes (40 814) had significant similarity with proteins in the National Center of Biotechnology Information (NCBI) non-redundant protein database and Swiss-Prot database (*E*-value $<10^{-5}$). Of these annotated unigenes, 9 156 and 11 947 unigenes were assigned to 52 gene ontology categories (GO) and 25 clusters of orthologous groups (COG), respectively. In total, 26 496 (35.83%) unigenes were assigned to 242 pathways using the Kyoto Encyclopedia of Genes and Genomes pathway database (KEGG). In addition, we found a number of highly expressed genes involved in the regulation of *P. sinensis* unsaturated fatty acid biosynthesis and collagen formation, including desaturases, growth factors, transcription factors, and extracellular matrix components. Our data represent the most comprehensive sequence resource available for the Chinese soft-shelled turtle and could provide a basis for new research on this turtle as well as the molecular genetics and functional genomics of other terrapins. To our knowledge, we report for the first time, the large-scale RNA sequencing (RNA-Seq) of terrapin animals and would enrich the knowledge of turtles for future research.

Key words: *Pelodiscus sinensis*, Illumina RNA-Seq, Transcriptome, Gene expression
doi:10.1631/jzus.B1200219 **Document code:** A **CLC number:** Q781

1 Introduction

The soft-shelled turtle *Pelodiscus sinensis* is a commercially important aquatic reptile species because of its high nutritional and medicinal value in Asian countries including China, Japan, and Korea (Li *et al*., 2008; Gong *et al*., 2011). In China, it is considered to be a rich delicious food with medical benefits. In recent years, the farming of this species has developed rapidly in China with a yield of more than 265 721 tons in 2010 (Ministry of Agriculture and Fisheries Bureau of China, 2011). Compositional analysis has shown that this animal is rich in both

‡ Corresponding author

* Project supported by the National Basic Research Program (973) of China (No. 2011CB111500), the Natural Science Foundation of Ningbo City, China (Nos. 2011A610012 and 2011A610017), the Zhejiang Provincial Project of Selective Breeding of Aquatic New Varieties, China (No. 2012C12907), and the Zhejiang Provincial Top Key Discipline of Modern Microbiology and Application, China (Nos. KF2010005 and KF2012003)

Electronic supplementary materials: The online version of this article (doi:10.1631/jzus.B1200219) contains supplementary materials, which are available to authorized users

collagen and unsaturated fatty acids (Wang *et al*., 1997; Zhan *et al*., 2000; Huang *et al*., 2005). The collagen has multiple functions in Chinese traditional medicine. Compared with collagen from land animals (such as pig and cow), aquatic collagen has drawn extensive attention for its excellent physiochemical properties and physiologically active functions (Liu *et al*., 2010). A closer examination of the in vivo gene expression responsible for the development of these nutritional quality traits of *P. sinensis* is needed to obtain a complete understanding of the mechanisms. Unfortunately, very limited genomic information is available for *P. sinensis*. Currently, there are only about 353 nucleotides (nt) and 214 expressed sequence tag (EST) sequences available in the National Center of Biotechnology Information (NCBI) database for the soft-shelled turtle *P. sinensis*. Publicly available datasets are of use for future *P. sinensis* research, such as elucidating the molecular mechanisms of specific traits and gene expression regulation.

Compared with the conventional methods for gene cloning and sequencing, next-generation sequencing (NGS) technologies, such as the 454 (Roche), Supported Oligo Ligation Detetion (SOLiD) (ABI), and Solexa/Illumina (Illumina) platforms are not only time-saving and less expensive but also yield a great deal of genetic information about certain species. Over the past several years, NGS technologies have emerged as powerful tools for high-throughput sequence analysis and dramatically improved the speed and efficiency of gene discovery (Schuster 2008; Metzker, 2010). Moreover, these technologies have shown great potential for expanding sequence databases of not only model species (Hegedűs *et al*., 2009; Hillier *et al*., 2009; Li *et al*., 2010), but also non-model organisms (Collins *et al*., 2008; Wang *et al*., 2010a; Gao *et al*., 2012). Recently, the development of high-throughput DNA sequencing strategies has provided a new means of both mapping and quantifying transcriptomes, which is the complete set of all transcripts for certain types of cells or tissues in a specific developmental stage or physiological condition. Transcriptome analysis is a way to yield comprehensive functional elements of the genome and reveal the molecular mechanisms involved in specific biological processes on gene structure and function (Wei *et al*., 2011).

The method, termed RNA sequencing (RNA-Seq) is an efficient and powerful method to analyze transcriptome data and has clear advantages over the existing approaches (Wang *et al*., 2009; 2010a). A primary platform being used to generate transcriptomic resources is the Illumina short read sequencing platform, followed by de-novo assembly of sequence reads. This approach has been used to characterize transcriptomes of diverse taxa, including plants (Chang *et al*., 2011; Garg *et al*., 2011), insects (Wang *et al*., 2010b; Xue *et al*., 2010), mollusks (Feldmeyer *et al*., 2011), fish (Xiang *et al*., 2010), and mammals (Yao *et al*., 2012a; 2012b). Despite their obvious benefits, NGS methods have not yet been applied to turtle research. It is essential to produce enormous transcript sequences of *P. sinensis* for gene discovery and molecular marker development by using high-throughput transcriptome sequencing.

In the present study, we utilized Illumina pair-end sequencing technology to perform RNA-Seq on six tissues from *P. sinensis*, and demonstrated the suitability of pair-end for de-novo assembly and gene annotation of *P. sinensis* that does not have sequenced genomes. Over three billion bases (nt) of high-quality complementary DNA (cDNA) sequences and 73954 unique transcripts were generated. The annotated sequences for genes involved in gene ontology (GO) classifications, cluster of orthologous groups (COG) classifications, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were then assigned. In addition, we also analyzed the gene expression profiles of growth factors, transcriptional factors, extracellular matrix proteins that were related to unsaturated fatty acid biosynthesis and collagen peptide formation. Undoubtedly, these results provide an invaluable resource for further research in *P. sinensis*.

2 Materials and methods

2.1 Sample collection and RNA extraction

One male and one female soft-shelled turtle *Pelodiscus sinensis* were subject to transcriptomic analysis. Adult individuals of *P. sinensis* (Yellow River population, body weight of (500±50) g) were obtained from a turtle farm in Zhejiang, China. The Yellow River population of *P. sinensis* is a typical commercial population and widely used in the production of soft-shelled turtles. All experimental

procedures were approved by the Animal Ethics Committee of Zhejiang Wanli University. Tissues including the liver, muscle, spleen, ovary, testis, and calipash were dissected and cut into small pieces, and immediately frozen in liquid nitrogen for further use.

Total RNA was isolated using Trizol reagent (Invitrogen, CA, USA) according to the manufacturer's instructions. RNA integrity (Schroeder *et al*., 2006) was verified using the Agilent 2100 Bioanalyzer (Agilent Technologies) and all six samples had an RNA integrated number (RIN) value more than 8.0. The samples for transcriptome analysis were prepared using the Illumina kit according to the manufacturer's instructions. Briefly, messenger RNA (mRNA) was firstly purified from 10 μg of total RNA using oligo(dT) magnetic beads and then fragmented into small pieces using fragmentation buffer. The cleaved RNA fragments were used for reverse transcription followed by second-strand cDNA synthesis using DNA polymerase I and RNase H, after which QiaQuick polymerase chain reaction (PCR) kits (Qiagen, CA, USA) were used for end repair and adapter ligation. Equal quantities of high-quality RNA from the six tissues were pooled for cDNA library preparation.

2.2 cDNA library construction, sequencing, and assembly

The cDNA library was sequenced on the Illumina sequencing platform (Illumina HiSeq™ 2000, BGI, Shenzhen, China) using paired-end sequencing technology. The average size of inserts in the library was 200 bp, and both ends of the libraries were sequenced. The raw reads were cleaned by removing adaptor sequences, empty reads, and low-quality sequences. The data filtering process was previously described in Wang *et al*. (2010a) and Xie *et al*. (2012). Reads were then assembled using Trinity method (Grabherr *et al*., 2011). The longest assembled sequences are called contigs. Then the reads are mapped back to contigs. With paired-end reads we are able to detect contigs from the same transcript as well as the distances between these contigs. Next, contigs were connected to form scaffolds using "N" to represent unknown sequences. Finally, we arrive at sequences without Ns that cannot be extended on either end. Such sequences are defined as unigenes. Gene-expression-level analyses were performed by the reads per kilobase of the exon model per million reads (RPKM) method (Mortazavi *et al*., 2008) using the formula $\mathrm{RPKM}=10^9C/(NL)$, where C is the number of mappable reads that are uniquely aligned to a unigene, N is the total number of reads that are uniquely aligned to all unigenes, and L is the sum of a unigene in base pairs.

2.3 Functional annotation by sequence comparison with public databases

For annotation of assembled unigenes, a sequence similarity search was conducted against the NCBI non-redundant (Nr) protein database and the Swiss-Prot protein database using the BLASTx algorithm (Altschul *et al*., 1997; Cameron *et al*., 2004; Camacho *et al*., 2009) with a cut-off E-value of 10^{-5}. On the basis of the Nr annotation, the Blast2GO program (Conesa *et al*., 2005) was used to obtain GO annotation against the GO database (Harris *et al*., 2004) for unigenes annotated by Nr. Then the WEGO software (Ye *et al*., 2006) was used to perform GO functional classification for these unigenes. Annotation with the COG and KEGG pathways was performed using BLASTx against the COGs database (Tatusov *et al*., 2001) and the KEGG database (Kanehisa *et al*., 2004). If results of different databases conflicted with one another, a priority order of Nr, Swiss-Prot, KEGG, and COG was followed when deciding sequence direction of unigenes.

3 Results

3.1 Illumina sequencing, assembly, and sequence analyses

In order to obtain a wide range of genes associated with formation and development of nutritional quality traits, RNA was isolated from the liver, muscle, ovary, testis, and calipash of the Chinese soft-shelled turtle *Pelodiscus sinensis*. A mixed cDNA sample, representing diverse adult tissues of *P. sinensis* was prepared and sequenced using the Illumina platform for a single sequencing run. A total of 46655766 raw reads of 90 bp in length were generated. After stringent data cleaning and quality checks, we obtained more than 41 million high-quality reads with 96.98% bases had quality greater than 20. Then, a total of 178773 contigs were assembled. The length

of contigs varied from 100 to 7933 bp, with a mean size of 338 bp (Table 1). The proportion of contigs with length more than 200 bp was 38.69%. Finally, 73954 unigenes with an average size of 754 bp and a total length of 55.76M bp were yielded using the de-novo assembly (Table 1). The length of assembled unigenes ranged from 200 to 19616 bp. The distribution revealed that 49.77% of unigenes were between 200 and 400 bp, and unigenes with length more than 400 bp accounted for 50.23% including 21.29% unigenes larger than 1000 bp (Table 1).

The transcript levels of each unigene could be quantified in RPKM method. In Fig. 1, it shows that the expression of each unigene was correlated to sequencing depth. The expression of unigenes varied from 0 to 16390.72 RPKM with an average of 10.98 RPKM. Fifty-four thousand of 73954 (73.02%) unigenes had very low expression levels of less than 10 RPKM.

3.2 Functional annotation by searching against public databases

The results indicate that out of 73954 assembled unigenes, 35131 unigenes (47.5%) showed significant similarity to known proteins in Nr databases. Of all the unigenes, 33276 (45.0%) had BLAST hits in the Swiss-Prot database. Compared with Wang *et al*. (2010b), in which only 16.2% had BLAST hits in Nr database, the higher percentage in the present study was partially ascribed to the higher frequency of long sequences (Wang *et al*., 2010b). As reported by Parchman *et al*. (2010), longer contigs were more likely to have BLAST matches in the protein database.

3.3 Functional classification by GO and COG

GO is an international standardized gene functional classification system which has three ontologies: molecular function, cellular component, and

Table 1 Length distribution of assembled contigs and unigenes

Nucleotides length (bp)	Contigs	Unigenes
100–200	109627	0
201–300	24574	24020
301–400	12599	12787
401–500	6684	6996
501–600	4307	4609
601–700	3141	3444
701–800	2311	2565
801–900	1781	2063
901–1000	1511	1724
1001–1200	2363	2744
1201–1400	1884	2380
1401–1600	1488	1833
1601–1800	1221	1532
1801–2000	946	1215
2001–2200	773	1028
2201–2400	655	819
2401–2600	529	751
2601–2800	428	595
2801–3000	359	505
>3000	1590	2340
Total	178773	73954
Minimal length (bp)	100	200
Maximal length (bp)	7933	19616
N_{50} (bp)	629	1288
Average length (bp)	338	754
Total nucleotide length (bp)	60508133	55762804

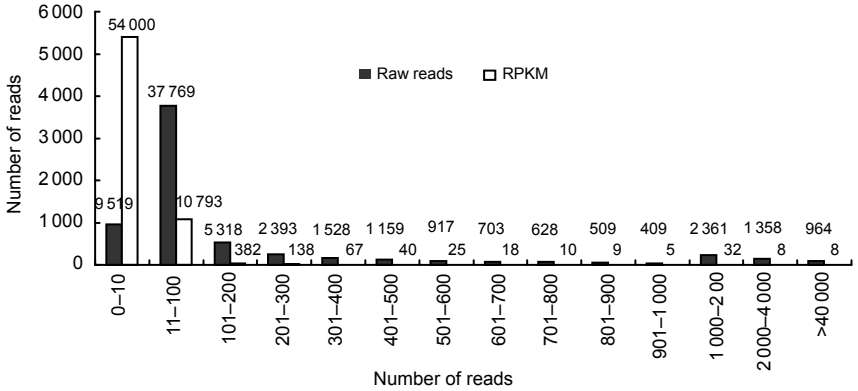

Fig. 1 Assessment of assembly quality
Distribution of unique-mapped reads and RPKM (reads per kb per million reads) of the assembled unigenes

biological process (Ashburner *et al*., 2000). Based on GO assignment, 9156 unigenes were categorized into 52 functional groups with 67893 terms (Fig. 2). The most abundant assignment was biological processes, (34318, 50.55%) followed by cellular components (22932, 33.78%), and molecular function (10643, 15.68%, Fig. 2). Under the biological process category, cellular processes (5492 unigenes, 16.00%), and metabolic processes (4241 unigenes, 12.36%) were predominant, indicating that some important cellular and metabolic activities may occur in *P. sinensis*. Moreover, 3343 and 3113 unigenes were assigned to the biological regulation (9.74%) and regulation of biological processes (9.07%), respectively. For the cellular component category, cell (7226 unigenes, 31.51%) and cell part (6842 unigenes, 29.84%) represented the majority of the category. Under the category of molecular function, binding (5061 unigenes, 47.55%) represented the most abundant classification, followed by catalytic activity (3002 unigenes, 28.21%).

The COG is a database that builts on coding proteins with complete genomes. Every protein in COG is thought to be orthologous. The purpose of COG is to serve as a platform for functional classification and annotation for many new sequences (Tatusov *et al*., 2001). In this study, out of 35131 Nr hits, a total of 11947 sequences were aligned to COG to predict and classify possible functions (Fig. 3). Among the 25 COG categories, the cluster for general function prediction only (4737, 14.56%) was the largest group, followed by translation, ribosomal structure, and biogenesis (3551, 10.91%) and replication, recombination, and repair (2938, 9.03%); only a few unigenes were assigned to nuclear structure and extracellular structures.

3.4 Functional classification by KEGG

The KEGG pathway database deposits the networks of molecular interactions in the cells and is widely used as a reference canonical database for integration and interpretation of large-scale dataset. Therefore, according to KEGG pathway mapping, gene functions with the emphasis on biochemical pathways can be categorized (Hou *et al*., 2011). Based on a comparison against the KEGG database, we assigned 26496 (35.83%) sequences to 242 KEGG pathways. The numbers of unigenes in different pathway ranged from 3 to 2674. Among them, the metabolic pathways (2674 unigenes) represented the

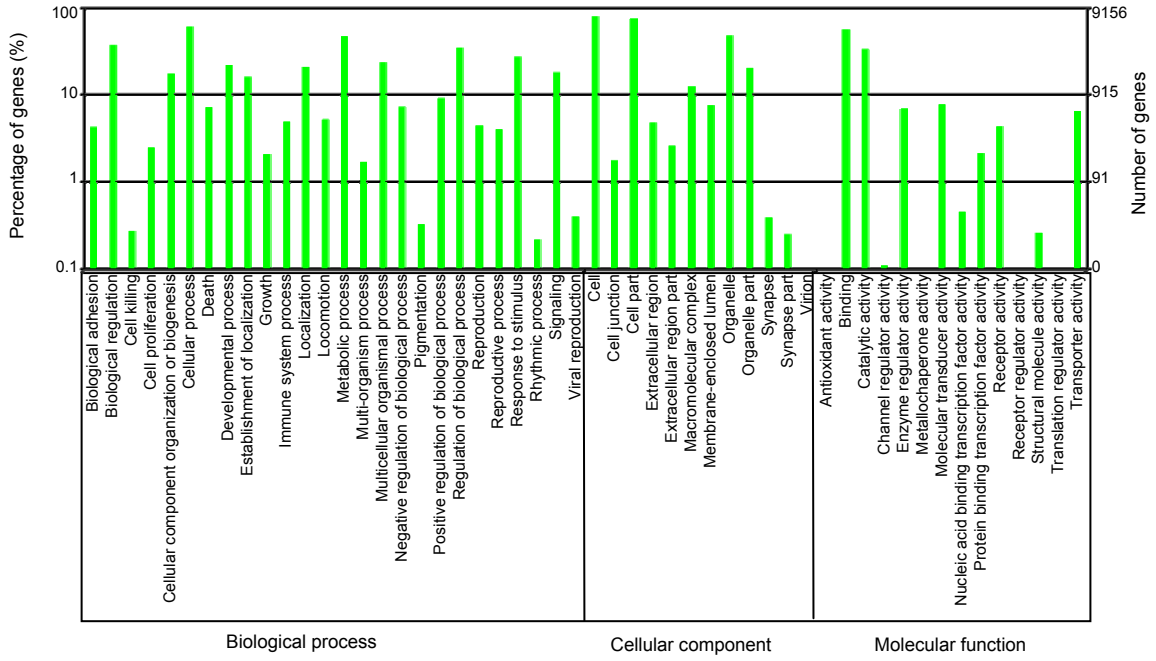

Fig. 2 Gene ontology classification of assembled unigenes

The results are summarized in three main categories: biological process, cellular component, and molecular function. The right *y*-axis indicates the number of genes in a category. The left *y*-axis indicates the percentage of a specific category of genes in that main category

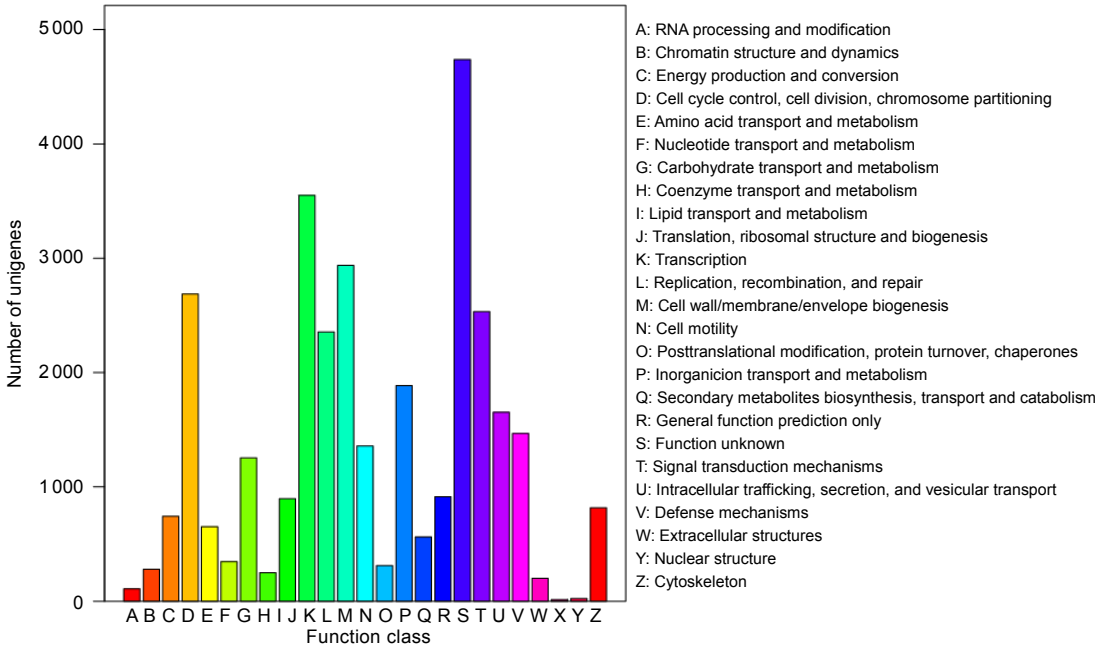

Fig. 3 Histogram presentation of clusters of orthologous groups (COG) classification

All unigenes were aligned to the COG database to predict and classify possible functions. Out of 35 131 Nr hits, 11 947 sequences were assigned to 25 COG classifications

most abundant group followed by those related to focal adhesion (1 627 unigenes) and regulation of actin cytoskeleton (1 348 unigenes) (Table S1).

3.5 Candidate genes involved in fatty acid biosynthesis

As shown in Table 2, a number of highly expressed genes related to unsaturated fatty acid biosynthesis of *P. sinensis* were identified. The genes included fatty acid binding protein, acyl Coenzyme A (CoA) desaturase, fatty acid desaturase, and elongation of very long chain fatty acids protein, etc. For example, the stearoyl CoA (D9) desaturase, one of the best studied desaturases to date, was highly expressed in the tissues of *P. sinensis*.

3.6 Candidate genes involved in collagen formation

A large number of highly expressed genes involved in collagen production, including transcription factors, growth factors, and extracellular matrix components of *P. sinensis* were also obtained. The results are listed in Table 3. It is shown that the most highly expressed transcripts were activating transcription factor-4 (ATF-4), followed by those encoding members of the collagens family and glycoproteins.

Table 2 Expression of genes involved in fatty acid biosynthesis

Gene name	RPKM[a]
Fatty acid binding protein 5 (FABP5)	2 321.44
Adipocyte fatty acid-binding protein (A-FABP)	254.06
Stearoyl-CoA desaturase (SCD)	152.37
Elongation of very long chain fatty acids protein 5 (ELO5)	55.31
Trans-2-enoyl-CoA reductase	48.89
Elongation of very long chain fatty acids protein 1 (ELO1)	43.24
Liver fatty acid-binding protein 1 (L-FABP1)	42.45
Enoyl-CoA hydratase	37.93
Heart fatty acid-binding protein (H-FABP)	34.76
Elongation of very long chain fatty acids protein 4 (ELO4)	17.03
Fatty acyl-CoA reductase	12.23
Fatty acid desaturase 6 (delta 6 FAD)	11.19
Fatty acid desaturase 1 (FAD1)	11.43
Acetyl-CoA carboxylase	10.92

[a] Reads per kilobase of exon model per million mapped reads

4 Discussion

With the rapid development of cost efficient and high throughput sequencing technologies, RNA-Seq

Table 3 Expression of genes involved in collagen formation [b]

Gene name	RPKM[a]
Transcription factor-4 (ATF-4)	341.38
Collagen alpha-1 (XVI) chain (Col 16α1)	154.31
Clathrin light chain A (Clta)	148.91
Fibronectin type 3 (Fn3)	90.44
Aggrecan core protein (Acan)	78.96
Collagen alpha-1 (III) chain (Col 3α1)	67.72
Collagen alpha-1 (VI) chain (Col 6α1)	59.51
Collagen alpha-3 (VI) chain (Col 6α3)	51.62
Connective tissue growth factor (CTGF)	38.03
Tenascin X (Tnx)	33.23
Insulin-like growth factor 2 receptor	31.87
Collagen alpha-1 (I) chain (Col 1α1)	29.89
Transforming growth factor β (TGF-β)	25.20
Matrix gla protein	24.05
Collagen alpha-1 (XVIII) chain (Col 18α1)	24.47
Collagen alpha-2 (VI) chain (Col 6α2)	23.88
Collagen alpha-3 (IX) chain (Col 9α3)	19.70
Collagen alpha-2 (I) chain (Col 1α2)	15.19
Type IV alpha 6 collagen (Col 4α6)	15.01
Collagen IV alpha 4 chain (Col 4α4)	13.97
Collagen alpha-1 (IV) chain (Col 4α1)	12.01
Collagen alpha-2 (VIII) chain (Col 8α2)	11.36
Collagen, type XVII, alpha 1 (Col 17α1)	10.70

[a] Reads per kilobase of exon model per million mapped reads. [b] Extracellular matrix proteins related

has emerged as a fast, powerful, and effective approach for novel gene discovering and gene expression profiling, especially in non-model organisms without prior genomic information (Yao *et al*., 2012b). In this study, we performed RNA-Seq by Illumina platform to characterize the transcriptome of *P. sinensis*, a species for which genomic and transcriptomic data is very limited in the public databases. In total, more than 41 million high-quality reads with 3.7G bp sequence coverage were obtained. By de-novo assembly, 73954 unigenes (≥200 bp) were generated, and further, 40814 assembled unigenes were annotated. Our coverage is approximately 130-fold more than all *P. sinensis* sequences deposited in GenBank combined (as of April 2012). To our knowledge, this study reports the first characterization of the complete transcriptome of terrapin animals by analyzing large-scale transcript sequences using an Illumina paired-end sequencing strategy.

In the present study, a number of the unigenes were assigned to GO functional groups and COG classifications (Figs. 2 and 3). Most representative unigenes were mapped to specific KEGG pathways, such as metabolism pathways, focal adhesion, and regulation of actin cytoskeleton. using the KEGG database (Table S1). Considering that *P. sinensis* is an important aquatic animal rich in nutrients and also has a great application potential in health-care products and anti-cancer drug development, pathways pertaining to poly unsaturated fatty acid biosynthesis and collagen formation were focused on. Quite a number of genes were enriched to these promising pathways, including fatty acid metabolism (89), biosynthesis of unsaturated fatty acid (64), fatty acid biosynthesis (38), and fatty acid elongation (16).

Based on de-novo sequencing and analysis of the transcriptome, we found several transcription factors related to fatty acid biosynthesis that were highly expressed. The fatty acid binding protein 5 (FABP5), which was the most highly expressed transcript, also known as EFABP and KFABP, belongs to the family of fatty acid-binding proteins (FABPs), a group of small intracellular 14–15 kDa cytoplasmic proteins that bind and transport long-chain free fatty acids. FABP5 was originally identified in the skin (Watanabe *et al*., 1994) and is expressed in different tissues and organs including the skin, liver, and the nervous system (Storch and McDermott, 2009). Adipocyte fatty acid-binding protein (AFABP) is one of the members of the FABP family and is mainly expressed in adipocytes (Storch and McDermott, 2009). Regarding the functions of KFABP and AFABP, Cao *et al*. (2008) reported that compared to the wild-type mice, lipid profiles from adipose tissue and plasma of double KFABP/AFABP knockout mice had elevated levels of the monounsaturated fatty acid (MUFA) palmitoleate (16:1D9). The stearoyl CoA (D9) desaturase was also enriched in the transcriptome. This enzyme catalyzes the first step in the polyunsaturated fatty acid (PUFA) biosynthetic pathway, namely the incorporation of a double bond at carbon nine of stearic acid to generate oleic acid, and is one of the best studied desaturases to date (Pereira *et al*., 2003).

Previous studies on *P. sinensis* have mainly focused on the nutritional value, such as preparation and characterization of collagen for biomaterial applications (Nobuhiro *et al*., 2009; Liu *et al*., 2010; Lu *et al*., 2010). The molecular mechanisms behind collagen

formation remain unclear. In this study, several collagen formation-related genes were found to be highly expressed. ATF-4, the most highly expressed transcription factor, is expressed in proliferative and prehypertrophic growth plate chondrocytes and regulates chondrocyte proliferation and differentiation via up-regulation of Indian hedgehog expression (Wang *et al*., 2009). Transgenic studies have demonstrated ATF-4 to be involved in lens and skeletal development, fertility, and proliferation (Ameri and Harris, 2008). We also found that many of the extracellular matrix components were highly expressed (Table 2), including fibronectin, tenascin, clathrin, and collagen. Among these proteins, members of collagen types I, III, IV, VI, VIII, IX, XVI, XVII, and XVIII were highly expressed. The collagens represent a large and complex family of structurally diverse extracellular matrix molecules. Collagen XVI, which was the most highly expressed transcript, is a member of the fibril-associated collagens with interrupted triple helices (FACIT-collagens) (Myers *et al*., 1994). Of potential physiological relevance, collagen XVI and collagen-binding integrin interactions may connect cells with specialized fibrils, contributing to the organization of fibrillar and cellular components within connective tissues (Eble *et al*., 2006). Collagen types I and III represent the major fibrillar collagen types in skin, and are possibly regulated by transforming growth factor (TGF)-β1, which is known to be the prototype of a large super family of proteins that control various aspects of differentiation (Heine *et al*., 1990). Type I collagen is the most plentiful and well-studied member of the collagen family and is one of the most widely expressed protein in the body being a major constituent of the skin, ligaments, tendons, bone, and numerous interstitial connective tissues (Bhogal *et al*., 2005). Lu *et al*. (2010) obtained the collagen from calipash of *P. sinensis* by pepsin digestion and reported that the extracted collagen belonged to typical collagen I with two α and one β chains. Whereas Liu *et al*. (2010) found that the collagen type of *P. sinensis* belongs to collagen V with an approximate molecular weight of 400–410 kDa, we did not find high expression of collagen V. Further research remains to be done to explain the absence of collagen V.

5 Conclusions

By using Illumina RNA-Seq techonology and de-novo analysis, we have generated more than 70000 unigenes with an average of 754 bp in length, of which 40814 sequences had a significant BLAST hit. We assigned 9156 sequences to 52 GO functional groups, 11947 sequences to 25 COG classifications and 26496 sequences to 242 KEGG pathways. A large number of transcript sequences obtained in this study were the first representatives of these transcripts for Chinese soft-shelled turtle *P. sinensis*. Many candidate genes potential involved in fatty acid biosynthesis and collagen formation were identified and are worthy of further investigation. These findings provide a substantial contribution to the existing sequences resources for the *P. sinensis* and other terrapin animals.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**(17): 3389-3402. [doi:10.1093/nar/25.17.3389]

Ameri, K., Harris, A.L., 2008. Activating transcription factor 4. *Int. J. Biochem. Cell Biol*., **40**(1):14-21. [doi:10.1016/j.biocel.2007.01.020]

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., 2000. Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**(1):25-29. [doi:10.1038/75556]

Bhogal, R.K., Stoica, C.M., McGaha, T.L., Bona, C.A., 2005. Molecular aspects of regulation of collagen gene expression in fibrosis. *J. Clin. Immunol*., **25**(6):592-603. [doi:10.1007/s10875-005-7827-3]

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST$^+$: architecture and applications. *BMC Bioinf*., **10**(1):421. [doi:10.1186/1471-2105-10-421]

Cameron, M., Williams, H.E., Cannane, A., 2004. Improved gapped alignment in BLAST. *IEEE/ACM Trans. Comput. Biol. Bioinf*., **1**(3):116-129. [doi:10.1109/TCBB.2004.32]

Cao, H., Gerhold, K., Mayers, J.R., Wiest, M.M., Watkins, S.M., Hotamisligil, G.S., 2008. Identification of a lipokine, a lipid hormone linking adipose tissue to systemic metabolism. *Cell*, **134**(6):933-944. [doi:10.1016/j.cell.2008.07.048]

Chang, L., Chen, J.J., Xiao, Y.M., Xia, Y.P., 2011. De novo characterization of Lycoris sprengeri transcriptome using

Illumina GA II. *Afr. J. Biotechnol.*, **10**(57):12147-12155. [doi:10.5897/AJB11.1761]

Collins, L.J., Biggs, P.J., Voelckel, C., Joly, S., 2008. An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Inf.*, **21**:3-14. [doi:10.1142/9781848163324_0001]

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**(18):3674-3676. [doi:10.1093/bioinformatics/bti610]

Eble, J.A., Kassner, A., Niland, S., Mörgelin, M., Grifka, J., Grässel, S., 2006. Collagen XVI harbors an integrin α1β1 recognition site in its C-terminal domains. *J. Biol. Chem.*, **281**(35):25745-25756. [doi:10.1074/jbc.M509942200]

Feldmeyer, B., Wheat, C.H., Krezdorn, N., Rotter, B., Pfenninger, M., 2011. Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. *BMC Genomics*, **12**(1):317. [doi:10.1186/1471-2164-12-317]

Gao, X.G., Han, J.B., Lu, Z.C., Li, Y.F., He, C.B., 2012. Characterization of the spotted seal Phoca largha transcriptome using Illumina paired-end sequencing and development of SSR markers. *Comp. Biochem. Physiol. Part D: Genom. Proteom.*, **7**(3):277-284. [doi:10.1016/j.cbd.2012.05.001]

Garg, R., Patel, R.K., Tyagi, A.K., Jain, M., 2011. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.*, **18**(1):53-63. [doi:10.1093/dnares/dsq028]

Gong, X., Niu, C.J., Zhang, Z.B., 2011. cDNA cloning and tissue expression for L-gulonolactone oxidase gene in soft-shelled turtle Pelodiscus sinensis a species with the ability to synthesize ascorbic acid. *Fish Sci.*, **77**(4):547-555. [doi:10.1007/s12562-011-0370-7]

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.D., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotech.*, **29**(7):644-652. [doi:10.1038/nbt.1883]

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., *et al.*, 2004. The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**:D258-D261. [doi:10.1093/nar/gkh036]

Hegedűs, Z., Zakrzewska, A., Ágoston, V.C., Ordas, A., Ŕacz, P., Mink, M., Spaink, H.P., Meijer, A.H., 2009. Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Mol. Immunol.*, **46**(15):2918-2930. [doi:10.1016/j.molimm.2009.07.002]

Heine, U.I., Munoz, E.F., Flanders, K.C., Roberts, A.B., Sporn, M.B., 1990. Colocalization of TGF-β1 and collagen I and III, fibronectin and glycosaminoglycans during lung branching morphogenesis. *Development*, **109**(1):29-36.

Hillier, L.W., Reinke, V., Green, P., Hirst, M., Marra, M.A., Waterson, R.H., 2009. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res.*, **19**(4):657-666. [doi:10.1101/gr.088112.108]

Hou, R., Bao, Z.M., Wang, S., Su, H.L., Li, Y., Du, H., Hu, J., Wang, S., Hu, X., 2011. Transcriptome sequencing and de novo analysis for Yesso Scallop (*Patinopecten yessoensis*) using 454 GS FLX. *PLoS One*, **6**(6):e21560. [doi:10.1371/journal.pone.0021560]

Huang, C.H., Lin, W.Y., Chu, J.H., 2005. Dietary lipid level influences fatty acid profiles, tissue composition, and lipid peroxidation of soft-shelled turtle, *Pelodiscus sinensis*. *Comp. Biochem. Physiol. A: Mol. Integr. Physiol.*, **142**(3):383-388. [doi:10.1016/j.cbpa.2005.09.004]

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**(S1):D277-D280. [doi:10.1093/nar/gkh063]

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., *et al.*, 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**(2):265-272. [doi:10.1101/gr.097261.109]

Li, X.L., Zhang, C.L., Fang, W.H., Lin, F.C., 2008. White-spot disease of Chinese soft-shelled turtles (*Trionyx sinensis*) caused by *Paecilomyces lilacinus*. *J. Zhejiang Univ. Sci. B*, **9**(7):578-581. [doi:10.1631/jzus.B0720009]

Liu, C.C., Liu, Y., Jin, Y.Z., 2010. Extraction and antioxidant activity of collagen from the Chinese soft-shelled turtle (*Pelodiscus sinensis*). *Adv. Mater. Res.*, **152-153**:1788-1792. [doi:10.4028/www.scientific.net/AMR.152-153.1788]

Lu, J.F., Wan, Q., Yin, Z.M., Lin, L., Weng, S.B., Ye, Y.W., Jiang, S.T., 2010. Extraction and characterization of collagen from calipash of Chinese soft-shelled turtle (*Pelodiscus sinensis*). *J. Fish China*, **34**:981-988 (in Chinese).

Metzker, M.L., 2010. Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**(1):31-46. [doi:10.1038/nrg2626]

Ministry of Agriculture and Fisheries Bureau of China, 2011. China Fishery Statistical Yearbook. China Agriculture Press, Beijing, China, p.41 (in Chinese).

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**(7):621-628. [doi:10.1038/nmeth.1226]

Myers, J.C., Yang, H.Y., D'Ippolito, J.A., Presente, A., Miller, M.K., Dion, A.S., 1994. The triple-helical region of human type XIX collagen consists of multiple collagenous subdomains and exhibits limited sequence homology to alpha 1(XVI). *J. Biol. Chem.*, **269**(28):18549-18557.

Nobuhiro, N., Hatsumi, K., Shizuka, K., Masanobu, M., 2009. Preparation and characterization of collagen from soft-shelled turtle (*Pelodiscus Sinensis*) skin for biomaterial applications. *J. Biomat. Sci., Polymer Ed.*, **20**(5-6):567-576. [doi:10.1163/156856209X426394]

Parchman, T.L., Geist, K.S., Grahnen, J.A., Benkman, C.W., Buerkle, C.A., 2010. Transcriptome sequencing in an

ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**(1):180. [doi:10.1186/1471-2164-11-180]

Pereira, S.L., Leonard, A.E., Mukerji, P., 2003. Recent advances in the study of fatty acid desaturases from animals and lower eukaryotes. *Prostaglandins Leukot. Essent. Fatty Acids*, **68**(2):97-106. [doi:10.1016/S0952-3278(02)00259-4]

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., Ragg, T., 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.*, **7**(1):3. [doi:10.1186/1471-2199-7-3]

Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**(1):16-18. [doi:10.1038/nmeth1156]

Storch, J., McDermott, L., 2009. Structural and functional analysis of fatty acid-binding proteins. *J. Lipid Res.*, **50**(Suppl.):S126-S131. [doi:10.1194/jlr.R800084-JLR200]

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E.V., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**(1):22-28. [doi:10.1093/nar/29.1.22]

Wang, D.Z., Tang, Z.R., Tan, Y.J., 1997. Biochemical compositions of Chinese soft-shelled turtle (*Trionyx sinensis*). I. Contents of normal nutrients and composition of muscle fatty acids. *Acta Hydrobiol. Sin.*, **21**:199-305 (in Chinese).

Wang, W., Lian, N., Li, L., Moss, H.E., Wang, W., Perrien, D.S., Elefteriou F., Yang, X., 2009. Atf4 regulates chondrocyte proliferation and differentiation during endochondral ossification by activating Ihh transcription. *Development*, **136**(24):4143-4153. [doi:10.1242/dev.043281]

Wang, X.W., Luan, J.B., Li, J.M., Bao, Y.Y., Zhang, C.X., Liu, S.S., 2010b. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*, **11**(1):400. [doi:10.1186/1471-2164-11-400]

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**(1):57-63. [doi:10.1038/nrg2484]

Wang, Z.Y., Fang, B.P., Chen, J.Y., Zhang, X.J., Luo, Z.X., Huang, L.F., Chen, X.L., 2010a. De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics*, **11**(1):726. [doi:10.1186/1471-2164-11-726]

Watanabe, R., Fujii, H., Odani, S., Sakakibara, J., Yamamoto, A., Ito, M., Ono, T., 1994. Molecular cloning of a cDNA encoding a novel fatty acid-binding protein from rat skin. *Biochem. Biophys. Res. Commun.*, **200**(1):253-259. [doi:10.1006/bbrc.1994.1442]

Wei, W.L., Qi, X.Q., Wang, L.H., Zhang, Y.X., Hua, W., Li, D.H., Lv, H.X., Zhang, X.R., 2011. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics*, **12**(1):451. [doi:10.1186/1471-2164-12-451]

Xiang, L.X., He, D., Dong, W.R., Zhang, Y.W., Shao, J.Z., 2010. Deep sequencing-based transcriptome profiling analysis of bacteria-challenged *Lateolabrax japonicus* reveals insight into the immune relevant genes in marine fish. *BMC Genomics*, **11**(1):472. [doi:10.1186/1471-2164-11-472]

Xie, F.L., Burklew, C.E., Yang, Y.F., Liu, M., Xiao, P., Zhang, B.H., Qiu, D.Y., 2012. De novo sequencing and a comprehensive analysis of purple sweet potato (*Impomoea batatas* L.) transcriptome. *Planta*, **236**(1):101-113. [doi:10.1007/s00425-012-1591-4]

Xue, J., Bao, Y.Y., Li, B.L., Cheng, Y.B., Peng, Z.Y., Liu, H., Xu, H.J., Zhu, Z.R., Lou, Y.G., Cheng, J.A., Zhang, C.X., 2010. Transcriptome analysis of the brown planthopper Nilaparvata lugens. *PloS One*, **5**(12):e14233. [doi:10.1371/journal.pone.0014233]

Yao, B.J., Zhao, Y., Wang, Q., Zhang, M., Liu, M.C., Liu, H.L., Li, J., 2012a. De novo characterization of the antler tip of Chinese Sika deer transcriptome and analysis of gene expression related to rapid growth. *Mol. Cell Biochem.*, **364**(1-2):93-100. [doi:10.1007/s11010-011-1209-3]

Yao, B.J., Zhao, Y., Zhang, H.S., Zhang, M., Liu, M.C., Liu, H.L., Li, J., 2012b. Sequencing and de novo analysis of the Chinese Sika deer antler-tip transcriptome during the ossification stage using Illumina RNA-Seq technology. *Biotechnol. Lett.*, **34**(5):813-822. [doi:10.1007/s10529-011-0841-z]

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R., Bolund, L., Wang, J., 2006. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.*, **34**:W293-W297. [doi:10.1093/nar/gkl031]

Zhan, X.A., Xu, Z.R., Qian, L.C., 2000. Muscle and fat quality of Chinese soft-shelled turtle. *J. Zhejiang Univ. (Agric. & Life Sci.)*, **26**(4):457-460 (in Chinese).

List of electronic supplementary materials

Table S1 KEGG pathways for *Pelodiscus sinensis*