GMR
Genetics and Molecular Research
ISSN 16765680

Transcriptome analysis of the grass carp (*Ctenopharyngodon idella*) using 454 pyrosequencing methodology for gene and marker discovery

L.Y. Yu, J.J. Bai, J.J. Fan, D.M. Ma, Y.C. Quan and P. Jiang

Key Laboratory of Tropical & Subtropical Fishery Resource Application
& Cultivation, Ministry of Agriculture;
Pearl River Fisheries Research Institute of Chinese Academy of Fishery Sciences,
Guangzhou, China

Corresponding author: J.J. Bai
E-mail: jjbai@163.net

Genet. Mol. Res. 14 (4): 19249-19263 (2015)
Received August 12, 2015
Accepted October 25, 2015
Published December 29, 2015
DOI http://dx.doi.org/10.4238/2015.December.29.35

ABSTRACT. Total RNA isolated from the brain, muscle, liver, gonad, and intestinal tissues of grass carp was pooled to construct cDNA libraries. Using 454 pyrosequencing, a total of 738,604 high-quality reads were generated from the normalized cDNAs of the pooled individuals. Clustering and assembly of these reads produced a set of 37,086 all-unigene sequences after BLAST. Of these, 24,010 (64.74%) were annotated in the National Center for Biotechnology Information database, and 3715 simple sequence repeats and 2008 single nucleotide polymorphisms were identified in this EST dataset as potential molecular markers. This study provides new data for functional genomic and biological research on grass carp. The markers identified in this study will enrich the currently used molecular markers and facilitate marker-assisted selection in grass carp-breeding programs. These results also demonstrate that transcriptomic analysis based on 454 sequencing is a powerful tool for gene discovery and molecular marker development in non-model species.

Key words: Transcriptome analysis; Grass carp; Pyrosequencing; Molecular marker

INTRODUCTION

The grass carp *(Ctenopharyngodon idella)* is the most common freshwater fish species with an annual production of 4 million tons in China, and occurs naturally in the basins of large water systems such as the Yangtze, Pearl, and Amur Rivers (Li et al., 1998; Ye et al., 2010). Its herbivorous, low-protein diet and excellent growth traits make the grass carp highly suitable for aquaculture in developing countries (Frimodt, 1995; Cudmore and Mandrak, 2004). However, because the adult grass carp is large and has a long breeding cycle, this species is quite difficult to breed, and to date, there is no artificially selected cultivar. Farmed grass carp currently consist primarily of a form domesticated from the wild species, so seedling degeneration is common, with reductions in the seedling survival and growth rates and low antiviral responses. Therefore, the use of genomic tools for the selection of elite bloodstocks will potentially enhance the productivity and value of this species. Furthermore, despite its importance, little genetic or genomic information for the grass carp is available, and the expressed sequence tag (EST) resource in particular is developing only slowly.

The development of an EST resource is a very useful approach for describing the gene expression profiles and mRNA sequences of specific organisms and stages (Wang et al., 2013). Because they provide comprehensive information about the transcriptome, ESTs are a valuable sequence resource for research and breeding purposes. They have played significant roles in functional genomic research, in the discovery of novel genes, and in the identification of different protein groups beyond those provided by the whole genome (Yamada-Akiyama et al., 2009). Among the high-throughput sequencing technologies used for EST profiling, 454 pyrosequencing is of particular interest in evolutionary studies and ecology primarily because it yields longer sequencing reads than any other method (up to 600 bp), which allows more-accurate *de novo* sequence assemblies, which are often required for non-model organisms (Elmer et al., 2010). Recently, 454 pyrosequencing has been used to analyze gene expression in the common carp (*Cyprinus carpio*) (Savan and Sakai, 2002), and channel catfish (*Ictalurus punctatus*) (Karsi et al., 1998). Several functional genomic studies of the grass carp have also been reported. For example, 1141 ESTs were obtained from the grass carp's intestinal cDNA library, this study provided a significant number of ESTs for gene to study digestive system (Zhang et al., 2007); another pair of studies gained 3027 high-quality unigenes from abundant ESTs of head kidney in grass carp, which enriched the molecular basis to identify the immune-relevant genes in grass carp (Chen, et al., 2012). An additional 6269 ESTs have also been isolated as putative markers of intestinal, gill, and hepatopancreatic tissues (Xu et al., 2010). However, functional genes are yet to be characterized, and putative markers are still limited.

This paper reports the large-scale EST analysis in the grass carp using 454 pyrosequencing. It includes an attempt to annotate the EST data and identify the markers in ways that should allow the discovery of new functional genes. This project extends our own research to the transcriptome level, and these sequences will form the basis for future microarray studies. Furthermore, the simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs) of the identified markers will provide a theoretical foundation for grass carp breeding based on a molecular design.

MATERIAL AND METHODS

Ethics statement

Collection of wild grass carp from the Yangtze River and the Pearl River system was permitted by the Department of Fishery, Ministry of Agriculture, China (permission No. 2009HB000320). The current study as a whole, in addition to its protocol, was approved by the Ethics Committee of Animal Experiments of the Pearl River Fishery Research Institute before the study began. All efforts were made to minimize the suffering of the grass carp.

Sample collection and RNA extraction

We utilized 63-month-old individuals from two natural populations in this study: we assigned 28 individuals from the Yangtze River system as the northern population, and 32 individuals from the Pearl River system constituted the southern population. The brain, muscle, liver, gonad, and intestinal tissues of each individual fish were analyzed. Total RNA was isolated from each tissue using the TRIzol reagent (Invitrogen, Carlsbad, CA, USA). mRNA was purified from the total RNA samples using the MicroPoly(A)Purist™ mRNA Purification Kit (Ambion, Cat. No. 1919, Austin, TX, USA), according to manufacturer instructions.

cDNA library synthesis

The mRNA isolated from each tissue pool was combined in a single southern or northern pool to maximize the diversity of the genomes using Trizol Reagent (Invitrogen) according to the manufacturer's instructions. The concentration of total RNA was estimated by measuring the absorbance at 260 nm using a 2100 Bioanalyzer (Agilent Technologies, USA), and RNA integrity was checked by ethidium bromide staining of 28S and 18S ribosomal bands on a 1% agarose gel. Double-stranded cDNA was synthesized following the manufacturer's protocol (Ng et al., 2005). The first-strand cDNA synthesis included a *Gsu*I-oligodT primer, 10 μg mRNA, and 1000 U Superscript II reverse transcriptase (Invitrogen). After incubation at 42°C for 1 h, the 5'-mRNA CAP structure was oxidized by $NaIO_4$ (Sigma) and ligated to biotin hydrazide, which was used bind complete mRNA/cDNA to Dynal M280 beads (Invitrogen). After second-strand cDNA synthesis, the polyA tail and 5'-adaptor were removed by *Gsu*I digestion. cDNA size fractionation was performed with a cDNA size fractionation column (Agencourt). Prepared cDNAs were modified into single-stranded template DNA (sstDNA) libraries with a GS DNA Library Preparation kit (Roche Applied Science). sstDNA libraries were clonally amplified in a bead-immobilized form with a GS emPCR kit (Roche Applied Science). After the bead-enrichment efficiency was examined, a whole-plate sequencing run was performed with Roche 454 GS FLX Titanium chemistry (Roche Diagnostics, Indianapolis, IN, USA). The 454 GS FLX Titanium series reagents can sequence 400-600 x 10^6 bp per run, with read lengths of 400-500 bp.

Output statistics and assembly

Image data output from the sequencing machine were transformed by base calling into sequence data, called “raw data” or “raw reads”, and stored in the FASTQ format. The raw reads produced by sequencing machines contain dirty reads, which contain adapters and unknown or low-quality bases (Ns). These data will negatively affect subsequent bioinformatic analyses. Therefore, dirty raw reads were discarded. Transcriptome assembly was performed with the short-reads assembly program Trinity (Grabherr et al., 2011). Trinity first combines reads that overlap over a certain length to form longer fragments without Ns, which are called “contigs”. These contigs are then further processed into sequence clusters with the sequence-clustering software to form longer sequences without Ns. These sequences are defined as “unigenes”. When multiple samples from the same species are sequenced, unigenes from the assembly of each sample can be processed further with sequence splicing and redundancy removal using the sequence-clustering software to generate non-redundant unigenes that are as long as possible.

Protein-coding region prediction

Unigenes were first aligned with BLASTx (e-value < 0.00001) to protein databases in the priority order: NCBI non-redundant (nr), Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), and clusters of orthologous group (COG). That is, we first aligned the unigenes to the nr database, then to Swiss-Prot, then to KEGG, and finally to COG. Unigenes aligned to a higher-priority database were not aligned to a lower-priority database. Alignment was complete when all the alignments had been made. Proteins ranking highest in the BLAST results were used to determine the coding-region sequences of the unigenes. The coding-region sequences were then translated into amino acid sequences with the standard codon table, so that both the nucleotide sequences (5'→3') and amino acid sequences of the unigene coding regions were acquired. Unigenes that could not be aligned to any database were scanned with ESTScan (Iseli et al., 1999) to obtain the nucleotide sequence (5'→3') direction and the amino acid sequence of the predicted coding region.

Functional annotation and gene ontology (GO) classification

Unigene annotation provides information about the expression and functional classification of the unigenes. Information from functional annotation allows GO functional annotation, COG functional annotation, and protein functional annotation of the unigenes. GO is an international standardized gene function classification system that offers a dynamically updated controlled vocabulary and a strictly defined concept for comprehensively describing the properties of genes and their products in any organism. GO has three ontologies: molecular function, cellular component, and biological process. The basic unit of GO is the “GO-term”. Each GO-term belongs to a type of ontology.

SSR and SNP marker identification

The ESTs were screened for SSRs using the MIcroSAtellite (MISA) software identification tool (http://pgrc.ipk-gatersleben.de/misa/) with the following parameters: at least eight repeats were required for di-, six repeats for tri- and tetra-, and three repeats for pentanucleotides (Iseli et al., 1999). To identify putative SNPs, the reads were mapped onto the transcripts using SSAHA2 (Iseli et al., 1999) using the default parameters, and the SNPs were then extracted using VarScan set to the default parameters. A sequence variation was deemed to be a putative SNP when a mismatch was identified in reads with four or more sequences and when the minor allele sequence existed at least twice within the reads. To identify high-quality SNPs, putative SNPs (identified as described above) were further screened with specific criteria based on the read depth, minor allele frequency, quality of the flanking regions, and absence of other SNPs within 20-bp flanking sequence. Only those SNPs with minor allele sequences representing more than 15% of the reads aligned at the polymorphic loci were deemed to be quality SNPs. No additional SNPs within the 20-bp flanking regions were allowed.

RESULTS AND DISCUSSION

Statistics and assembly

Roche 454 GS FLX high-throughput sequencing technology was used to sequence the cDNA libraries of two grass carp populations. The southern population contained 353,529 ESTs with an average length of 357 bp, and the northern population contained 385,075 ESTs with an average length of 371 bp. The average length of the ESTs in the combined library was 385 bp (Table 1). The raw reads obtained with sequencing were not always valid because they included repeated and low-quality reads, and these reads might negatively influence the assembly and subsequent analysis. Therefore, the impurities in the reads were removed to generate clean reads. All cleaned reads generated in this study have been deposited in the NCBI Sequence Read Archive database (accession number: SRA111136). Comparison of the two cDNA libraries after purification indicated that the amount of valid data for the northern population was greater than that available for the southern population. This indicated that the quality of the reads obtained by sequencing the northern population was better than that for the southern population.

Table 1. Sequencing output statistics.

Samples	Total reads	Total nucleotides	Q20 percentage	N percentage	GC percentage
Southern	353,529	132,430,729	98.84%	0.00%	46.28%
Northern	385,075	138,659,867	98.38%	0.00%	47.04%

Total nucleotides = total reads 1 x read 1 size + total reads 2 x read 2 size; Total reads = total reads 1 + total reads 2.

The assembly software Trinity was used for transcriptome assembly of the reads from the two libraries, and the unigenes of the northern and southern populations were obtained. Further splicing and redundancy removal were then performed on the unigenes of the two libraries. A statistical analysis of the length distributions of the unigenes and all-unigenes was performed (Figure 1). Most of the ESTs were distributed in the 500-600-bp range. The data volume of the all-unigenes obtained after the assembly of the two libraries (37,086) was significantly higher than that for either the northern (24,066) or the southern population library (25,466) alone.

Transcriptome sequencing is one of the most important tools in gene discovery. However, large-scale EST sequencing using the traditional Sanger method is time consuming and expensive (Wang et al., 2010). With advances in sequencing technologies, next-generation sequencing offers tremendous advantages for high-throughput gene discovery on a genome-wide scale in non-model organisms (Cheung et al., 2006). Because the Roche GS FLX system generates long reads and was commercially developed before the other platforms, it is the most widely used platform for transcriptome sequencing in many organisms, including the blunt snout bream (Gao et al., 2012), chestnut (Barakat et al., 2009), pine (Parchman et al., 2010), maize (Vega-Arreguin et al., 2009), and insects (Hahn et al., 2009). Consistent with this research, our results indicate that the relatively short reads produced with GS FLX sequencing can be effectively assembled and used for novel gene discovery and marker development in a non-model organism. Here, approximately 37,000 100-bp reads were generated using Genome Analyzer. This huge number of reads resulted in a relatively high depth of coverage. These sequences, on basis of the length compare of N50 parameter, also produced longer unigenes (950 bp) than those produced in previous studies using earlier 454 technology with shorter reads (e.g., 197 bp (Vera et al., 2008); 440 bp (Meyer et al., 2009); 500 bp (Parchman et al., 2010); and 422 bp (Roeding et al., 2009). Furthermore, a large proportion of our unigenes were greater than 500 bp in length and the high quality of the assembled unigenes was reflected in the high proportion of unigenes that matched known proteins when analyzed with BLAST searches.

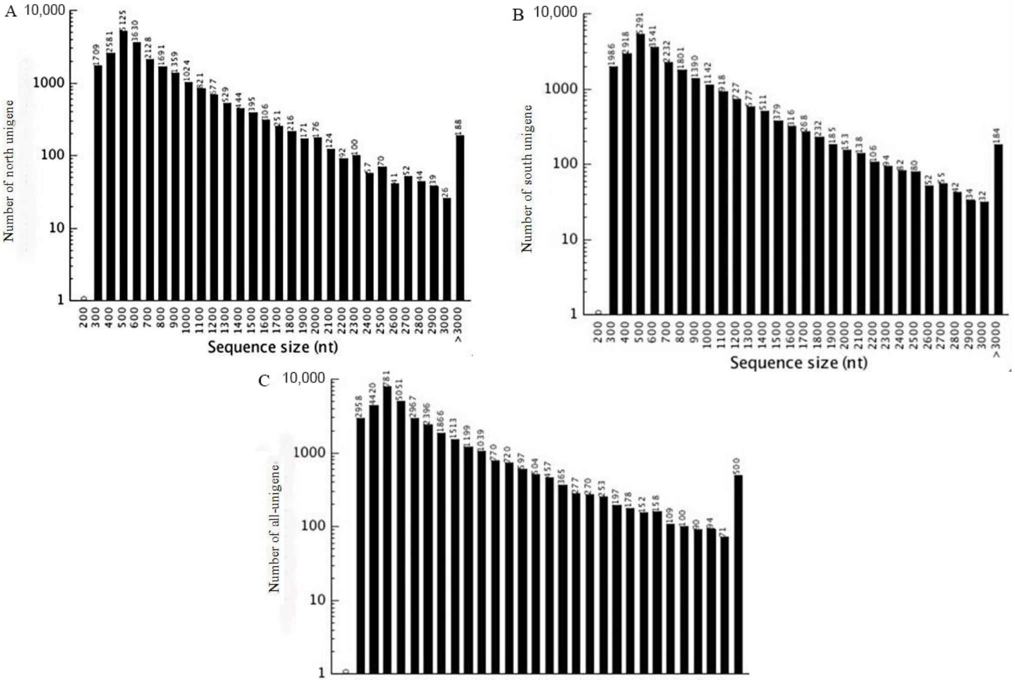

Figure 1. Characteristics of the assembled unigenes of the grass carp. **A.** Length distributions of the northern unigenes. **B.** Length distributions of southern unigenes. **C.** Length distributions of all-unigenes.

Functional annotation

Estimating the number of genes and the level of transcript coverage represented in an EST collection are important issues in transcriptome sequencing projects, but are difficult or impossible without a completely annotated reference genome sequence (Gao et al., 2012). To assess the extent of transcript coverage provided by the unigenes and to evaluate how the coverage depth affected the assembly of the unigenes, we plotted the ratio of assembled unigene length to zebrafish length against the coverage depth. Most of the zebrafish coding regions were covered by our individual unigenes. A large proportion of the grass carp unigenes showed their best BLAST matches to zebrafish proteins, with a smaller proportion of unigenes matching proteins of other taxonomic groups (Figure 2). Some of the unigenes or singletons without BLAST matches probably represent additional genes not represented in the annotated protein databases searched, or genes that could not be matched with BLAST because they are too short.

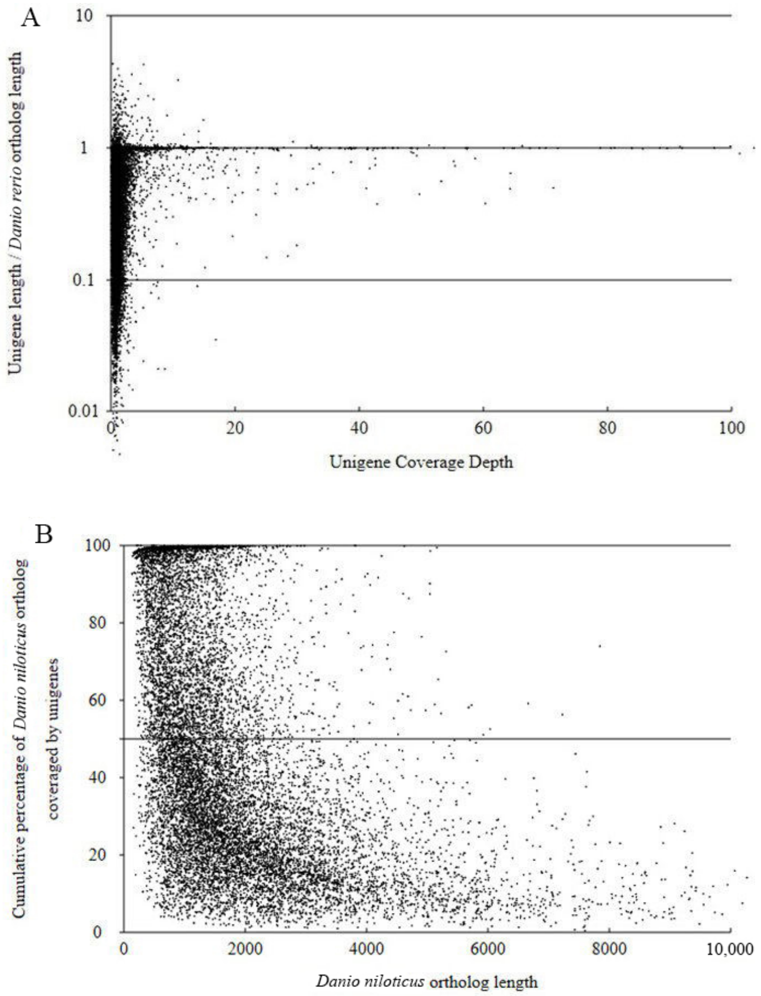

Figure 2. Comparison of grass carp unigenes and orthologous *Danio rerio* coding sequences. **A.** Ratio of grass carp unigene lengths to *D. rerio* orthologue lengths plotted against the grass carp unigene coverage depth. **B.** Total percentage of *D. rerio* orthologous coding sequences that were covered by all the grass carp unigenes.

To validate and annotate the assembled unigenes, sequence similarity searches were conducted against the NCBI nr, Swiss-Prot, COG, and KEGG databases using the BLASTx algorithm. The results indicated that of 37,086 unigenes, 23,977 (64.65%), 21,597 (58.23%), 12,497 (33.70%), and 7088 (19.11%) showed significant similarity to known proteins in the nr, Swiss-Prot, KEGG, and COG databases, respectively. Among all the unigenes, 24,010 (64.74%) were annotated in the NCBI database. Finally, the annotated unigenes were classified in the four databases (Figure 3). These results indicated that 5361 unigenes were annotated in all four databases (nr, Swiss-Prot, KEGG, and COG); 1924, 13, and 13 unigenes were annotated to only the nr, Swiss-Prot, or KEGG databases, respectively; and no unigenes were annotated only to COG. The unigenes annotated to the nr database were used for homology comparisons in 298 species, and 73.92% of the unigenes were annotated as zebrafish orthologues.

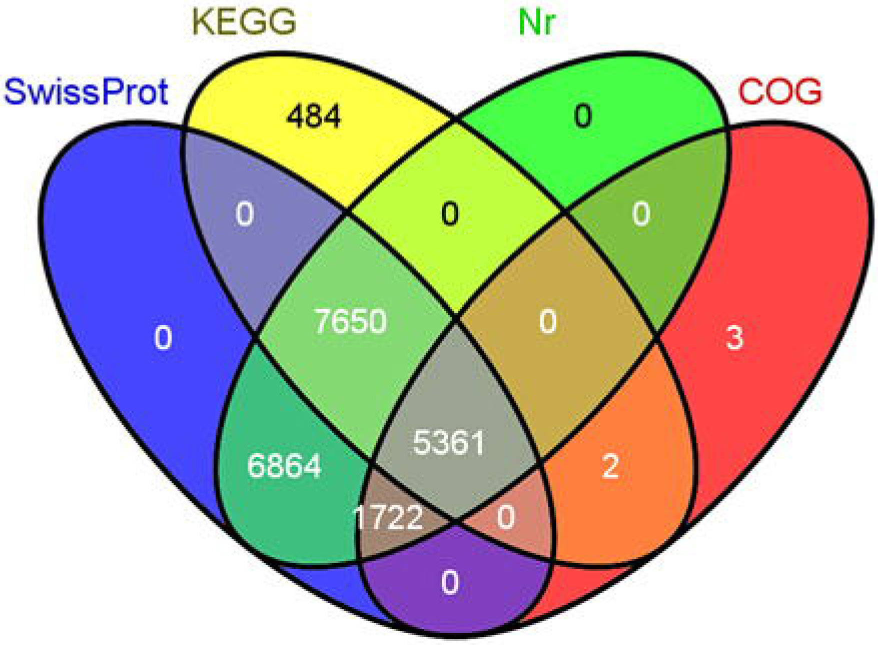

Figure 3. Functional annotation classification of all-unigenes of the grass carp in the non-redundant (nr), Swiss-Prot, KEGG, and COG databases.

Protein-coding region prediction

A BLASTx alignment between the unigene sequences and the protein libraries cited above was conducted according to descending priorities: nr, Swiss-Prot, KEGG, and COG databases. The CDSs of 23,746 unigenes were predicted and translated into amino acid sequences according to the standard codon table. The ESTScan software was used to predict the CDSs for the 320 unigene sequences that could not be matched to any protein library. The nucleic acid sequences (5'→3') and amino acid sequences corresponding to the CDSs were obtained (Iseli et al., 1999).

Among all the predicted CDSs, 75.52% were shorter than 700 bp, of which most were approximately 400 bp (17.49%). Some CDSs were over 3000 bp in length (0.32%). The length distributions of the CDSs and amino acids are shown in Figure 4.

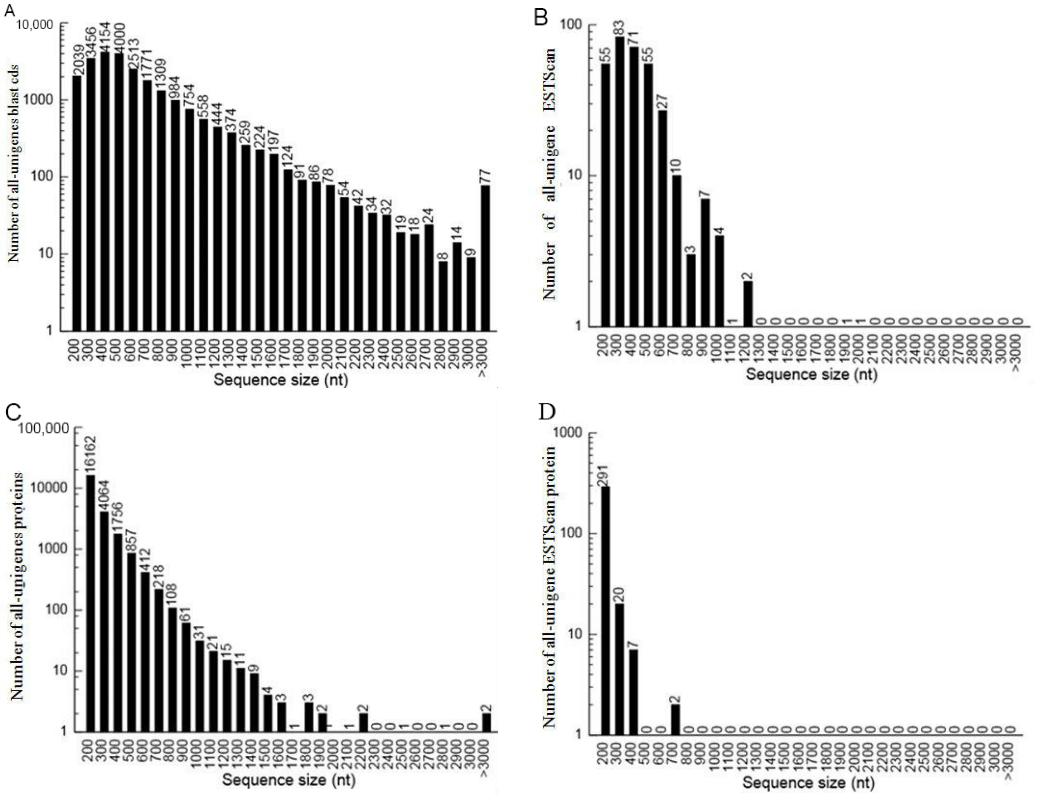

Figure 4. Characteristics of all assembled unigenes (all-unigenes) of the grass carp. **A.** Length distribution of all-unigene BLAST CDSs. **B.** Length distribution of all-unigene ESTScan CDSs. **C.** Length distribution of all-unigene BLAST proteins. **D.** Length distribution of all-unigene ESTScan proteins. CDS, coding DNA sequence.

GO classification

GO is an international standardized functional gene classification system that offers a dynamically updated controlled vocabulary and a strictly defined concept with which the properties of the genes and their products in any organism can be comprehensively described. With nr annotation, we used the Blast2GO program (Conesa et al., 2005) to generate GO annotations for the unigenes. After GO annotation was obtained for every unigene, we used the WEGO software (Ye et al., 2006) to determine the GO functional classifications of all the unigenes and to understand the distribution of the gene functions in this species on the macro level. In total, 41,077 unigenes could be annotated to 46 functional groups in the GO database, including 19,260 into “biological processes” (46.89%), 14,573 into “cell components” (35.48%), and 7244 into “molecular functions” (17.64%). The distribution of the unigenes in the 46 functional groups is shown in Figure 5.

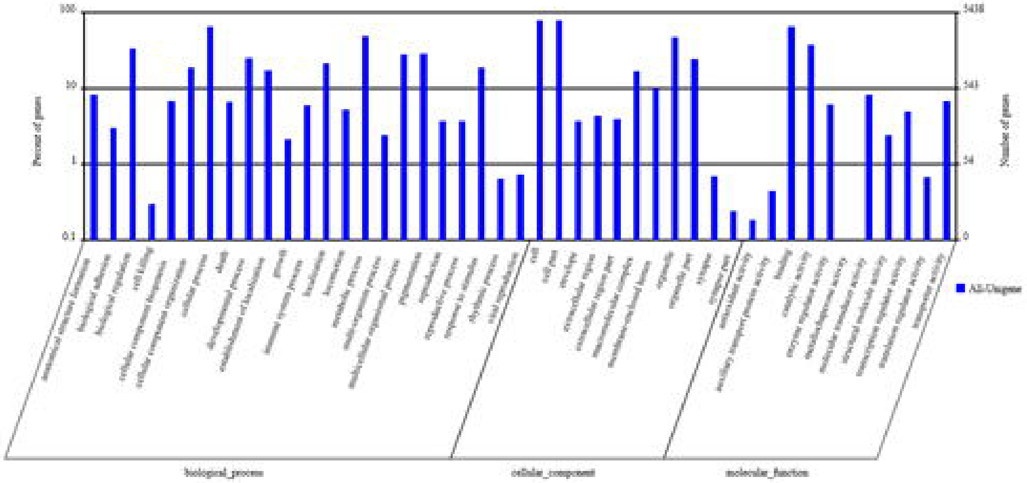

Figure 5. Functional gene ontology (GO) classification of all-unigenes of the grass carp.

COG annotation

COG is a database in which orthologous gene products are classified. Every protein in COG is assumed to have evolved from an ancestral protein, and the whole database is built on coding proteins within complete genomes and the systematic evolutionary relationships of bacteria, algae, and eukaryotic organisms. Unigenes are aligned to the COG database to predict and classify the possible functions of their products. Among the 37,086 unigenes of the grass carp obtained with transcriptome sequencing and assembly, 14,504 unigenes were successfully annotated to the COG database. These unigenes were classified into 25 COG categories (Figure 6). The unigenes that clustered in the "general function" category represented the largest proportion (2530 unigenes, 17.44%), followed by those in "transcription" (1101 unigenes, 7.59%), and "transcription, ribosomal structure, and biogenesis" (1077 unigenes, 7.43%). The categories with the least annotated unigenes were "nuclear structure" (12 unigenes, 0.08%) and "extracellular structure" (12 unigenes, 0.08%).

Functional classifications by KEGG

The KEGG pathway database records the networks of molecular interactions in cells and the variants specific to particular organisms. This pathway-based analysis extends our understanding of the biological functions and interactions of genes. Based on a comparison with the KEGG database using BLASTx, 13,497 of the 37,086 unigenes had significant matches in the database and were assigned to 274 KEGG pathways. Of these, 604 (4.48%) unigenes were assigned to the "cancer" pathway, and 526 (3.9%) unigenes were assigned to the "focal adhesion" pathway. COG analysis and KEGG pathway analysis are useful in predicting potential genes and their functions at the level of the whole transcriptome. COG, together with the predicted metabolic pathways, can be used for further investigation of gene functions in future studies.

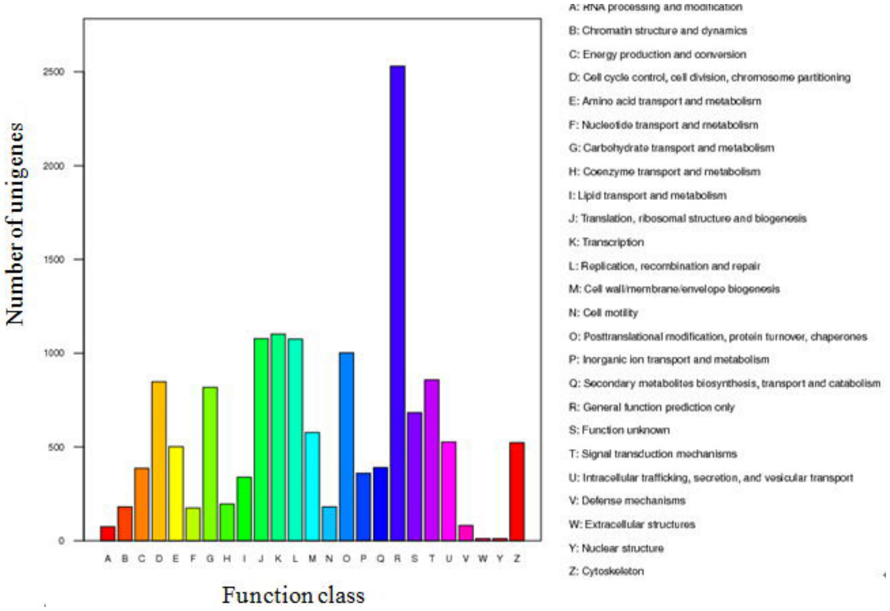

Figure 6. Clusters of orthologous group (COG) functional classifications of the all-unigenes of the grass carp.

Development and characterization of cDNA-derived SNP and SSR markers

To develop new molecular markers, all of the 37,086 unigenes generated in this study were screened for potential microsatellites, which are defined as di- to hexanucleotide SSRs with a minimum of four repetitions of all motifs. Using the MISA identification tool, 3715 potential combined SSRs (cSSRs) were identified in 3253 unigenes; of these, 409 sequences contained more than one cSSR and 218 cSSRs were present in a compound form (Table 2). The frequencies, types, and distributions of the potential 3715 cSSRs were also analyzed in this study. The compilation of all cSSRs revealed that on the average, one cSSR could be found every 7.86 kb in the unigenes, and the frequency of cSSRs was 10.02%. Among the 3715 cSSRs, the di- and trinucleotide repeat motifs were most abundant (2143, 57.69% and 1059, 28.51%, respectively), followed by the tetra- (421, 11.33%), penta- (68, 1.83%), and hexanucleotide repeat motifs (24, 0.65%). Di- to hexanucleotide motifs were further analyzed for cSSR length (number of repeat units; Table 3).

Table 2. Summary of cSSR search results.

Search Item	Number
Total number of sequences examined	37,086
Total size of sequences examined (bp)	29,197,047
Total number of cSSRs identified	3715
Number of cSSR-containing sequences	3253
Number of sequences containing more than 1 cSSR	409
Number of cSSRs present in compound formation	218
Di-nucleotide	2143
Tri-nucleotide	1059
Tetra-nucleotide	421
Penta-nucleotide	68
Hexa-nucleotide	24

cSSR = combined simple sequence repeats.

Table 3. Length distributions of cSSRs based on the number of repeated units.

Number of repeated units	Di-	Tri-	Tetra-	Penta-	Hexa-
4	0	0	280	58	23
5	0	556	112	7	1
6	722	283	26	2	0
7	434	171	2	0	0
8	269	43	1	0	0
9	230	0	0	0	0
10	274	1	0	1	0
11	203	3	0	0	0
12	11	1	0	0	0
13	0	0	0	0	0
14	0	1	0	0	0

cSSR = combined simple sequence repeats.

Within the cSSRs examined, 73 sequence-type motifs were identified, and the di-, tri-, penta-, and hexanucleotide repeats contained 4, 10, 23, 20, and 16 types, respectively. The AC/GT dinucleotide repeat was the most abundant motif detected in our cSSRs (1107, 29.8%), followed by the motifs AG/CT (527, 14.2%), AT/AT (505, 13.6%), AAT/ATT (277, 7.5%), ATC/ATG (186, 5%), AGG/CCT (154, 4.1%), AGC/CTG (151, 4.1%), and AAAT/ATTT (113, 3%). The frequencies of the remaining 65 types of motifs accounted for 18.7% of the cSSRs (Figure 7).

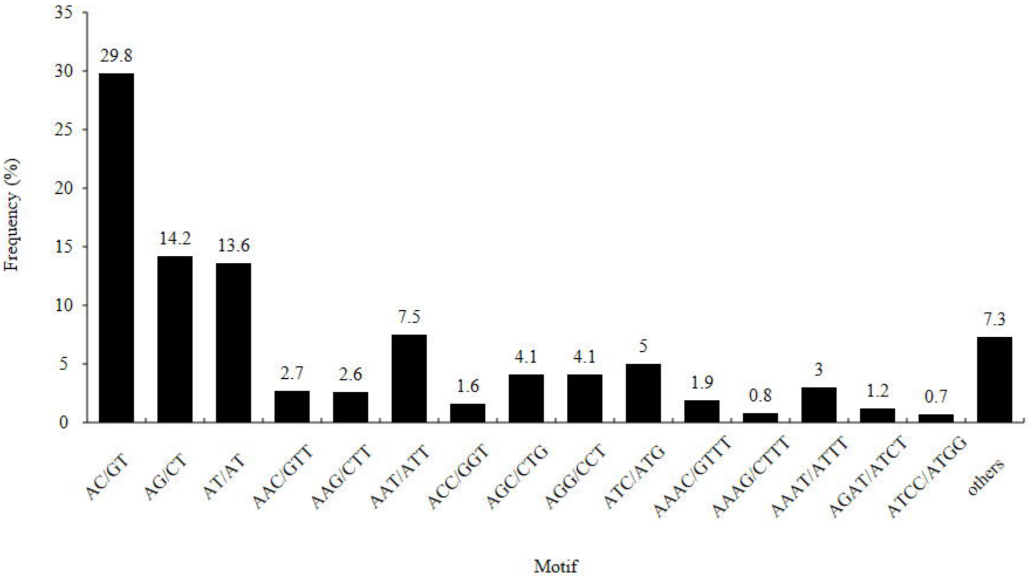

Figure 7. Frequency distribution of cSSRs based on sequence-type motif. cSSR, combined simple sequence repeat.

To analyze the differential SSR sites within and between the populations, SSR screening and analysis were conducted on the reads, the assembled unigenes within each population, and the unigenes between the populations. According to the results, 65 and 75 differential SSR markers were identified in the southern and northern populations, respectively, and 11 differential SSR markers were identified between the populations. Based on these analyses, we designed 453 pairs of SSR primers according to the unigene sequences in which each marker was located, laying the foundation for SSR marker verification and marker-assisted breeding of the grass carp.

The GS Reference Mapper (454 Life Science) software was used to identify polymorphisms among the ESTs by aligning individual reads against unigenes from the assembly. For a sequence difference to be deemed a true polymorphism, at least two individual reads aligned to the consensus must have contained the variant allele and at least two others must have contained the allele with the consensus sequence. Applying this criterion, 4616 and 4543 SNP sites were identified in the northern and southern populations, and 12,608 SNP sites were identified altogether, including 3449 common sites. In all the types of SNP mutations, the ratio of transition to transversion (t_s/t_v) was 1.72:1 (2183:1266). To improve the reliability and accuracy of the analysis, we conducted a depth filter with a coverage depth of ≥8X to the low-quality loci. Finally, 1239 and 1304 SNP sites were obtained in the northern and southern populations, respectively, and 2008 SNP sites were obtained altogether, including 535 common sites. Transitions accounted for the largest proportion of SNP mutations specific to the two populations (Figure 8).

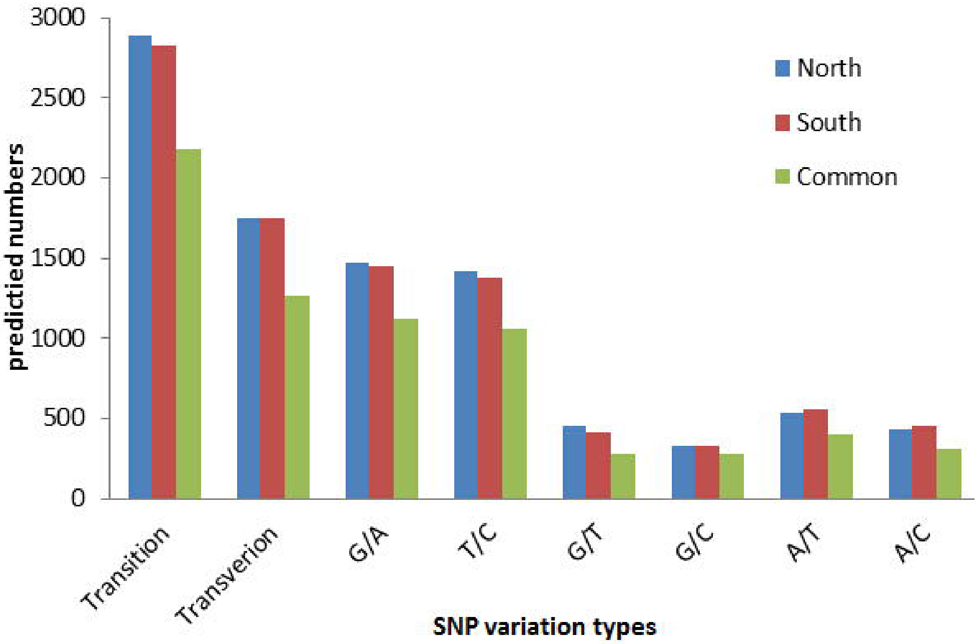

Figure 8. Distribution of putative SNPs in the transcriptomes of the two populations of grass carp. SNP, single nucleotide polymorphism.

CONCLUSION

In this study, *de novo* transcriptome sequencing was applied to the grass carp genome using the 454 GS FLX system. A large number of high-quality transcriptome reads were obtained. The assembly of these reads produced a set of 37,086 unigene sequences after BLAST, and 24,010 of the total unigenes were annotated in the NCBI database. These will be useful for future comparative genomic studies of the grass carp. Large numbers of SSRs and SNPs were predicted, which will benefit breeding programs and whole-genome association studies to enhance the performance and production traits of the grass carp.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by grants from the National High-Tech Research and Development Program (‘‘863’’ Plans) of China (#2011AA100403), the National Key Technology R&D Program (#2012BAD26B02), and the Modern Agroindustry Technology Research System entitled “Staple Freshwater Fishery Industry Technology System” (#CARS-46-03). We thank the Fishery Seeds Limited Company of Nanhai Bairong and Sanshui Baijin for help with grass carp sample collection.

REFERENCES

Barakat A, DiLoreto DS, Zhang Y, Smith C, et al. (2009). Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biol.* 9: 51.

Chen J, Li C, Huang R, Du F, et al. (2012). Transcriptome analysis of head kindney in grass carp and discovery of immune-related genes. *BMC Vet. Res.* 8: 108

Cheung F, Haas BJ, Goldberg SM, May GD, et al. (2006). Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* 7: 272.

Conesa A, Götz S, García-Gómez JM, Terol J, et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676.

Cudmore B and Mandrak NE (2004). Biological synopsis of grass carp (*Ctenopharyngodon idellus*). Canadian Manuscript Report of Fisheries and Aquatic Sciences 2705. Available at [http:sbisrvntweb.uqac.ca/archivage/24061712.pdf].

Elmer KR, Fan S, Gunter HM, Jones JC, et al. (2010). Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Mol. Ecol.* 19: 197-211.

Frimodt C (1995). Multilingual illustrated guide to the world’s commercial warm water fish. Wiley-Blackwell, Oxford.

Gao ZX, Luo WL, Liu H, Zeng C, et al. (2012). Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (*Megalobrama amblycephala*). *PLoS One* 7: e42673.

Grabherr MG, Hass BJ, Yassour M, Levin JZ, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29: 644-652.

Hahn DA, Ragland GJ, Shoemaker DD and Denlinger DL (2009). Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophagi crassipalpis*. *BMC Genomics* 10: 234.

Iseli C, Jongeneel CV and Bucher P (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1999: 138-148.

Karsi A, Li P, Dunham R and Liu ZJ (1998). Transcriptional activities in the pituitaries of channel catfish before and after induced ovulation by injection of carp pituitary extract as revealed by expressed sequence tag analysis. *J. Mol. Endocrinol*. 21: 121-129.

Li SF, Lv GQ and Louis BA (1998). Diversity of mitochondrial DNA in the population of silver carp, bighead carp, grass carp and black carp in the middle and lower reaches of the Yangtze River. *Acta Zool.Sin.* 4: 82-93.

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter JA, et al. (2009). Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFLx. *BMC Genomics* 10: 219.

Ng P, Wei CL, Sung WK, Chiu KP, et al. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* 2: 105-111.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, et al. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* 11: 180.

Roeding F, Borner J, Kube M, Klages S, et al. (2009). A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol. Phylogenet. Evol.* 53: 826-834.

Savan R and Sakai M (2002). Analysis of expressed sequence tags (EST) obtained from common carp, *Cyprinus carpio* L., head kidney cells after stimulation by two mitogens, lipopolysaccharide and concanavalin-A. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* 131: 71-82.

Vega-Arreguin JC, Ibarra-Laclette E, Jiménez-Moraila B, Martínez O, et al. (2009). Deep sampling of the Palomero maize transcriptome by a high throughput strangy of pyrosequencing. *BMC Genomics* 10: 299.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, et al. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* 17: 1636-1647.

Wang R, Xu S, Jiang Y, Jiang J, et al. (2013). *De novo* sequence assembly and characterization of *Lycoris aurea* transcriptome using GS FLX titanium platform of 454 pyrosequencing. *PLoS One* 8: e60499.

Wang ZY, Fang BP, Chen JY, Zhang XJ, et al. (2010). *De novo* assembly and characterization of root transcriptome using llumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). *BMC Genomics* 11: 726.

Xu B, Wang S, Jiang Y, Yang L, et al. (2010). Generation of analysis of ESTs from the grass carp, *Ctenopharyngodon idellus*. *Anim. Biotechnol.* 21: 217-225.

Yamada-Akiyama H, Akiyama Y, Ebina M, Xu Q, et al. (2009). Analysis of expressed sequence tags in apomictic guineagrass (*Panicum maximum*). *J. Plant Physiol.* 166: 750-761.

Ye J, Fang L, Zheng H, Zhang Y, et al. (2006). WEGO: a web tool for plotting GO annotations: *Nucleic Acids Res.* 34: W293-W297.

Ye X, Zhang LL, Dong H, Tian Y, et al. (2010). Validation of reference genes of grass carp *Ctenopharyngodon idellus* for the normalization of quantitative real-time PCR. *Biotechnol. Lett*. 32: 1031-1038.

Zhang XJ, Qu G, Zhu WL, Zhang L, et al. (2007). Construction of intestinal cDNA library and analysis of some expressed sequence tags sequencing of *Ctenopharyngodon idellus*. *Acta Hydrobiol. Sin.* 31: 251-258.