Transcriptome Analysis of Tailfins From Grass-Goldfish And Egg-Goldfish to Identify Genes Involved in Artificial Selection for Ornamental Purposes

Shu-yan Wang (13074@ahu.edu.cn)
Anhui University https://orcid.org/0000-0002-1397-2314
Nai-yi Liu
Anhui university
Jie Fang
Anhui University School of Life Sciences

Research Article

Keywords: goldfish, high-throughput sequencing, morphological variation, gene expression difference

Posted Date: April 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-424631/v1

Abstract

Goldfish, one of the first animals domesticated for ornamental purposes, experience huge morphological variation, of which the number changes in tailfins and the loss of dorsal fins is the most impressive due to their important functions in fish. In the present study, a transcriptome analysis of the tailfins of Grass- (with complete dorsal fin and a single tailfin), and Egg-goldfish (loss of dorsal fin and double tail fins) was carried out to determine the sequence variants, to detect differential gene expression patterns in transcriptomes, and to characterize the effects of selection and identify target genes under selection. In total, 124,808,475 and 101,965,939 high quality reads were obtained from the tailfins of Grass- and Egg-goldfish, respectively, and 114,796 unigenes were further assembled from over 33.48 Gb nucleotides. A large portion of unigenes related to various primary and secondary metabolite pathways were identified, and the differentially expressed genes (DEGs) between the tailfins from Grass- and Egg-goldfish were also generally enriched in the metabolite pathways, for instant, the PI3K–Akt signalling pathway, the MAPK signalling pathway, osteoclast differentiation, and the dorso-ventral axis formation, all relating to cell proliferation, growth, and differentiation. These identified DEGs (*HOXA2b*, *HOXB13a*, paired mesoderm homeobox protein 1-like isoform X2 (*PRRX2*), zinc finger E-box-binding homeobox 1-like isoform X3 (*ZEB1*), and homeobox protein (*Meis1*), presumably played an important role in the development of the double tail fins during the artificial selection for ornamental purposes.

1. Introduction

Over the past 13,000 years of human history, domestication of all kinds of animals and plants has been fundamental. The successfully domesticated animals and plants not only provided for the development of human society in the necessary material foundation (e.g., clothing, food, and shelter), some domestic animals have also become the assistants and companions of human daily life (Diamond 2002). Compared with their wild ancestors, some domestic animals show huge differences in morphological characteristic such as goldfish and crucian carp (Smartt 2001; Wang 1985; Wang 2000). Understanding the genetic basis of phenotypic variation, especially identifying target genes under anthropogenic selection during domestication, can provide insight into the processes of rapid evolution and improvement (Diamond 2002).

Goldfish (*Carassius auratus*) belong to the Cypriniformes, Cyprinidae, and *Carassius* family, subfamily and genus, respectively. From unconscious breeding to the conscious anthropogenic selection for ornamental purposes, the morphological characteristics of the goldfish have experienced a huge variation compared with the wild crucian carp in less than 1, 000 years of being domesticated (Wang 1985). Long-term artificial selection for ornamental purposes makes the external morphology of goldfish (eye, fin, body shape, scales, etc.) considerably different from that of crucian carp, such as egg-shaped bodies, celestial or telescopic eyes, fancy tailfins, lionhead morphotypes, double tail fins, no dorsal fin, *inter alia* (Komiyama et al. 2009). Of all the morphological variations that have appeared during the domesticated history from wild crucian carp to goldfish, the number changes in the tailfins and the loss of dorsal fins has attracted general concern due to their important functions (Smartt 2001; Wang 2000). The biological function of the fish dorsal fin is to maintain the balance of the body while swimming and most fishes cannot stay upright without the dorsal fin (Drucker *et al*. 2001). The dorsal finlessness appeared after the attainment of double tail fins, which could compensate for the loss of dorsal fins (Wang et al. 2013). Depending on the condition of the dorsal (retained or loss), and tail (single or double) fins, two terms are used to designate the breeds: Grass-goldfish and Egg-goldfish (Wang 2000). The morphology of the Grass-goldfish (slender-shaped body with complete dorsal fin and a single tailfin; Fig. 1a) is less derived and more similar to the native *Carassius* than other breeds (Wang 1985; Wang 2000). The Egg-goldfish (Fig. 1b) always has an egg-shaped body and double tail fins, while their dorsal fins were lost in the process of artificial selection (Wang 1985; Wang 2000).

These morphological differences, caused by artificial selection for ornamental purposes, will leave traces on the genetic material (Diamond 2002). Analysis of the mitochondrial DNA sequences reveals more genetic and nucleotide diversity for Grass-goldfish comparing to Egg-goldfish, indicating that the Grass-goldfish could have been the first domesticated breed and the Egg-goldfish is likely to have a more recent origin (Komiyama et al. 2009; Wang et al. 2013). Further studies on the function of nuclear genes have shown that the homeobox genes play an important role in regulating the occurrence and development of fins. The expression of *HoxA11b* and *HoxA13b* genes in the growth and regeneration of pectoral and tail fins reveals that they are involved in the differentiation of fish osteoblasts (Shao *et al*. 2009). Additionally, research on the function of *HoxB1b* from *Megalobrama amblycephala* indicated it was closely related to the development of the pectoral fin on the anterior and posterior axes of the embryo (Pu et al. 2004). After introducing HoxD13 protein into zebrafish fins, the formation of new cartilage tissue and a decrease in fin tissue in zebrafish embryos were induced (Freitas et al. 2012). Research (Abe et al. 2014) on goldfish's chordin gene (osteogenic gene) revealed that a nonsense mutation in the chordin gene during the artificial domestication of goldfish might be the direct cause for the production of the double tailed goldfish (the shaft of the forking tail). Although goldfish have important biological significance in studying morphological differences caused by artificial selection for ornamental purposes, the overall transcriptomic or genomic analysis of goldfish of different breeds is still very limited. Over the last decade, benefiting from the progress in sequencing technology and the reduction of sequencing costs, a good opportunity presents itself to detect and compare the transcriptome sequences from different breeds of goldfish domesticated for ornamental purposes.

In the present study, a comprehensive analysis based on transcriptome sequencing from the tailfins of Grass-goldfish and Egg-goldfish was implemented to determine the sequence variants and detect differentially expressed genes (DEGs) involved in the development of double tail fins, and to identify the target genes artificial selected for ornamental purposes. In addition, the simple sequence repeats (SSR) loci and markers in the transcriptome from the tailfins were detected for further assistance in molecular marker-assisted breeding of goldfish. Through these extensive transcriptome analyses, we anticipate some progress in elucidating the regulatory effect of DEGs in tailfin variations caused by anthropogenic selection for ornamental purposes.

2. Material And Methods

2.1 RNA extraction, library construction and transcriptome sequencing

Live Grass- and Egg-goldfish specimens were collected from Beijing, Hangzhou, and Xuzhou, China. Total RNA isolation, the construction and sequencing of the cDNA library were carried out by the Map Biotechnology Co., Ltd. (Shanghai, China). Three individuals were selected from both Grass-goldfish and Egg-goldfish for total RNA extraction from the tailfin tissues of each sample using Trizol reagent (Invitrogen, Burlington, ON, Canada). Subsequently, the total isolated RNA was digested by DNase I (Takara, Shanghai, China), following the experimental procedures provided in the product manual. The mRNA was isolated from the total RNA using Dynabeads Oligo (dT) (Invitrogen, Burlington, ON, Canada), and afterwards broken into short fragments for the construction of the cDNA library.

Six mRNA-Seq Illumina libraries were constructed and sequenced through the HiSeq 2000 system (Illumina, San Diego, CA, USA) basing on the instructions from the manufacturer, and each tissue for sequencing contained three biological replicates. The raw sequence reads and assembled contigs were uploaded to the Sequence Read Archive (SRA) under the BioProject database of NCBI (http://www.ncbi.nlm.nih.gov) with the accession PRJNA666023.

2.2 De novo *assembly and annotation*

We abandoned the adapter sequences, reads with low quality (defined as reads containing > 50% bases with Q value ≤ 10), or more than 5% unknown ('N') nucleotides to obtain the cleaned reads from the raw data. Subsequently, Trinity software (version: trinityrnaseq-r2013-02-25; Grabherr et al. 2011) was used to generate the contigs by *de novo* assembly of high-quality cleaned reads, and the contigs were further assembled for the construction of unigenes. The total unigene number and length, as well as the average length, and the N50 length were calculated.

For the functional classification of the unigenes, the Blast2GO software (version 4.1.5, http://www.geneontology.org/) (Conesa et al. 2005) was used to establish the NR (NCBI non-redundant protein dataset), GO (Gene Ontology), COGs (Clusters of Orthologous Groups) and SWSS (SwissProt) annotations with a cut-off value of 1e-5the default parameters. We identified all the metabolic pathways that the unigenes might participate in to elucidate the metabolism and synthesis ability of the goldfish, based on the KEGG database (Kyoto Encyclopedia of Genes and Genomes Pathway, http://www.genome.jp/kegg/) (Kanehisa et al. 2004). The number of unigenes for each KEGG orthology (KO) was calculated and presented for further analysis.

2.2 Analyses of differential gene expression

In mRNA-Seq analysis, the gene expression level is evaluated by the number of cleaned reads located in specific regions of the genome. An RPKM (reads per kilobase of exon model per million mapped reads) value, a standard to measure gene expression level, was calculated for each unigene by the RSEM v1.2.11 program (Li and Dewey 2011). The unigene RPKM value was counted for each individual, and the RPKM distribution for all individuals were computed by using the gene density value and $\log_2$ (RPKM + 1) value.

We used the R package, DEseq2 (Anders and Huber 2010), to perform the differential expression analyses between Grass-goldfish and Egg-goldfish. Using negative binomial generalized linear models, DESeq2 provides methods to test for differential expression and to estimate the variance-mean dependence. We used the $\log_2$ values of normalized mean counts to represent fold changes of transcript expression levels, and the FDR (False Discovery Rate) value was calculated to adjust the P value. A significantly differential expression was distinguished by setting the threshold with the $\log_2$ (fold change) absolute value ≥ 2, and FDR value ≤ 0.05. For the visualization analyses of differential expression between Grass-goldfish and Egg-goldfish, a volcano plot was obtained by taking $\log_2$ (fold change) absolute value as the horizontal coordinate and the negative $\log_{10}$ (P value) as the ordinate.

Enrichment analyses of differentially expressed genes were respectively performed by using the software Goatools (https://github.com/tanghaibao/GOatools) for GO enrichment analysis (Klopfenstein et al. 2018) and KOBAS (http://kobas.cbi.pku.edu.cn/home.do) for KEGG pathway enrichment analysis (Wu et al. 2006; Xie et al. 2011), with a threshold of adjusted (FDR) P value ≤ 0.05.

Clustering analysis of the expression patterns for 2946 DEGs between Grass-goldfish and Egg-goldfish were implemented by using the R package heatmap2 (with Spearman's correlation coefficient for different individuals and Pearson correlation between unigenes) to infer the functional similarities between the DEGs.

2.3 Simple sequence repeats (SSRs) analysis

The MISA program (Beier et al. 2017) was used to identify the SSRs in the unigenes (length > 1000 bp) of goldfish, and to design the primers for the predicted SSRs. Furthermore, the numbers of the predicted SSRs (mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides) were calculated and revealed by a histogram.

3. Results

3.1 Transcriptome sequencing and de novo *assembly*

Three biological replicate libraries were generated and sequenced from the tailfins of Grass-goldfish and Egg-goldfish, respectively (Table 1), and all the raw reads were deposited in the SRA with accession number PRJNA666023. A total number of 54,123,840, 52,650,612 and 54,797,604 raw reads were respectively obtained from the three Grass-goldfish specimens. Three Egg-goldfish specimens separately produced 45,551,450, 49,700,170 and 50,184,718 raw reads. The adaptors, the reads with more than 5% N content, or the bases with lower quality (defined as reads containing > 50% bases with Q value ≤ 10) were removed. After quality control, approximately 18.40 and 15.08 Gb clean nucleotides in total were obtained for Grass-goldfish and Egg-goldfish, respectively.

Table 1
Summary statistics of original and assembled data from transcriptome of three Grass-goldfish and Egg-goldfish

	Grass-1	Grass-2	Grass-3	Egg-1	Egg-2	Egg-3
Raw sequencing reads						
Total reads	54,123,840	52,650,612	54,797,604	45,551,450	49,700,170	50,184,718
Total bases (bp)	8,118,576,000	7,897,591,800	8,219,640,600	6,832,717,500	7,455,025,500	7,527,707,700
Clean sequencing reads						
Total reads	48,160,910	47,175,288	29,472,277	41,018,144	31,511,232	29,436,563
Total bases (bp)	7,105,892,795	6,963,276,072	4,329,394,413	6,055,892,155	4,666,822,327	4,361,478,274
Percentage of clean reads (%)	88.98	89.60	53.78	90.04	63.40	58.65
Percentage of clean bases (%)	87.53	88.17	52.67	88.63	62.60	57.94
GC % of clean reads	42.41	43.36	42.57	41.96	43.28	45.16
Q30 (%)	95.54	95.43	94.97	95.30	95.64	95.85
Alignment information						
Aligned reads	37,065,166	36,090,133	22,552,540	32,203,590	25,137,076	23,154,977
Aligned ratio (%)	76.96	76.50	76.52	78.51	79.77	78.66
Unigene information						
Total number of unigenes	114,796					
Number of bases	89,545,010					
Mean length of unigene (bp)	780.04					
N50 length of unigene (bp)	861					

	Grass-1	Grass-2	Grass-3	Egg-1	Egg-2	Egg-3
Largest unigene (bp)	18,054					

Subsequently, 114,796 unigene sequences, with the total length of 89,545,010 bp, were *de novo* assembled from the clean reads of the combined six mRNA-Seq Illumina libraries using Trinity software (version: trinityrnaseq-r2013-02-25) (Conesa et al. 2005). The proportions of reads that could map back to the assembled transcript were 76.68% for Grass-goldfish, and 78.94% for Egg-goldfish. The overall GC contents of the transcriptomes were 42.78% and 43.47% in Grass-goldfish and Egg-goldfish, respectively (Table 1). Furthermore, the mean (780 bp), N50 (861 bp) and largest (18, 054 bp) length of unigenes were also calculated.

3.2 Functional annotation of unigenes

The assembled 114,796 unigenes were further annotated in the NR (ftp://ftp.ncbi.nih.gov/blast/db; Pruitt et al. 2012), GO (http://geneontology.org; Ashburner et al. 2000), COG (http:// https://www.ncbi.nlm.nih.gov/research/cog-project; Galperin et al. 2021), KEGG (http://www.genome.jp/kegg; Kanehisa et al. 2004) and SWSS databases (http://www.uniprot.org; Bairoch and Boeckmann 1991) (Table 2; Fig. 2). Overall, 49.45%, 1.33%, 32.41%, 25.02%, and 34.39% unigenes were identified and categorized in the NR, GO, COG, KEGG, and SWSS databases, respectively, and 1,096 unigenes annotated to all databases are shown at the intersections of all circles in Fig. 2.

Table 2
Summary of unigene annotation against public databases

Category	Number of annotated unigenes	Percentage of annotated unigenes (%)
Annotated in NR	56,761	49.45
Annotated in GO	1,530	1.33
Annotated in COG	37,204	32.41
Annotated in KEGG	28,723	25.02
Annotated in SWSS	39,476	34.39
Annotated in at least one database	57,584	50.16
Annotated in all database	1,096	0.95

Note. NR, NCBI non-redundant protein dataset (ftp://ftp.ncbi.nih.gov/blast/db; Pruitt et al., 2012); GO, Gene Ontology (http://geneontology.org; Ashburner et al., 2000); COG, Clusters of Orthologous Groups (https://www.ncbi.nlm.nih.gov/research/cog-project; Galperin et al., 2021); KEGG, Kyoto Encyclopedia of Genes and Genomes Pathway (http://www.genome.jp/kegg; Kanehisa et al., 2004); SWSS, SwissProt (http://www.uniprot.org; Bairoch and Boeckmann, 1991).

In the GO category, 1,530 unigenes were classified into three primary GO terms, “biological processes”, “cellular components”, and “molecular function”, and further subdivided into 56 secondary-class terms (Fig. 3). In the GO term “biological processes”, “cellular process (1045 genes)”, “single-organism process (829 genes)”, and “metabolic process (803 genes)”, were the top three, secondary-class terms with the most unigenes. The secondary categories “cell part (883 genes)”, “cell (883 genes)”, and “organelle (640 genes)”, were the top three secondary categories in the GO term “cellular components”. In the “molecular function” GO term, the secondary-class terms “binding (820 genes)”, and “catalytic activity (586 genes)” occupied the top two positions with the most unigenes.

In the COG annotation analyses, 37,204 unigenes were mapped into 26 categories, and the “U: intracellular trafficking, secretion, and vesicular transport (3784 genes)”, “O: Posttranslational modification, protein turnover, chaperones (2482 genes)”, “T: Signal transduction mechanisms (2099 genes)”, “K: Transcription (1950 genes)”, and “Z: Cytoskeleton (770 genes)” were substantially enriched (Fig. 4).

In the KEGG pathway annotation, 28,723 unigenes were classified into five categories mapping to 340 pathways (Fig. 5A), with 8237 genes located in “A: Cellular Process”, 3441 genes in “B: Environmental Information Processing”, 7200 genes in “C: Genetic Information Processing”, 5837 genes in “D: Metabolism”, and 10,180 genes in “E: Organismal Systems”. The top 20 enriched KEGG pathways are displayed by a histogram (Fig. 5B), and the top three pathways with the most unigenes enrichment were “metabolic pathways (2824 unigenes)”, “pathways in cancer (1245)”, and “PI3K – Akt signalling pathway (1147 unigenes)”.

3.3 Gene quantification in Grass-goldfish and Egg-goldfish

The number of mapped reads for each gene was calculated and standardized to RPKM (reads per kilobase of exon model per million mapped reads) to quantify the transcriptome expression level in the tailfins from Grass-goldfish and Egg-goldfish. In total, ~ 76.68% (Grass-goldfish), and 78.93% (Egg-goldfish) clean reads from each library were mapped to the prior-assembled 114,796 unigenes. According to the calculated RPKM value for each unigene, the expression levels of unigenes were subsequently defined as no transcription (RPKM < 1), low (1 ≤ RPKM < 5), medium (5 ≤ RPKM < 10), and high (RPKM ≥ 10). The number of genes with medium (14,651 vs. 14,464), and high expression (7398 vs. 7331) in Grass-goldfish and Egg-goldfish tailfins was similar, while the number of unexpressed (99,537 vs. 85,489), and low-expressed (60,474 vs. 45,147) genes in the tailfins of Grass-goldfish was much greater than in Egg-goldfish, suggesting that more genes function in the development of the double tail fin.

3.4 Analysis of the DEGs (differentially expressed genes)

To identify the DEGs between the Grass-goldfish and Egg-goldfish, a paired comparison of the expression quantity for each unigene in the tailfin of the different breeds was conducted and the unigenes with fold change ≥ 2 and FDR < 0.01 were filtered (Table 3). A total of 7516 DEGs were detected, out of which 2946 unigenes demonstrated substantially differential expression in the tailfins between Grass-goldfish and Egg-goldfish. In these differentially expressed genes, compared with Egg-goldfish, 1922 unigenes were up-regulated and 1024 unigenes were down-regulated in Grass-goldfish.

Table 3
Summary of DEG (differentially expressed gene) number between different pairs of Grass-goldfish and Egg-goldfish.

Comparison pairs	DEG number	Up-regulation	Down-regulation
Grass vs. Egg	2946	1922	1024
Grass1 vs. Gass2	1065	502	563
Grass1 vs. Grass3	1067	496	571
Grass2 vs. Grass3	1126	531	595
Egg1 vs. Egg2	422	232	190
Egg1 vs. Egg3	457	219	238
Egg2 vs. Egg3	431	190	241

The differentially expressed unigenes from tailfins of Grass-goldfish vs. Egg-goldfish were further mapped and classified in GO categories and KEGG pathways. The significantly up-regulated DEGs (240 unigenes) annotated in GO categories were mainly involved in the single organism process (GO: 0044699, 23 unigenes), and cellular process (GO: 0009987, 23 unigenes) of biological processes (121 unigenes), the cell part (GO: 0044464, 14 unigenes) and membrane part (GO: 0044425, 13 unigenes) of the cellular component (86 unigenes), and the binding (GO: 0005488, 16 unigenes) and catalytic activity (11 unigenes) of molecular function (33 unigenes). The significantly down-regulated DEGs (198 unigenes) were generally involved in the cellular process (GO: 0009987, 22 unigenes) and metabolic process (GO: 0008152, 20 unigenes) of the biological process (78 unigenes), the cell part (GO: 0044464, 20 unigenes) and organelle (GO: 0043226, 18 unigenes) of the cellular component (97 unigenes), and the binding (GO: 0005488, 11 unigenes) and catalytic activity (9 unigenes) of molecular function (23 unigenes). Furthermore, in the metabolic process category, the DEG analyses revealed that the *HOXA2b* (TRINITY_DN71748_c0_g1), *HOXB13a* (TRINITY_DN74023_c0_g1), paired mesoderm homeobox protein 1-like isoform X2 (*PRRX2*, TRINITY_DN89108_c0_g1), zinc finger E-box-binding homeobox 1-like isoform X3 (*ZEB1*, TRINITY_DN93814_c0_g1), and homeobox protein, *Meis1* (TRINITY_DN92032_c0_g1) were up-regulated, while the H2.0-like homeobox protein (*HLX*, TRINITY_DN86311_c0_g1) was down-regulated in the tailfin from the Grass-goldfish.

Large numbers of DEGs were also mapped to various primary and secondary metabolite pathways (e.g., ko01100, ko04151, ko04010, and ko04014) in the enrichment analyses of DEGs between the Grass- and Egg-goldfish KEGG pathways (Table S1 and S2), including the PI3K-Akt signalling pathway (36 unigenes), MAPK signalling pathway (17 unigenes), and Jak-STAT signalling pathway (13 unigenes). The osteoclast differentiation (ko04380) and dorso-ventral axis formation (ko04320), probably related to the formation of the goldfish tailfin, were also identified in the KEGG pathway enrichment analyses of DEGs. In the osteoclast differentiation (ko04380) pathway, our analyses demonstrated that the GRB2-associated-binding protein 2-like isoform X2 (TRINITY_DN94490_c0_g1), serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform X2 (TRINITY_DN90198_c0_g2), autophagy-related protein 9A-like (TRINITY_DN92800_c0_g2), and calcium/calmodulin-dependent protein kinase type IV-like (TRINITY_DN98117_c0_g1) were up-regulated (Table S1), while the interferon gamma receptor 1–2 (TRINITY_DN94125_c1_g2), fos-related antigen 2-like (TRINITY_DN85588_c0_g1), interleukin-1 beta-2 (TRINITY_DN104072_c0_g1), fos-related antigen 1-like isoform X2 (TRINITY_DN76364_c0_g1), retinoic acid receptor responder protein 3-like (TRINITY_DN16362_c0_g1), and suppressor of cytokine signalling 1-like (TRINITY_DN83317_c0_g1) were down-regulated (Table S2) in the tailfin from Grass-goldfish. In addition, the GTPase, HRas (TRINITY_DN90348_c2_g2), was down-regulated in the Grass-goldfish's tailfin (Table S2), within the dorso-ventral axis formation (ko04320) pathway.

3.5 Polymorphism analysis of simple sequence repeats (SSR) markers

A total of 15931 SSRs (simple sequence repeats) were identified including di- (12,511, 78.53%), tri- (2919, 18.32%), and tetra-nucleotide repeats (438, 2.75%), (Table 4, Fig. 6). Among the di-nucleotide repeats, AC/GT (8669), AG/CT (2909), and AT/AT (910) were the dominant motifs. Within the tri- nucleotide repeats, AAT/ATT (537), followed by AGG/CCT (470), and AAC/GTT (444), were the most repeated motifs (Table 4).

Table 4
The SSRs (simple sequence repeats) identified from the unigene sequences of Goldfish.

SSR parameters	Number
Total number of sequences examined	114,796
Total size of examined sequences	88,134,359
Total number of identified SSRs	15,931
Di-nucleotide	12511
Tri-nucleotide	2919
Tetra-nucleotide	438
Penta-nucleotide	49
Hexa-nucleotide	14
Number of SSRs containing sequences	24,594
Number of sequences containing more than 1 SSR	5,307
Number of SSRs present in compound formation	2,785

4. Discussion

Compared with the wild crucian carp, the variation in the tail and dorsal fins is the most attractive, owing to its important biological function in fish. In the present study, based on RNA-Seq technology, transcriptome analyses of tailfins from Grass-goldfish (single), and Egg-goldfish (double) were performed. After quality control, approximately 77.25% and 70.11%clean reads were obtained for Grass-goldfish and Egg-goldfish, respectively. The overall GC contents of the Grass- (42.78%) and Egg-goldfish (43.47%) transcriptomes were both slightly lower than that of AT, which was consistent with the 42.9% GC content in the mitochondrial genome of the wild crucian carp (Ge et al. 2020). Subsequently, 114,796 unigene sequences, with the total length of 89,545,010 bp, were *de novo* assembled from the clean reads, and the proportion of reads that could map back to the assembled transcript was 76.68% for Grass-goldfish, and 78.94% for Egg-goldfish. The average (780 bp), N50 (861 bp), and largest (18, 054 bp) length of unigenes were also calculated and ensured reliable transcriptome data.

The KEGG pathway analysis results indicated that the significantly enriched pathways were involved in four aspect including “Signal transduction”, “Global and overview maps”, “Endocrine system”, and “Transport and catabolism”. In the “Signal transduction” group, more than 73% unigenes were distributed in the “PI3K − Akt signalling pathway (1147 unigenes)”, “MAPK signalling pathway (857 unigenes)”, “Rap1 signalling pathway (809 unigenes)”, and “Ras signalling pathway (656 unigenes)”. Phosphatidylinositide-3-kinases (PI3K, an intracellular phosphatidylinositol kinase), and Akt (a serine/threonine specific protein kinase) play a key role in cell growth processes, such as glucose metabolism, apoptosis, cell proliferation, transcription and cell migration (Van de Sande et al. 2002;Voskas et al. 2014). The Ras-Raf-MAPK signal transduction pathway is involved in signal transduction of various growth factors, cytokines, mitogens and hormone receptors, and also plays an important role in cell proliferation, growth, and differentiation (Haston et al. 2017). These pathways probably contributed to the morphological variation in tailfins apparent in wild crucial carp and Grass- and Egg-goldfish.

For further understanding of the functions and metabolic pathways involved in the artificial selection of goldfish tailfins, the GO and KEGG pathway enrichment analyses of differentially expressed genes (DEGs) were performed. The GO

enrichments analyses revealed that the HOXA2b, HOXB13a, paired mesoderm homeobox protein 1-like isoform X2 (PRRX2), zinc finger E-box-binding homeobox 1-like isoform X3 (ZEB1,) and homeobox protein Meis1 were up-regulated, while the H2.0-like homeobox protein was down-regulated in the tailfin of Grass-goldfish vs. Egg-goldfish. Homeobox genes (*HOX* genes), an evolutionarily highly conserved family of polygenes, are a class of transcription factors containing homologous allomorphic domains, which are involved in controlling somatic features, regulating the central nervous system, and determining the relationship between the anterior and posterior somatic axis differentiation in vertebrate embryos (Moghadam et al. 2005; Luo et al. 2007). HoxA2b participates in the multicellular organism development through regulating the transcriptional process (Scemama et al. 2006). The previous study also revealed that the *HoxB13* and *PRRX2* genes in the human were expressed in patterns consistent with roles in cutaneous regeneration and foetal skin development (Stelnicki et al. 1998). Homeobox protein Meis1 is involved in the negative regulation of myeloid cell differentiation (Moskow et al. 1995), and the H2.0-like homeobox protein participates in positive regulation of cell population proliferation and organ growth. Furthermore, recent research (Yuan et al. 2010; Freitas et al. 2012; Paco and Freitas 2018) has identified that the *Hox* genes (e.g. *HoxA*, *HoxB*, and *HoxD*) play an important role in regulating the occurrence and development of fins. Large numbers of differentially expressed unigenes were found in various primary and secondary metabolite pathways (e.g., the PI3K-Akt signalling pathway, MAPK signalling pathway, Jak-STAT signalling pathway). In addition, osteoclast differentiation (ko04380), and dorso-ventral axis formation (ko04320) were also found in the enrichment analyses of DEGs in KEGG pathways. Osteoclasts (OC), derived from the monocyte/macrophage haematopoietic lineage, are the main functional cell in bone resorption and play an important role in bone development, growth, repair, and reconstruction (Boyle et al. 2003). A previous study (Abe et al. 2014) on goldfish also revealed that a nonsense mutation in the osteogenic gene might be a direct cause of the production of the double-tailed goldfish. The down-regulated GTPase, HRas, in dorso-ventral axis formation (ko04320) is also involved in regulating cell division in response to growth factor stimulation (Guil et al. 2003). These analyses indicate that the DEGs between the tailfins from Grass-goldfish and Egg-goldfish are generally found in metabolite pathways relating to cell proliferation, growth, and differentiation, and probably played an important role in the development of the double-tailed fin during artificial selection for ornamental purposes.

5. Conclusion

In total, 124,808,475 and 101,965,939 high quality reads were obtained from the tailfins of Grass- and Egg-goldfish, respectively, and 114,796 unigenes were further assembled from over 33.48 Gb nucleotides. A large portion of unigenes related to various primary and secondary metabolite pathways were identified, and the DEGs between the tailfins of Grass- and Egg-goldfish were also generally enriched in the metabolite pathways, for instant, the PI3K–Akt signalling pathway, the MAPK signalling pathway, osteoclast differentiation, dorso-ventral axis formation, all of which relate to cell proliferation, growth, and differentiation. The DEGs that were identified, including *HOXA2b*, *HOXB13a*, paired mesoderm homeobox protein 1-like isoform X2 (*PRRX2*), zinc finger E-box-binding homeobox 1-like isoform X3 (*ZEB1*,) and homeobox protein *Meis1*, probably played an important role in the development of the double-tailed fins during artificial selection for ornamental purposes. The results of the present study provide key information on the candidate genes that potentially regulate double-tailed fin development, and would facilitate marker-assisted selection and breeding of goldfish.

Declarations

The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. All data generated or analysed during this study are included in this published article (and its supplementary information files).

Funding

This study was supported by the National Natural Science Foundation of China (Grant Nos. 31402288).

Competing interests

The authors declare no competing interests.

Authors' contributions

Conceptualization, W.S.Y.; methodology, L.N.Y., and W.S.Y.; data collection, curation, and analysis, F.J., and W.S.Y.; writing-original draft preparations, W.S.Y.; writing, review, and editing, W.S.Y.; funding acquisition, W.S.Y. All authors have read and agreed to the published version of the manuscript.

Ethics approval (include appropriate approvals or waivers)

The use and care of the experimental animals complied with the Regulations for the Administration of Affair Concerning Experimental Animals by State Science and Technology Commission of China.

Acknowledgments

The authors would like to thank Editage (www.editage.cn) for English language editing.

References

1. Abe G, Lee SH, Chang M, Liu SC, Tsai HY, Ota KG (2014) The origin of the bifurcated axial skeletal system in the twin-tail goldfish. Nat Commun 5:3360. https://doi.org/10.1038/ncomms4360
2. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106. https://doi.org/10.1038/npre.2010.4282.1
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29. https://doi.org/10.1038/75556
4. Bairoch A, Boeckmann B (1991) The SWISS-PROT protein sequence data bank. Nucleic Acids Res 19:2247–2249
5. Beier S, Thiel T, Munch T, Scholz U, Mascher M (2017) MISA-web: A web server for microsatellite prediction. Bioinformatics 33:2583–2585. https://doi.org/10.1093/bioinformatics/btx198
6. Boyle WJ, Simonet WS, Lacey DL (2003) Osteoclast differentiation and activation. Nature 423:337–342. https://doi.org/10.1038/nature01658
7. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676. https://doi.org/10.1093/bioinformatics/bti610
8. Diamond J (2002) Evolution, consequences and future of plant and animal domestication. Nature 418:700–707. https://doi.org/10.1038/nature01019
9. Drucker EG, Lauder GV (2001) Locomotor function of the dorsal fin in teleost fishes: Experimental analysis of wake forces in sunfish. J Exp Biol 204:2943–2958
10. Freitas R, Gomez-Marin C, Wilson JM, Casares F, Gomez-Skarmeta JL (2012) Hoxd13 contribution to the evolution of vertebrate appendages. Dev Cell 23:219–1229. https://doi.org/10.1016/j.devcel.2012.10.015
11. Galperin MY, Wolf YI, Makarova KS, Alvarez RV, Landsman D, Koonin EV (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49:D274–D281. https://doi.org/10.1093/nar/gkaa1018

12. Ge QY, Cai Y, Wang GF, Zhao SG (2020) Complete genome analysis of mtDNA in carp and crucian. Genom Appl Biol 39:37–43
13. Grabherr MG, Haas BJ, Yassour M, Thompson DA, Amit I (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644. https://doi.org/10.1038/nbt.1883
14. Guil S, de La Iglesia N, Fernandez-Larrea J, Cifuentes D, Ferrer JC, Guinovart JJ, Bach-Elias M (2003) Alternative splicing of the human proto-oncogene c-H-ras renders a new Ras family protein that trafficks to cytoplasm and nucleus. Cancer Res 63:5178–5187
15. Haston S, Pozzi S, Carreno G, Manshaei S, Panousopoulos L, Gonzalez-Meljem JM, Apps JR, Virasami A, Thavaraj S, Gutteridge A, Forshew T, Marais R, Brandner S, Jacques TS, Andoniadou CL, Martinez-Barbera JP (2017) MAPK pathway control of stem cell proliferation and differentiation in the embryonic pituitary provides insights into the pathogenesis of papillary craniopharyngioma. Development 144:2141–2152. https://doi.org/10.1242/dev.150490
16. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–D280. https://doi.org/10.1093/nar/gkh063
17. Klopfenstein DV, Zhang L, Pedersen BS, Ramirez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, Dampier W, Dessimoz C, Flick P, Tang H (2018) GOATOOLS: a python library for Gene Ontology analyses. Sci Rep 8:10872. https://doi.org/10.1038/s41598-018-28948-z
18. Komiyama T, Kobayashi H, Tateno Y, Inoko H, Gojobori T, Ikeo K (2009) An evolutionary origin and selection process of goldfish. Gene 430:5–11. https://doi.org/10.1016/j.gene.2008.10.019
19. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323. https://doi.org/10.1186/1471-2105-12-323
20. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. https://doi.org/10.1093/bioinformatics/btr509
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btr509
22. Luo J, Stadler PF, He S, Meyer A (2007) PCR survey of *hox* genes in the goldfish *Carassius auratus*. J Exp Zool B Mol Dev Evol 308:250–258. https://doi.org/10.1002/jez.b.21144
23. Moghadam HK, Ferguson MM, Danzmann RG (2005) Evolution of Hox clusters in Salmonidae: a comparative analysis between Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*). J Mol Evol 61:636–649. https://doi.org/10.1007/s00239-004-0338-7
24. Moskow JJ, Bullrich F, Huebner K, Daar IO, Buchberg AM (1995) *Meis1*, a PBX1-related homeobox gene involved in myeloid leukemia in BXH-2 mice. Mol Cell Biol 15:5434–5443. https://doi.org/10.1128/MCB.15.10.5434
25. Paco A, Freitas R (2018) *HoxD* genes and the fin-to-limb transition: insights from fish studies. Genesis 56. https://doi.org/10.1002/dvg.23069
26. Pruitt KD, Tatusova T, Brown RG, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40:D130–D135. https://doi.org/10.1093/nar/gkr1079
27. Pu JW, He DZ, Zou SM (2004) Cloning and function of morphology-related gene HoxB1b in *Megalobrama amblycephala*. J Shanghai Ocean Univ 20:1–7
28. Scemama JL, Vernon JL, Stellwag EJ (2006) Differential expression of *hoxa2a* and *hoxa2b* genes during striped bass embryonic development. Gene Expr Patterns 6:843–848. https://doi.org/10.1016/j.modgep.2006.02.004
29. Shao JH, Xu ZL (2009) Regeneration of fins. Chemistry of Life 29:265–267
30. Smartt J (2001) Goldfish varieties and genetics: handbook for breeders. Blackwell Science, Oxford

31. Stelnicki EJ, Komuves LG, Kwong AO, Holmes D, Klein P, Rozenfeld S, Lawrence HJ, Adzick NS, Harrison M, Largman C (1998) HOX homeobox genes exhibit spatial and temporal changes in expression during human skin development. J Invest Dermatol 110:110–115. https://doi.org/10.1046/j.1523-1747.1998.00092.x
32. van de Sande T, De Schrijver E, Heyns W, Verhoeven G, Swinnen JV (2002) Role of the phosphatidylinositol 3'-kinase/PTEN/Akt kinase pathway in the overexpression of fatty acid synthase in LNCaP prostate cancer cells. Cancer Res 62:642–646
33. Voskas D, Ling LS, Woodgett JR (2014) Signals controlling undifferentiated states in embryonic stem and cancer cells: Role of the phosphatidylinositol 3' kinase pathway. J Cell Physiol 229:1312–1322. https://doi.org/10.1002/jcp.24603
34. Wang CY (1985) Origin of goldfish. Bull Biol 11–13
35. Wang CY (2000) Chinese goldfish. Jindun Press, Beijing
36. Wang SY, Luo J, Murphy RW, Wu SF, Zhu CL, Gao Y, Zhang YP (2013) Origin of Chinese goldfish and sequential loss of genetic diversity accompanies new breeds. PLoS One 8:e59571. https://doi.org/10.1371/journal.pone.0059571
37. Wu J, Mao X, Cai T, Luo J, Wei L (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Res 34:W720–W724. https://doi.org/10.1093/nar/gkl167
38. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39:W316–W322. https://doi.org/10.1093/nar/gkr483
39. Yuan J, He Z, Yuan X, Jiang X, Sun X, Zou S (2010) Speciation of polyploid Cyprinidae fish of common carp, crucian carp, and silver crucian carp derived from duplicated *Hox* genes. J Exp Zool B Mol Dev Evol 314:445–456. https://doi.org/10.1002/jez.b.21350

Figures

Figure 1

The morphological characteristics of the fins of Grass- and Egg-goldfish used in this study. (a) Grass-goldfish (b) Egg-goldfish

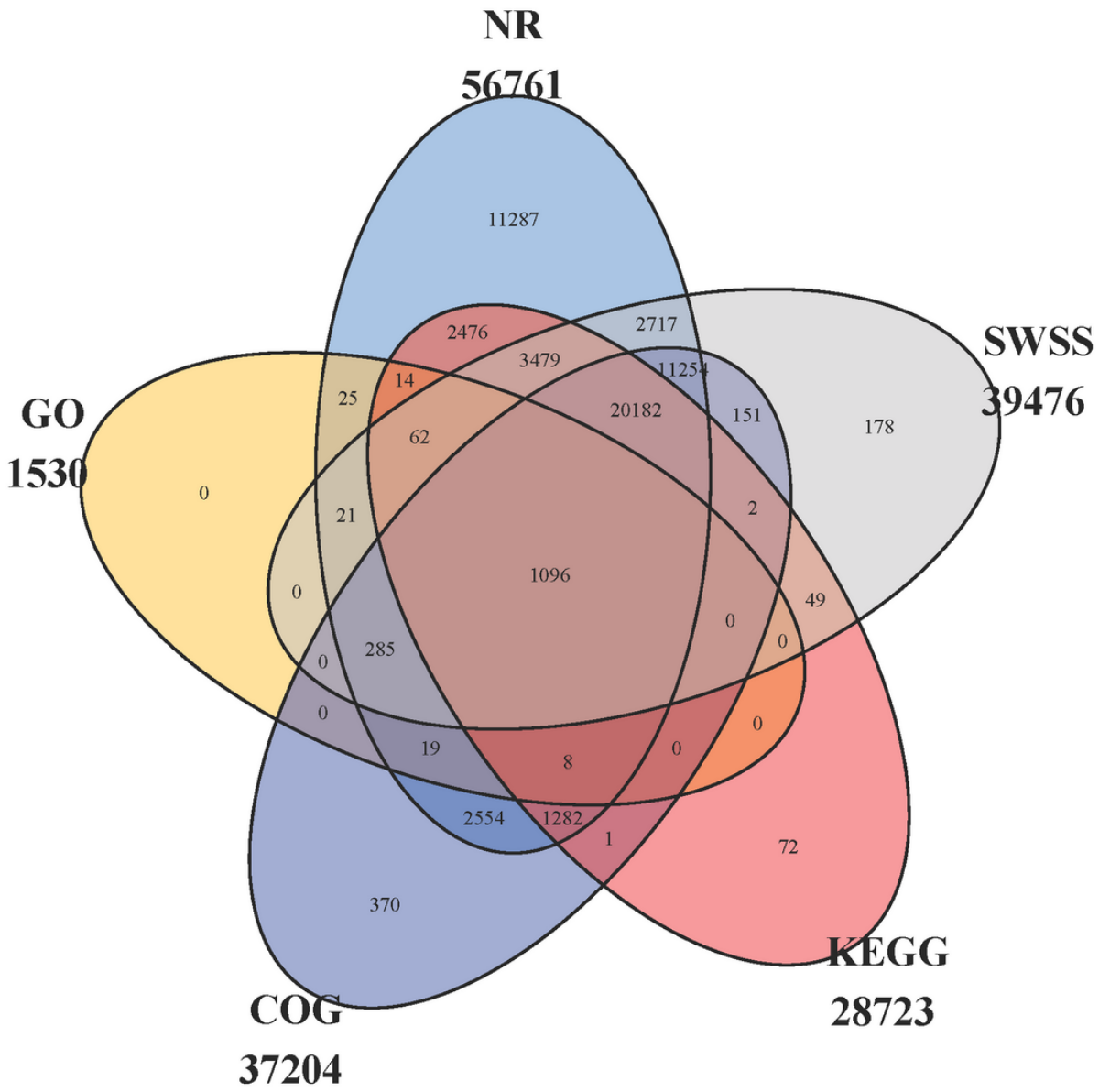

Figure 2

Venn diagram of unigenes from tail fins of goldfish annotated using various databases. The Venn diagram shows the overlapping unigenes annotated in the NR (NCBI non-redundant protein dataset), GO (Gene Ontology), COGs (Clusters of Orthologous Groups), KEGG (Kyoto Encyclopedia of Genes and Genomes Pathway), and SWSS (SwissProt) databases, and the numbers in overlaps show the unigenes annotated to more than one database. Note. NR, ftp://ftp.ncbi.nih.gov/blast/db, Pruitt et al., 2012; GO, http://geneontology.org, Ashburner et al., 2000; COG, https://www.ncbi.nlm.nih.gov/research/cog-project, Galperin et al., 2021; KEGG, http://www.genome.jp/kegg, Kanehisa et al., 2004; SWSS, http://www.uniprot.org, Bairoch and Boeckmann, 1991.

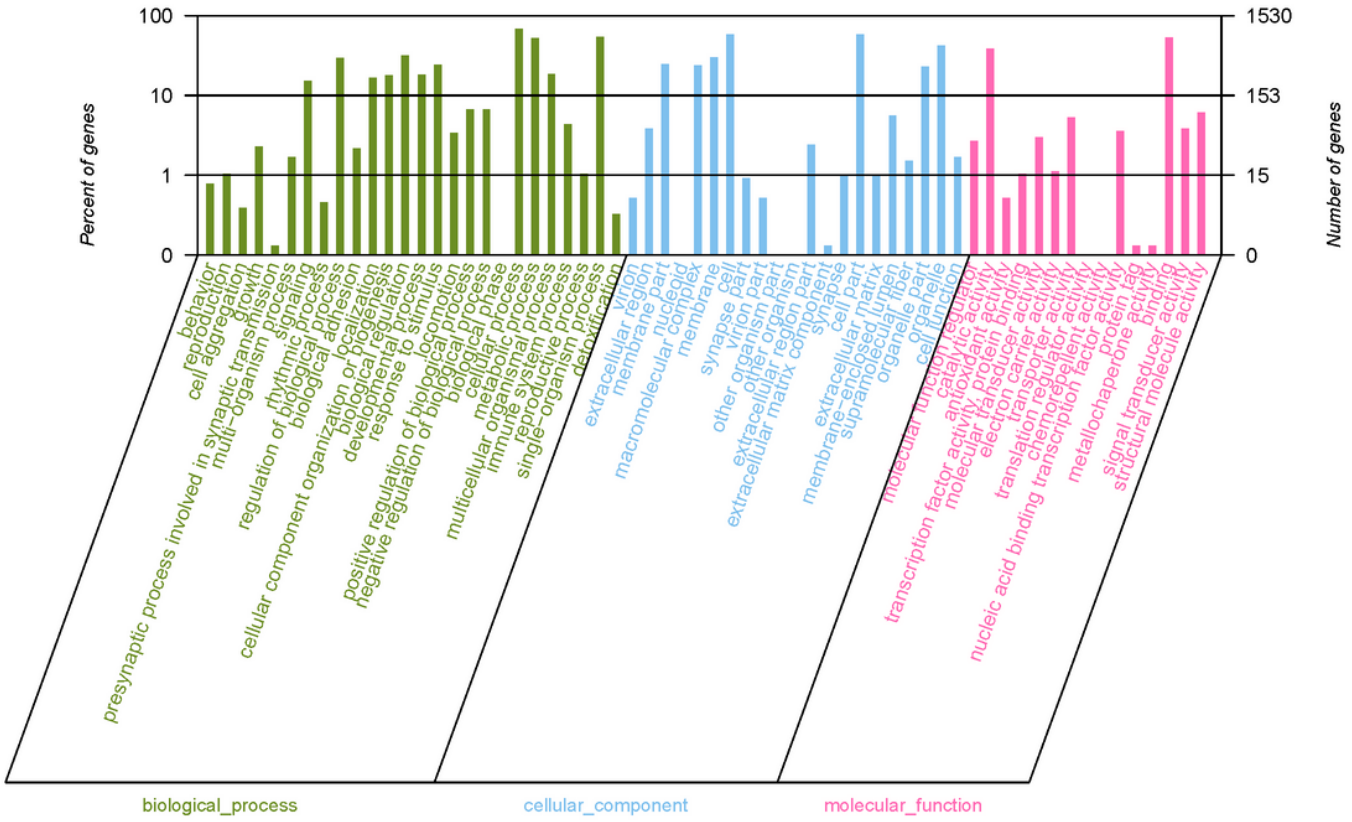

Figure 3

Functional GO (Gene Ontology) annotation of assembled unigenes. The 1,530 unigenes were categorized into 56 GO functional classes belonging to biological processes (green), cellular components (blue) and molecular function (red) Note. GO, http://geneontology.org, Ashburner et al., 2000.

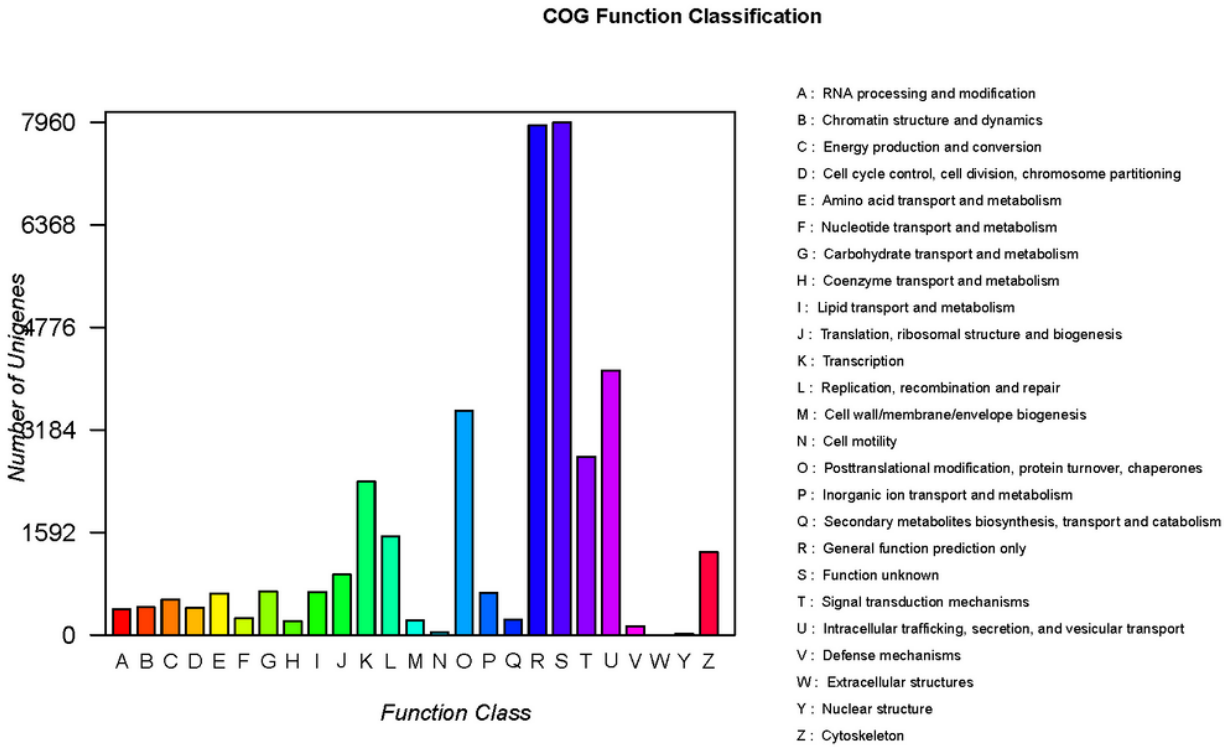

Figure 4

Functional COG (Clusters of Orthologous Groups) annotation of assembled unigenes. The 37,204 unigenes were mapped into 26 categories Note. COG, https://www.ncbi.nlm.nih.gov/research/cog-project, Galperin et al., 2021.

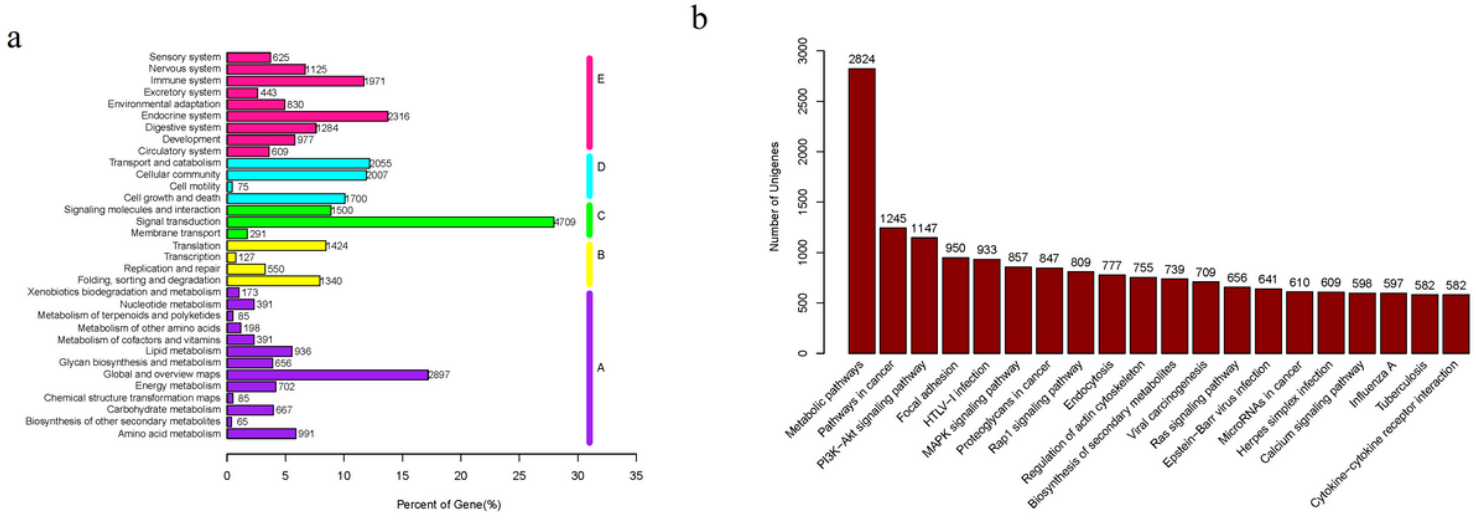

Figure 5

Functional KEGG (Kyoto Encyclopedia of Genes and Genomes Pathway) annotation of assembled unigenes. (a) The 28,723 unigenes were mapped into 340 pathways classified into “A: Cellular Process”, “B: Environmental Information Processing”, “C: Genetic Information Processing”, “D: Metabolism”, and “E: Organismal Systems” (b) The top 20

pathways with the most unigenes in the KEGG annotation Note. KEGG, http://www.genome.jp/kegg, Kanehisa et al., 2004.

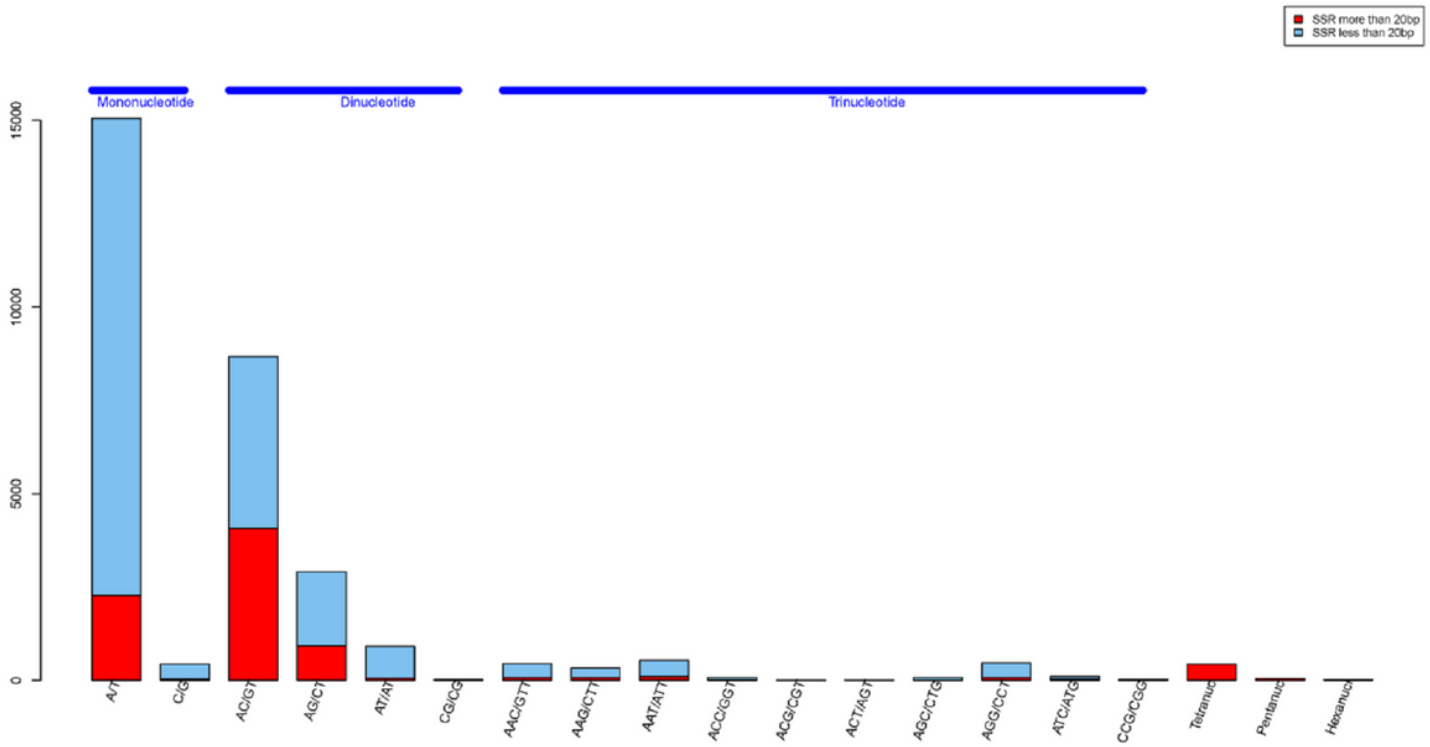

Figure 6

Number of SSRs (simple sequence repeats) discovered in the unigenes from goldfish based on motif sequence type. The red bar represents SSR with length ≥ 15 bp; the blue bar represents SSR with length < 15 bp, and each bar represents a type of SSR

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- tableS1.xlsx
- tableS2.xlsx