Genome Biology and Evolution

SMBE

Comprehensive transcriptome analysis reveals accelerated genic evolution in a Tibet fish, *Gymnodiptychus pachycheilus*

Liandong Yang[1,2]
Email: yangliandong1987@163.com

Ying Wang[1,2]
Email: xinyuanwangying@163.com

Zhaolei Zhang[3,4]
Email: zhaolei.zhang@utoronto.ca

Shunping He[1*]
Email: clad@ihb.ac.cn

[1] The Key Laboratory of Aquatic Biodiversity and Conservation of Chinese Academy of Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, Hubei 430072, People's Republic of China

[2] University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

[3]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

[4] Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada

[*] Corresponding author. Tel. +86 27-68780071. Fax. +86 27-68780430.

Genome Biology and Evolution

Abstract

Elucidating the genetic mechanisms of organismal adaptation to the Tibetan Plateau at a genomic scale can provide insights into the process of adaptive evolution. Many highland species have been investigated and various candidate genes that may be responsible for highland adaptation have been identified. However, we know little about the genomic basis of adaptation to Tibet in fishes. Here we performed transcriptome sequencing of a schizothoracine fish (*Gymnodiptychus pachycheilus*) and used it to identify potential genetic mechanisms of highland adaptation. We obtained totally 66,105 assembled unigenes, of which 7,232 were assigned as putative one-to-one orthologs in zebrafish. Comparative gene annotations from several species indicated that at least 350 genes lost and 41 gained since the divergence between *G. pachycheilus* and zebrafish. An analysis of 6,324 orthologs among zebrafish, fugu, medaka, and spotted gar identified consistent evidence for genome-wide accelerated evolution in *G. pachycheilus* and only the terminal branch of *G. pachycheilus* had an elevated Ka/Ks ratio than the ancestral branch. Many functional categories related to hypoxia and energy metabolism exhibited rapid evolution in *G. pachycheilus* relative to zebrafish. Genes showing signature of rapid evolution and positive selection in the *G. pachycheilus* lineage were also enriched in functions associated with energy metabolism and hypoxia. The first genomic resources for fish in the Tibetan Plateau and evolutionary analyses provided some novel insights into highland adaptation in fishes and served as a foundation for future studies aiming to identify candidate genes underlying the genetic bases of adaptation to Tibet in fishes.

Keywords: Tibetan Plateau, Adaptation, Positive selection, Schizothoracine fish, Transcriptome

SMBE

Introduction

Understanding how species adapt to extreme environments is a central goal in evolutionary biology (Smith and Eyre-Walker 2002). As the world's highest and largest plateau, the Tibetan Plateau, with an average elevation of 4,500 metres above sea level, imposes many inhospitable living conditions on most organisms, including cold temperatures, low oxygen concentrations, and strong ultraviolet radiation (Bickler and Buck 2007; Thompson, et al. 2000). Nevertheless, several species have well adapted to these harsh living challenges. Indeed, recent genome-wide studies on multiple species have identified various adaptive processes that may be responsible for highland adaptation, including humans (Beall, et al. 2010; Bigham, et al. 2010; Peng, et al. 2011; Simonson, et al. 2010; Xing, et al. 2013; Xu, et al. 2011; Yi, et al. 2010), yak (Qiu, et al. 2012), Tibetan antelope (Ge, et al. 2013), the ground tit (Cai, et al. 2013; Qu, et al. 2013), and Tibetan Mastiff (Gou, et al. 2014; Li, et al. 2014). Among these adaptive processes, it was well-known that genes showing signals of positive selection and expansion were significantly enriched in hypoxia-inducible factor (HIF) and energy metabolic pathways. For example, hypoxia related genes (such as *EPAS1*, *EGLN1*, and *PPARA*) have experienced strongly positive selection and are significantly associated with the decreased hemoglobin concentration in Tibetans (Beall, et al. 2010; Simonson, et al. 2010; Yi, et al. 2010). However, almost all previous genome-wide studies were performed on endothermic terrestrial vertebrates. We know little about the genomic bases of adaptation to highland in fishes. Therefore, it may provide some novel insights by investigating the genetic mechanisms of adaptation to the Tibetan Plateau in fishes.

The schizothoracine fishes (Teleostei: Cyprinidae), which are distributed throughout the Tibetan Plateau and its peripheral regions, are the largest and most diverse taxon of the Tibetan Plateau ichthyofauna (Cao, et al. 1981; Chen and Cao 2000). These fishes are the only taxon within the most successful family Cyprinidae that have well adapted to the hostile environment of the Tibetan Plateau (He, et al. 2004). The schizothoracine fishes dominate the plateau lakes and torrential mountain streams of the Tibetan Plateau (He and Chen 2006) and have evolved a number of unique traits to adapt to the hypoxia and cold environment (Wu and Wu 1991). Therefore, they have been considered as excellent models to investigate high altitude adaptation of fishes. According to the degree of specialization of the scales, pharyngeal teeth, and barbels, the schizothoracine fishes are divided into three grades: primitive, specialized, and highly specialized schizothoracine fishes and the orderly reductions of these morphological characteristics in these groups were closely associated with the drastic environmental changes caused by three stages of violent upheaval of the Tibetan Plateau (Cao, et al. 1981). Thus, it was suggested that the three phases of uplift of the Tibetan Plateau have contributed to the speciation of the schizothoracine fishes. The species *Gymnodiptychus pachycheilus*, belonging to the specialized schizothoracine fishes, distributes only in the headwater area in the northeast of the Tibetan Plateau with elevations of 2750-3750 m (He, et al. 2004) and is the most dominant group of the ichthyofauna of the Yellow River (Wu and Wu 1991). However, human activities, including overexploitation and habitat destruction, have affected this species considerably, which makes *G. pachycheilus* listed as an endangered species in the "China Species Red List" (Wang and Xie 2004). Thus, characterization and evolutionary analyses of its transcriptome resources can not only provide information of highland adaptation of fishes, but also help protect its population.

The recent rapid advances in sequencing technologies have offered the opportunity to generate transcriptomes in almost any species of interest. When genome sequence is not available, transcriptome sequencing is a rapid and effective approach to obtain massive protein-coding genes and molecular makers. In this study, we generated the first transcriptome of a schizothoracine fish (*G. pachycheilus*) endemic to the Tibetan Plateau using high-throughput sequencing technology. We then characterized the transcriptome comprehensively and performed evolutionary analyses together with other previously available fish genomes to investigate the potential mechanisms of highland adaptation of fishes.

Materials and Methods

Fish sampling, RNA extraction and sequencing

All animal experiments were performed in accordance with the ethics committee of Institute of Hydrobiology, Chinese Academy of Sciences. One wild schizothoracine fish (*Gymnodiptychus pachycheilus*) was sampled from Gansu Fisheries and Science Research Institute, Lanzhou, Gansu, China. To obtain as many expressed genes as possible, five different types of organs (heart, brain, liver, kidney, and spleen) were sampled and stored in RNAlater (QIAGEN) immediately. Total RNA was isolated using the SV Total RNA Isolation System (Promega) according to the manufacturer's protocol and the quality of RNA was measured using electrophoresis and the BioPhotometer plus 6132 (Eppendorf, Germany). Poly (A) mRNA was purified using Oligo (dT) magnetic beads and interrupted into short fragments. Subsequently, the first-strand cDNA was synthesized using random hexamer primer and then second strand cDNA was generated. Finally, the paired-end cDNA library was prepared according to the Illumina's protocols and sequenced (101 bp read length) on Illumina HiSeq 2000 platform. The sequencing data have been deposited into the NCBI Sequence Read Archive database (Accession No. SRR1583887)

De novo assembly

The raw reads were first preprocessed and filtered by removing reads with sequencing adaptors, reads with unknown nucleotides and low quality (quality scores < 20). All subsequent analyses were based on these filtered reads. Next, transcriptome *de novo* assembly was performed using Trinity software (Grabherr, et al. 2011) with default parameters. Only contigs longer than 200 bp were kept for further analysis. Then, CD-HIT-EST program (Li and Godzik 2006) was used to further remove the redundancy in the final assembly.

Gene annotation

To annotate the assembled unigenes, we first downloaded the protein datasets of zebrafish (*Danio rerio*) from the Ensembl database (release-75) (Flicek, et al. 2013) and then using BLASTX (Altschul, et al. 1997) searches to map the unigenes to these proteins with an E-value cutoff of 1×10^{-10}. In order to identify genes that may be lost (or missing) in the zebrafish genome, unigenes without hits against zebrafish proteins were used to search against protein datasets from other model fishes *Astyanax mexicanus*, *Gadus morhua*, *Gasterosteus aculeatus*, *Oreochromis niloticus*, *Oryzias latipes*, *Takifugu rubripes*, *Tetraodon nigroviridis*, and *Xiphophorus maculatus* from the Ensembl database. Then, those unigenes with hits in other model fishes were further searched against the zebrafish genome with BLASTN and BLAT (Kent 2002) to confirm that these putative genes were lost in the zebrafish genome. Putative functions for assembled unigenes were assigned

by Blast2GO suit (Gotz, et al. 2008) using BLASTX against the non-redundant (NR) databases with a conservative E-value cutoff of 1 × 10-5. We then extracted the ORFs using getorf tool implemented in EMBOSS (Rice, et al. 2000) and predicted the protein coding potential for the assembled unigenes using CPAT (Wang, et al. 2013), with Zebrafish (Zv9/danRer7) as the assembly database and 0.38 as the coding probability cutoff.

Identification of orthologs

Putative orthologs between *G. pachycheilus* and zebrafish were determined using the reciprocal BLAST best-hit method with an E-value cutoff of 1×10^{-10}. Then one-to-one orthologs between zebrafish, fugu, medaka, and spotted gar (*Lepisosteus oculatus*) were obtained from Ensembl using Biomart (Durinck, et al. 2005). When genes had multiple transcripts, the longest one was used. Each orthologous gene set was aligned using PRANK (Loytynoja and Goldman 2005) with the parameter "-codon" and trimmed using GBlocks (Castresana 2000) with the parameter "-t = c". We further deleted all gaps and "N" from the alignments to lower the effect of ambiguous bases on the inference of positive selection. After the deletion process, the trimmed alignments shorter than 150 bp (50 codons) were discarded for subsequent analyses.

Substitution rate estimation and selection analyses

To estimate lineage-specific evolutionary rates for each branch of the five species, the Codeml program in the PAML package (Yang 2007) with the free-ratio model (model = 1) was run on each ortholog, a concatenation of all alignments of the orthologs, and 1000 concatenated alignments constructed from 10 randomly chosen ortholog. Parameters, including dN, dS, dN/dS, N*dN, and S*dS values, were obtained for each branch and genes were discarded if N*dN or S*dS < 1, or dS >2, according to previous study (Goodman, et al. 2009).

We used the branch model to identify fast evolving genes with the null model assuming that all branches have been evolving at the same rate and the alternative model allowing foreground branch to evolve under a different rate. The likelihood ratio test (LTR) with df = 1 was used to discriminate between alternative model for each ortholog in the gene set. Multiple testing was corrected by applying the false discovery rate method (FDR) implemented in R (Storey and Tibshirani 2003). We considered the genes as evolving with a significantly faster rate in foreground branch if the FDR-adjusted p-value less than 0.05 and a higher ω values in the foreground branch than the background branches.

To detect positive selection on a few codons along specific lineage, we used the optimized branch-site model (Zhang, et al. 2005) following the author's recommendation. A likelihood ratio test was constructed to compare a model that allows sites to be under positive selection on the foreground branch with the null model in which sites may evolve neutrally and under purifying selection. The p-values were computed based on the Chi-square statistic adjusted by the FDR method and genes with adjusted p-value < 0.05 were treated as candidates for positive selection.

Gene ontology (GO) functional enrichment analyses for both fast evolving genes and positively selected genes were carried out by DAVID (Dennis, et al. 2003; Huang da, et al. 2009).

Results

Sequence analysis and assembly

A mixed sample of cDNAs obtained from five tissues, including heart, brain, liver, kidney, and spleen, was prepared and sequenced using the Illumina HiSeq 2000 platform, which produced 22,805,393 raw 101 bp paried-end reads. After removing adaptors and low-quality reads, we obtained 22,728,725 quality filtered reads pairs with a median read length of 100 bp. With these high quality reads, 132,794 reconstructed contigs were generated using Trinity, with a median length of 745 base pairs (bp) and an N50 of 2,322 bp. We further used CD-HIT-EST to produce a non-redundant unigene datasets and obtained 66,105 unigenes ranging from 201 bp to 21,730 bp, with a median length of 710 bp and an N50 of 1,602 bp (supplementary Table S1). The length distribution of all unigenes is provided and nearly 60% of the unigenes are between 200 to 500 bp. (supplementary Fig. S1). A significantly positive relationship between the length of unigenes and number of reads covered was observed, with an average coverage depth of 263 reads (supplementary Fig. S2).

To assess the quality of our assembled unigenes, we downloaded all 13 mitochondrial protein coding genes available for *G. pachycheilus* from NCBI database as reference sequences and compared our assembled unigenes to these reference genes using BLASTN with an E-value cutoff of 1×10^{-10}. All these protein coding genes were found to be present in our assembled unigenes. The proportions of mismatching nucleotides relative to the reference sequences were calculated and only 0.74% mean nucleotide difference was observed. We further obtained the complete mitochondrial genome sequence (Wu, et al. 2014) to evaluate the completeness and continuity of our assembled unigenes. We found a total of 15,429 nucleotide identities out of 15,514 (99.5%) total nucleotide length of unigene relative to whole mitochondrial sequences. In addition to above computing method, we further performed reverse transcription polymerase chain reaction (RT-PCR) to validate the quality of our assembled unigenes. We randomly picked 10 unigenes with different expression levels (RPKM ranged from 51 to 889). Primers for these unigenes were designed and all these cDNAs can be successfully amplified (supplementary Table S2 and supplementary Fig. S3). These results demonstrated reliable transcriptome assembly quality, which is the foundation for subsequently comparative genomic analysis.

Functional annotation

We used several complementary approaches to annotate the assembled unigenes. First, a BLASTX search against zebrafish proteins returned 28,586 (43.2%) *G. pachycheilus* unigenes with significant hits to zebrafish genes. This percentage of unigenes with BLAST hits is comparable with previous *de novo* transcriptome studies for non-model organisms (Guo, et al. 2013), in which unigenes without significant hits may consist of orphan genes, non-coding RNAs, untranslated transcripts, or misassembled transcripts. Second, we used Blast2GO with the Gene Ontology (GO) annotation database to assign their putative functions and 24,131 unigenes have one or more GO terms (supplementary Fig. S4). Finally, Clusters of Orthologous Groups of protein (COGs) databases were used to further annotate these unigenes and produced good results for 9,740 putative proteins (supplementary Fig. S5).

Homologs to known proteins of any species in the non-redundant (NR) databases were identified

for 31,733 unigenes which represent 48% of the total *de novo* reference transcriptome assembly. Overall, 11,058 unigenes had an E-value of BLASTX results between 1×10^{-5} to 1×10^{-50} and 9,103 unigenes have an E-value of 0. More than half (52.7%, n = 16,726) of the homologs to known proteins have identity between 80% to 100%. And 55.3% (n = 17,560) of the best hits were with zebrafish, which may reflect the close phylogenetic relationship between these two species, or reflect the wealthy genomic resources for zebrafish (supplementary Fig. S6). We next divided our assembly unigenes into two subsets (unigenes with and without protein homology in NR database, named as "with hits" and "no hits" set) and characterized their sequence and expression features in detail. Overall, the "with hits" set had significantly larger unigene length (median 758 bp vs 318 bp) and longer ORFs (median 462 bp vs 132 bp) than the "no hits" set (Wilcoxon rank sum test, $P < 2.2 \times 10^{-5}$) (Fig. 1A). Analysis of the potential for protein coding with CPAT (Wang, et al. 2013) revealed a significantly lower protein coding potential in the group of unigenes without hits (Wilcoxon rank sum test, $P < 2.2 \times 10^{-6}$) (Fig. 1B). The distributions of GC content and normalized expression level also show that, in general, unigenes with BLAST hits have higher values than those without hits (Wilcoxon rank sum test, $P < 2.2 \times 10^{-6}$) (Fig. 1C, D). These characteristics between the "with hits" and "no hits" set of unigenes in *G. pachycheilus* were consistent with previous reports on non-model species without a reference genome (Ferreira, et al. 2013; Schunter, et al. 2014), which indicates that many novel unigenes may be non-protein coding sequences.

Additionally, we found that 1,065 unigenes, which did not have significant BLASTX hit against protein sequences from zebrafish, had significant hits against proteins from at least one of the other eight fish genomes obtained from Ensembl. After performing BLASTN and BLAT searches against zebrafish genome, 350 out of the 1,065 unigenes were confirmed to have no hits in the zebrafish genome (supplementary Table S3). Considering that these unigenes have orthologous genes in other teleost genomes, we thought that the orthologs of these unigenes are probably lost in the zebrafish genome instead of being gained in *G. pachycheilus*.

Orphan genes in *G. pachycheilus*

In the past few years, substantial progress has demonstrated that lineage-specific new genes can rapidly evolve indispensable biological roles and make a contribution to lineage-specific phenotype and adaptation (Chen, et al. 2013). Thus, it is meaningful to identify putative novel protein-coding genes (orphan genes) in *G. pachycheilus*, which might have evolved specific functional roles and contributed to their adaptation to Tibetan plateau. To investigate this, we first predicted the protein coding potential for each of the assembled unigenes in *G. pachycheilus* using the CPAT program. Out of the 66,105 assembled unigenes, 15,845 (24%) were predicted as protein coding genes and these included 1,565 (10%) unigenes that had no identifiable zebrafish ortholog. To exclude any orthologs of these unigenes in other species, we further searched them against the NR databases and identified 744 with orthologs in any other species in NR databases. Among the remaining 821 unigenes, we set several cutoffs by calculating the median size (758 bp), the median protein coding potential score (0.24), and the median expression level (RPKM = 2) of the 31,733 unigenes with identifiable orthologs against NR databases and detected 88 unigenes that were longer, had higher protein coding potential, higher expression level than the median values of the unigenes with known protein coding orthologs. Furthermore, as recommended by CPAT,

0.38 is the optimum cutoff to filter false protein coding genes in fishes (Wang, et al. 2013). Thus, we used this cutoff to further remove the candidate de novo protein coding genes that have coding potential lower than 0.38. The remaining 88 unigenes were further searched against the zebrafish and other fish genome sequences and resulted in significant blastn hit for 47 unigenes. Finally, we identified 41 putative orphan genes specific to *G. pachycheilus*, which originated around 50 million years ago (mya) after the split from zebrafish (Steinke, et al. 2006) (supplementary Table S4).

Accelerated evolution on the lineage leading to Tibet fish

To better understand the evolutionary dynamics of Tibet fish, we analyzed the putative single copy orthologs in *G. pachycheilus*, zebrafish, fugu, medaka, and spotted gar genomes. After alignment and trimming for quality control (see Materials and Methods), a total of 6,324 orthologs, ranging from 150 to 13,707 bp, were determined. Despite the lengths of the orthologs were shorter after trimming, their shapes of length distributions were generally similar (supplementary Fig. S7), which ensured subsequent evolutionary analyses.

First, to compare the overall difference in selective constraints in different branch at the gene level, each orthologous gene was evaluated for substitution rates including Ka, Ks, and Ka/Ks, using the species tree (Near, et al. 2012) (Fig. 2A). The free-ratio model (M1 model) in PAML was used, which allows an independent Ka/Ks ratio for each branch (Yang 2007). Averaged across all 6,324 orthologous genes, the *G. pachycheilus* branch had a significantly higher ratio of nonsynonymous to synonymous substitutions than other fish branches (Wilcoxon rank sum test, $P < 2.2 \times 10^{-16}$), suggesting accelerated function evolution in the *G. pachycheilus* lineage (Fig. 2B). Indeed, by examining the Ka/Ks ratio for each gene in the *G. pachycheilus* and zebrafish lineages, we found that 2,607 genes have higher Ka/Ks in *G. pachycheilus* while only 1,607 genes higher in zebrafish. We further calculated the Ka/Ks ratio for each branch for a concatenated alignment of all 6,324 orthologs and 1,000 concatenated alignments constructed from 10 randomly chosen orthologs, and found that both datasets exhibited a significantly higher Ka/Ks ratio for the *G. pachycheilus* branch than other fish branches (Wilcoxon rank sum test, $P < 2.2 \times 10^{-16}$) (Fig. 2C and 2D). Furthermore, comparison of Ka/Ks ratios between terminal and ancestral branches indicated that only the *G. pachycheilus* branch had an elevated Ka/Ks ratio than the ancestral branch (Wilcoxon rank sum test, $P < 2.2 \times 10^{-16}$) (Fig. 2A), implying accelerated evolution only in *G. pachycheilus* after their split from zebrafish.

To identify the GO categories that undergone rapid or slow evolution in *G. pachycheilus* than zebrafish, we calculated the average Ka/Ks ratios for each GO category with at least 10 orthologs in *G. pachycheilus* and zebrafish lineages, respectively. Among these GO categories, the number of GO category with average Ka/Ks ratios higher in *G. pachycheilus* lineage was significantly larger than the number of GO category with average Ka/Ks ratios higher in zebrafish lineage (1,108 vs 258) and there was significantly larger number of GO categories with statistically significantly higher average Ka/Ks ratios in *G. pachycheilus* than in zebrafish lineage (480 vs 2), confirming overall accelerated evolution in *G. pachycheilus*. Furthermore, many GO categories involved in energy metabolism, hypoxia response, and DNA repair showed significantly accelerated evolution in *G. pachycheilus* than zebrafish, such as ‘response to oxidative stress’,

'blood vessel morphogenesis', 'glucose metabolic process', 'NAD binding', and 'positive regulation of DNA repair' (Fig. 3 and supplementary Table S5).

Fast evolving and positively selected genes

To detect genes that might evolve adaptively in specific lineage, two types of gene sets were compiled: 1) fast evolving genes (FEGs), which have experienced a significantly higher Ka/Ks ratio in specific lineage compared with other lineages, and 2) positively selected genes (PSGs), which have been influenced by positive selection only on a few codons along particular lineage (see Materials and Methods). In total, we identified 883 FEGs in *G. pachycheilus* and 556 FEGs in zebrafish, and 123 PSGs in *G. pachycheilus* and 111 PSGs in zebrafish, respectively (supplementary Table S6 and Table S7). Functional enrichment analysis showed that the FEGs identified in *G. pachycheilus* lineage were significantly enriched for genes involved in energy metabolism and oxidation related functions, including 'ATP binding', 'mitochondrion', 'regulation of GTPase activity', and 'Oxidative phosphorylation', while FEGs detected in zebrafish were generally enriched in functions involved in structure components (Fig. 4 and supplementary Table S8). Similarly, the PSGs identified in *G. pachycheilus* lineage rather than in zebrafish were also enriched for genes potentially related to hypoxia response, including epidermal growth factor (EGF) (Fig. 4 and supplementary Table S9). In addition, we found that the putative PSGs, whose P-values were not corrected by FDR method, were also enriched for genes involved in adaptation to high-elevation environment, such as 'vasculature development' (supplementary Table S9).

To identify genes that may directly contribute to the adaptation to high altitude, we combined two approaches to detect all the candidate genes according to their functional roles. First, we compared our candidate genes (PSGs) to an *a priori* list proposed by Zhang *et al*. (Zhang, et al. 2014), which includes 1,351 putative hypoxia-related genes. Second, we made use of the functional annotated information for each PSG to identify the gene associated with hypoxia response reported in previous experimental studies. In total, we identified 9 candidate PSGs in *G. pachycheilus* that may be involved in hypoxia response: BYSL, HSF1, YES1, ARRDC2, SSPN, SEMA4D, VWF, COMP, and LAMB (Table 1).

Discussion

Over the past few years, comparative genomics has been widely employed as a tool to understand the genetic bases of many fundamental evolutionary questions, including adaptation (Axelsson, et al. 2013; Jones, et al. 2012; Yi, et al. 2010; Zhao, et al. 2013), speciation (Ellegren, et al. 2012; Poelstra, et al. 2014; Soria-Carrasco, et al. 2014), and genetic variation (Guo, et al. 2012). When the genome sequencing data is not available, transcriptome sequencing is an effective and accessible approach to initiate comparative genomic analyses on non-model organisms, because they contain large number of protein-coding genes likely enriched for targets of natural selection. Here, using the next-generation sequencing technology, we have generated and annotated the first comprehensive transcriptome resources for a schizothoracine fish (*G. pachycheilus*), which is endemic to the Tibetan Plateau and shows many unique traits to adapt to highland environments (Su, et al. 2014; Wu and Wu 1991). We generated more than seven thousand pairwise orthologous genes between zebrafish and over six thousand orthologous genes among other fish genomes, which are important bases for comparative genomic studies of adaptation in fishes. Therefore, the

transcriptome resources produced by our study are useful to understand the genetic makeup of fishes in high altitude and provide a foundation for further studies to identify candidate genes underlying adaptation to the Tibetan Plateau of fishes.

Gene losses and gains are important adaptive processes that have a contribution to evolutionary innovations (Ding, et al. 2012; Hahn, et al. 2007). Thus, we first attempted to identify genes that present in *G. pachycheilus* but lost in zebrafish and genes gained specific in *G. pachycheilus* through comparison of orthologous genes between *G. pachycheilus* and other fish genomes. By setting a strict set of cutoffs, we revealed that as many as 350 genes have a potential to have been lost in zebrafish because they exist both in *G. pachycheilus* and other fish genomes. There are also alternative possibilities that these genes have evolved too fast to resemble their orthologs in other fishes, or that they are missed from the current zebrafish genome assembly. Among the genes that might have been lost in zebrafish, many have GO categories associated with binding, including RNA, protein, and nucleic acid binding, which is similar to potentially lost genes in three-spined sticklebacks (Guo, et al. 2013). On the other hand, we identified at least 41 genes that are uniquely present in the *G. pachycheilus* transcriptome data set compared with other fishes. These genes are likely to have originated in the schizothoracine fish lineage around 50 mya after split from the zebrafish (Steinke, et al. 2006) and might have evolved novel functional roles that may be contributing to the adaptation to high altitude of schizothoracine fishes. Even though it is possible that these new genes may have evolved too fast only in schizothoracine fish lineage to be detected in other fishes, or represent ancestral genes that have lost function in other fishes and accumulated substitutions too fast to be identified as homologs by standard BLAST searches, they are still important and interesting, as fast evolution itself may be an adaptive process. These new genes should be important targets in future studies aiming at elucidating the genetic basis of adaptation to highland of fishes. In addition to de novo genes, gene gain can also be mediated by duplication. However, we could not infer such recent duplication event considering that there is no genome sequence. Therefore, a more thorough understanding of the number and function of genes lost and gained within schizothoracine fishes can only be achieved by increasing taxon sampling and whole genome sequencing.

In addition to loss and gain of genes, adaptive evolution may prefer to occur at the molecular level, expressed by an increased rate of nonsynonymous substitutions to synonymous substitutions (Bakewell, et al. 2007). The major adaptations to highland habitat of different endothermic organisms are expansion of gene families, increased evolutionary rate and positive selection on genes associated with hypoxia response and energy metabolism (Ge, et al. 2013; Qiu, et al. 2012; Qu, et al. 2013). Species living in similar ecological environment can be shaped by convergent evolution to form physiological or morphological similarities (Stern 2013). Just like previous studies in endothermic animals, our evolutionary analyses suggested that the schizothoracine fish can also be characterized by its adaptation to the extreme environment of the Tibetan plateau at the molecular level. First, the schizothoracine fish lineage showed genome-wide accelerated evolution relative to other fish lineages, which is independent of the dataset used. Thus, the schizothoracine fishes may have adaptively speeded up their evolutionary rates of genes overall to better adapt to the extreme environment of the Tibetan Plateau, as accelerated evolution is usually driven by positive selection. It is also possible that accelerated evolution could be caused by relaxation of

functional constraint, which yet needs to be further confirmed from population genomic analyses in future. Second, only the terminal branch of the schizothoracine fishes had undergone elevated evolutionary rates than the ancestral branch, suggesting that accelerated evolution only occurred in the schizothoracine fish lineage after split from zebrafish. Third, Functional GO categories related to hypoxia response and energy metabolism were found to have evolved faster in the schizothoracine fish lineage. Fourth, rapidly evolving and positively selected genes in the schizothoracine fish lineage were also enriched in categories involved in energy metabolism and hypoxia. All in all, these results indicated that the schizothoracine fishes may have experienced adaptive evolution to cope with the extremely inhospitable environment. However, our current evidence only showed accelerated protein sequence evolution in *G. pachycheilus* and whether gene content evolution (gene losses and gains) is also accelerated remains as an interesting question in the next stage.

The most extreme challenge for species living in high-altitude is low oxygen supply (Beall 2007). To identify the potential genes directly involved in hypoxia, we focused on the function of positively selected genes in the schizothoracine fish lineage and found several interesting candidate genes that may be involved in response to hypoxia. For example, the bystin-like (BYSL) gene that encodes an accessory protein for cell adhesion significantly up-regulated induced by hypoxia, suggesting an important role in hypoxia response (Fang, et al. 2008). The activation of heat shock proteins (Hsps) is critical to adaptation to hypoxia, which is regulated by HSF1 (heat shock transcription factor 1). And HSF1 is up-regulated directly by HIF-1 (hypoxia-inducible factor-1), suggesting a link of positively selected genes to hypoxia pathway (Baird, et al. 2006). YES1 (v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1), a member of the Src family of tyrosine protein kinases acting on focal adhesions and contacts (Gaudreault, et al. 2005; Lynch, et al. 2004), has a relation to HSP27 (Hansen, et al. 2001), which can be specifically unregulated by hypoxic signaling through HIF-1 (Whitlock, et al. 2005). Genes, ARRDC2 and SSPN are reported to be related to hypoxia and collected in the database of hypoxia-regulated proteins (HypoxiaDB) (Khurana, et al. 2013). SEMA4D can promote angiogenesis acting through Plexin-B1 on endothelial cells, which is regulated by HIF-1 (Sun, et al. 2009). VWF (von Willebrand factor), an adhesive glycoprotein that expressed exclusively in endothelial cells, increased expression levels in pulmonary hypertension caused by hypoxia (Caramuru, et al. 2003; Mojiri, et al. 2013). Although there are several candidate genes that are potentially involved in hypoxia showing signature of positive selection, none is shared with previously reported genes in other endothermic animals. This observation suggests that fishes may have employed different genic toolkit to adapt to the extreme environment of the Tibetan Plateau. However, this hypothesis needs to be further confirmed by population genomics in future.

Acknowledgements

This work was supported by the Pilot projects (Grant No. XDB13020100) and the Major Research plan of the National Natural Science Foundation of China (Grant No. 91131014).

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Axelsson E, et al. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495: 360-364.

Baird NA, Turnbull DW, Johnson EA 2006. Induction of the heat shock pathway during hypoxia requires regulation of heat shock factor by hypoxia-inducible factor-1. Journal of Biological Chemistry 281: 38675-38681.

Bakewell MA, Shi P, Zhang J 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci U S A 104: 7489-7494.

Beall CM 2007. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. Proc Natl Acad Sci U S A 104 Suppl 1: 8655-8660.

Beall CM, et al. 2010. Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci U S A 107: 11459-11464.

Bickler PE, Buck LT 2007. Hypoxia tolerance in reptiles, amphibians, and fishes: life with variable oxygen availability. Annual Review of Physiology 69: 145-170.

Bigham A, et al. 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. Plos Genetics 6: e1001116.

Cai Q, et al. 2013. Genome sequence of ground tit Pseudopodoces humilis and its adaptation to high altitude. Genome Biology 14: R29.

Cao W, Chen Y, Wu Y, Zhu S 1981. Origin and evolution of schizothoracine fishes in relation to the upheaval of the Xizang Plateau. Studies on the period, amplitude and type of the uplift of the Qinghai-Xizang Plateau: (ed. by Tibetan Expedition Team of the Chinese Academy of Science), pp. 118–130 (in Chinese). Science Press, Beijing.

Caramuru LH, Soares Rde P, Maeda NY, Lopes AA 2003. Hypoxia and altered platelet behavior influence von Willebrand factor multimeric composition in secondary pulmonary hypertension. Clin Appl Thromb Hemost 9: 251-258.

Castresana J 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17: 540-552.

Chen S, Krinsky BH, Long M 2013. New genes as drivers of phenotypic evolution. Nat Rev Genet 14: 645-660.

Chen Y, Cao W 2000. Schizothoracinae. Fauna Sinica,Osteichthyes, Cypriniformes III (ed. by P. Yue): 273–335 (in Chinese). Science Press, Beijing.

Dennis G, Jr., et al. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.

Ding Y, Zhou Q, Wang W 2012. Origins of New Genes and Evolution of Their Novel Functions. Annual Review of Ecology, Evolution, and Systematics 43: 345-363.

Durinck S, et al. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439-3440.

Ellegren H, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756-760.

Fang D, et al. 2008. Expression of bystin in reactive astrocytes induced by ischemia/reperfusion and chemical hypoxia in vitro. Biochimica Et Biophysica Acta-Molecular Basis of Disease 1782: 658-663.

Ferreira PG, et al. 2013. Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. Genome Biol 14: R20.

Flicek P, et al. 2013. Ensembl 2013. Nucleic Acids Res 41: D48-55.

Gaudreault E, Thompson C, Stankova J, Rola-Pleszczynski M 2005. Involvement of BLT1 endocytosis

and Yes kinase activation in leukotriene B4-induced neutrophil degranulation. J Immunol 174: 3617-3625.

Ge RL, et al. 2013. Draft genome sequence of the Tibetan antelope. Nat Commun 4: 1858.

Goodman M, et al. 2009. Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries. Proc Natl Acad Sci U S A 106: 20824-20829.

Gotz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36: 3420-3435.

Gou X, et al. 2014. Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. Genome Research 24: 1308-1315.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-652.

Guo B, Chain FJ, Bornberg-Bauer E, Leder EH, Merila J 2013. Genomic divergence between nine- and three-spined sticklebacks. BMC Genomics 14: 756.

Guo B, Zou M, Wagner A 2012. Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication. Mol Biol Evol 29: 3005-3022.

Hahn MW, Demuth JP, Han SG 2007. Accelerated rate of gene gain and loss in primates. Genetics 177: 1941-1949.

Hansen RK, Parra I, Hilsenbeck SG, Himelstein B, Fuqua SAW 2001. Hsp27-induced MMP-9 expression is influenced by the Src tyrosine protein kinase Yes. Biochemical and Biophysical Research Communications 282: 186-193.

He DK, Chen YF 2006. Biogeography and molecular phylogeny of the genus Schizothorax (Teleostei : Cyprinidae) in China inferred from cytochrome b sequences. Journal of Biogeography 33: 1448-1460.

He DK, Chen YF, Chen YY, Chen ZM 2004. Molecular phylogeny of the specialized schizothoracine fishes (Teleostei : Cyprinidae), with their implications for the uplift of the Qinghai-Tibetan Plateau. Chinese Science Bulletin 49: 39-48.

Huang da W, Sherman BT, Lempicki RA 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

Jones FC, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55-61.

Kent WJ 2002. BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.

Khurana P, Sugadev R, Jain J, Singh SB 2013. HypoxiaDB: a database of hypoxia-regulated proteins. Database (Oxford) 2013: bat074.

Li W, Godzik A 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.

Li Y, et al. 2014. Population variation revealed high-altitude adaptation of tibetan mastiffs. Molecular Biology and Evolution 31: 1200-1205.

Loytynoja A, Goldman N 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 102: 10557-10562.

Lynch G, et al. 2004. The tyrosine kinase Yes regulates actin structure and secretion during pancreatic acinar cell damage in rats. Pflugers Arch 447: 445-451.

Mojiri A, et al. 2013. Hypoxia results in upregulation and de novo activation of von Willebrand factor expression in lung endothelial cells. Arterioscler Thromb Vasc Biol 33: 1329-1338.

Near TJ, et al. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A 109: 13698-13703.

Peng Y, et al. 2011. Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. Molecular Biology and Evolution 28: 1075-1081.

Poelstra JW, et al. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science 344: 1410-1414.

Qiu Q, et al. 2012. The yak genome and adaptation to life at high altitude. Nat Genet 44: 946-949.

Qu Y, et al. 2013. Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. Nat Commun 4: 2071.

Rice P, Longden I, Bleasby A 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.

Schunter C, Vollmer SV, Macpherson E, Pascual M 2014. Transcriptome analyses and differential gene expression in a non-model fish species with alternative mating tactics. BMC Genomics 15: 167.

Simonson TS, et al. 2010. Genetic evidence for high-altitude adaptation in Tibet. Science 329: 72-75.

Smith NG, Eyre-Walker A 2002. Adaptive protein evolution in Drosophila. Nature 415: 1022-1024.

Soria-Carrasco V, et al. 2014. Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. Science 344: 738-742.

Steinke D, Salzburger W, Meyer A 2006. Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. Journal of Molecular Evolution 62: 772-784.

Stern DL 2013. The genetic causes of convergent evolution. Nature Reviews Genetics 14: 751-764.

Storey JD, Tibshirani R 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-9445.

Su J, et al. 2014. Genetic Structure and Demographic History of the Endangered and Endemic Schizothoracine Fish Gymnodiptychus pachycheilus in Qinghai-Tibetan Plateau. Zoolog Sci 31: 515-522.

Sun Q, Zhou H, Binmadi NO, Basile JR 2009. Hypoxia-inducible factor-1-mediated regulation of semaphorin 4D affects tumor growth and vascularity. Journal of Biological Chemistry 284: 32066-32074.

Thompson LG, et al. 2000. A high-resolution millennial record of the south asian monsoon from himalayan ice cores. Science 289: 1916-1920.

Wang L, et al. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res 41: e74.

Wang S, Xie Y 2004. China species red list. Beijing, China: Higher Education Press.

Whitlock NA, Agarwal N, Ma JX, Crosson CE 2005. Hsp27 upregulation by HIF-1 signaling offers protection against retinal ischemia in rats. Investigative Ophthalmology & Visual Science 46: 1092-1098.

Wu B, Deng Y, Wu J, Yan C, Song Z 2014. Complete mitochondrial genome of Gymnodiptychus pachycheilus (Teleostei: Cypriniformes: Cyprinidae). Mitochondrial DNA.

Wu Y, Wu C 1991. The fishes of the Qinghai - Xizang plateau. Chengdu: Sichuan Science and Technology Press.

Xing J, et al. 2013. Genomic analysis of natural selection and phenotypic variation in high-altitude mongolians. Plos Genetics 9: e1003634.

Xu S, et al. 2011. A genome-wide search for signals of high-altitude adaptation in Tibetans. Molecular Biology and Evolution 28: 1003-1011.

Yang Z 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.

Yi X, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:

75-78.
Zhang J, Nielsen R, Yang Z 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472-2479.
Zhang W, et al. 2014. Hypoxia Adaptations in the Grey Wolf (Canis lupus chanco) from Qinghai-Tibet Plateau. Plos Genetics 10: e1004466.
Zhao S, et al. 2013. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nat Genet 45: 67-71.

Data Accessibility

The sequencing reads were submitted through the NCBI SRA and can be accessible via NCBI BioProject accession SRP047392.

Tables

Table 1 Positively selected genes involved in hypoxia response in *G. pachycheilus*

Gene ID	Gene name	Description	Adjusted P-value
ENSDARG00000001057	BYSL	bystin-like	0.003
ENSDARG00000008818	HSF1	heat shock transcription factor 1	0.03
ENSDARG00000005941	YES1	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1	0.04
ENSDARG00000020761	ARRDC2	arrestin domain containing 2	0.02
ENSDARG00000041747	SSPN	sarcospan (Kras oncogene-associated gene)	0.009
ENSDARG00000067801	SEMA4D	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4D	0.01
ENSDARG00000077231	VWF	von Willebrand factor	0.04
ENSDARG00000053865	COMP	cartilage oligomeric matrix protein	0.03
ENSDARG00000059369	LAMB	laminin, beta 3	0.001

Figure legends

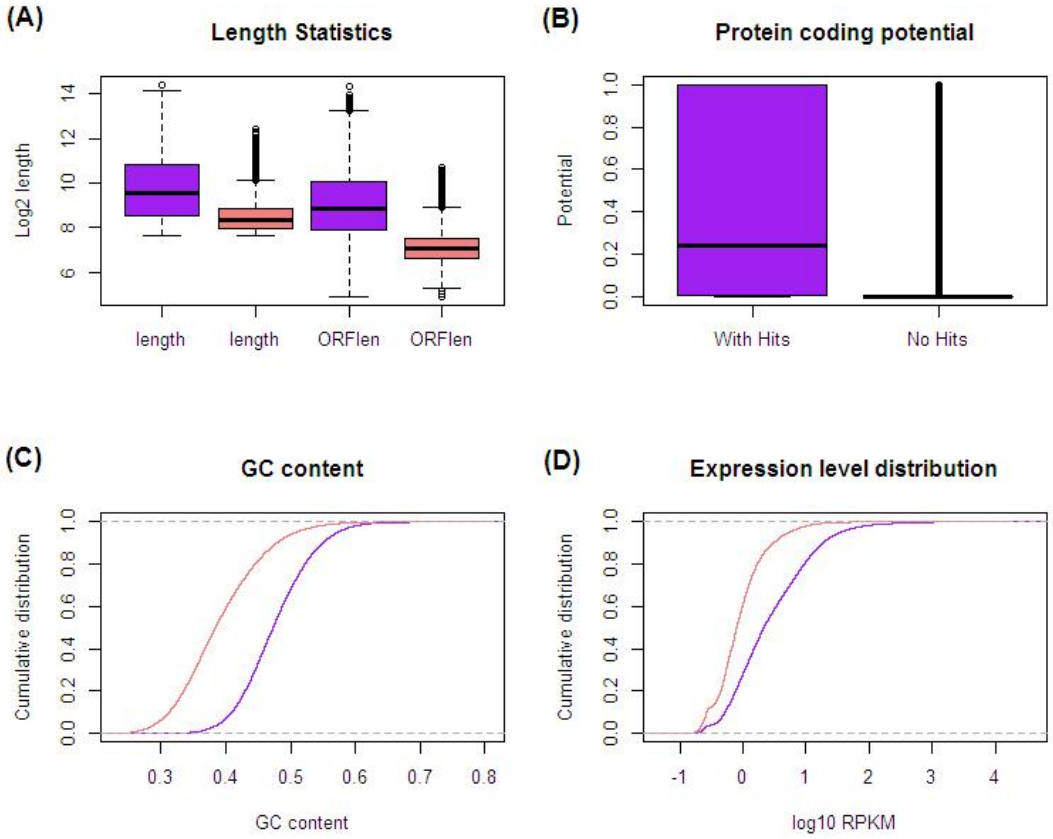

Fig. 1 Comparison between the set of unigenes with and without BLAST-hits. (A) Overall length and longest ORF length statistics, (B) Protein coding potential determined by CPAT, (C) Distribution of GC content, (D) Distribution of normalized expression level. RPKM, reads per kilobase per million mapped. Purple color: with BLAST hit, pink: without BLAST hit.

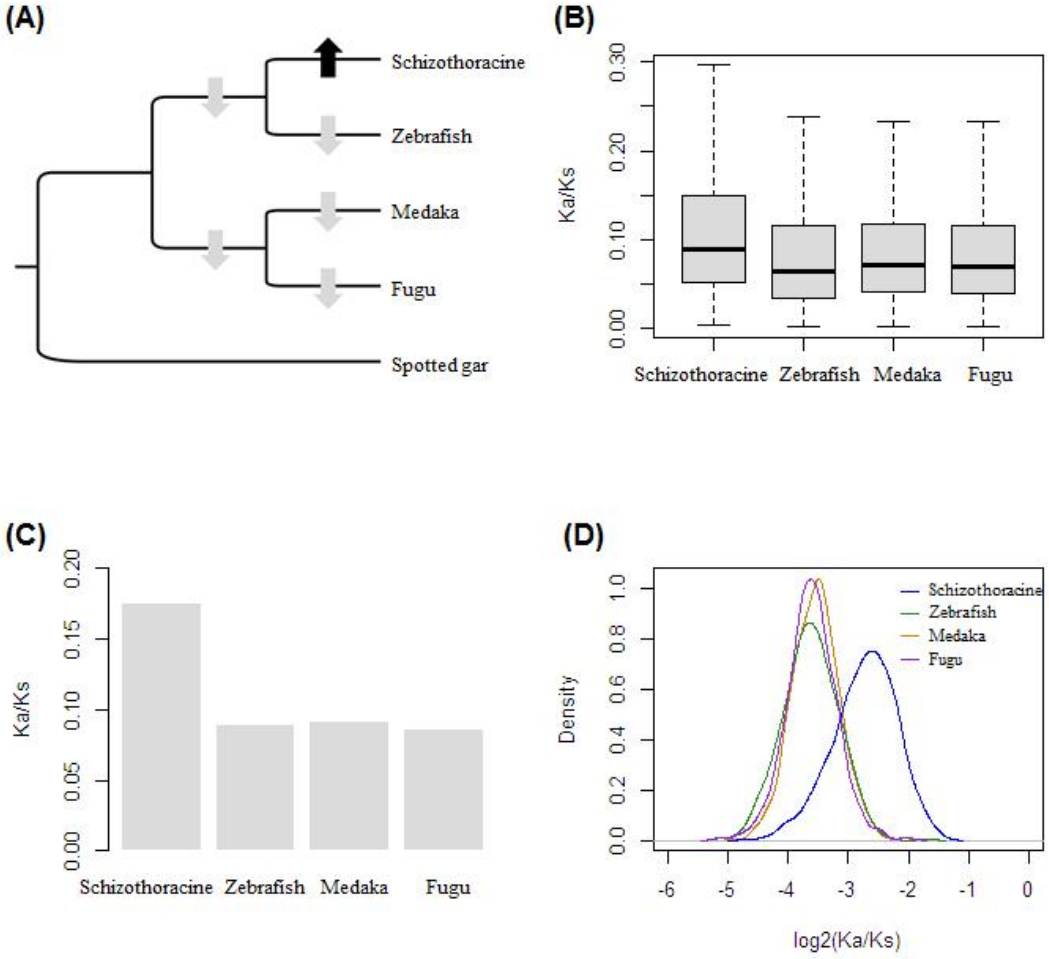

Fig. 2 Phylogenetic tree used in this study (A) and branch specific Ka/Ks ratios obtained from different datasets (B, C, D). Gray and black arrows in (A) indicate decreased or increased terminal Ka/Ks ratios compared with the ancestral branch. The Ka/Ks ratios for terminal branches were estimated from each ortholog (B), concatenated all orthologs (C), and 1,000 concatenated alignments constructed from 10 randomly chosen orthologs (D).

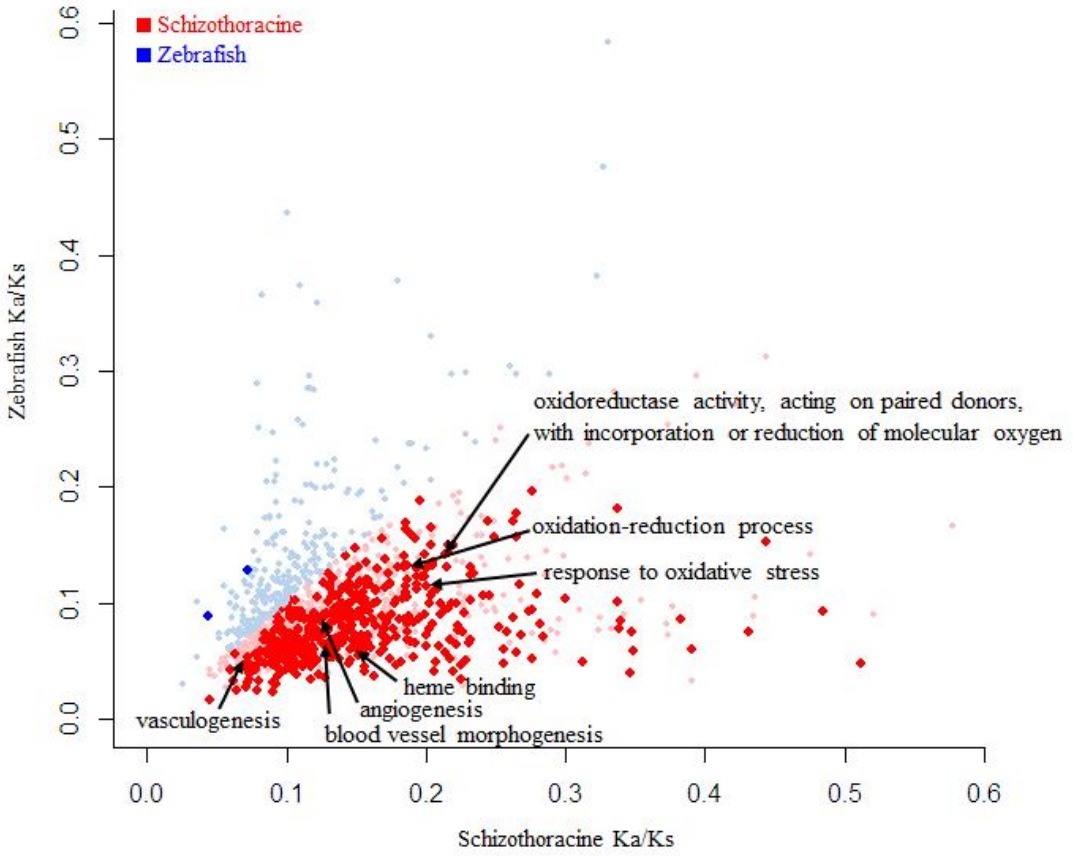

Fig. 3 Scatter plot of mean Ka/Ks ratios for each GO category in *G. pachycheilus* and zebrafish. GO categories with significantly higher mean Ka/Ks ratios in *G. pachycheilus* (red) and zebrafish (blue) are highlighted. Light red and light blue points represent the GO categories with higher but not statistically significant mean Ka/Ks ratios in *G. pachycheilus* and zebrafish.

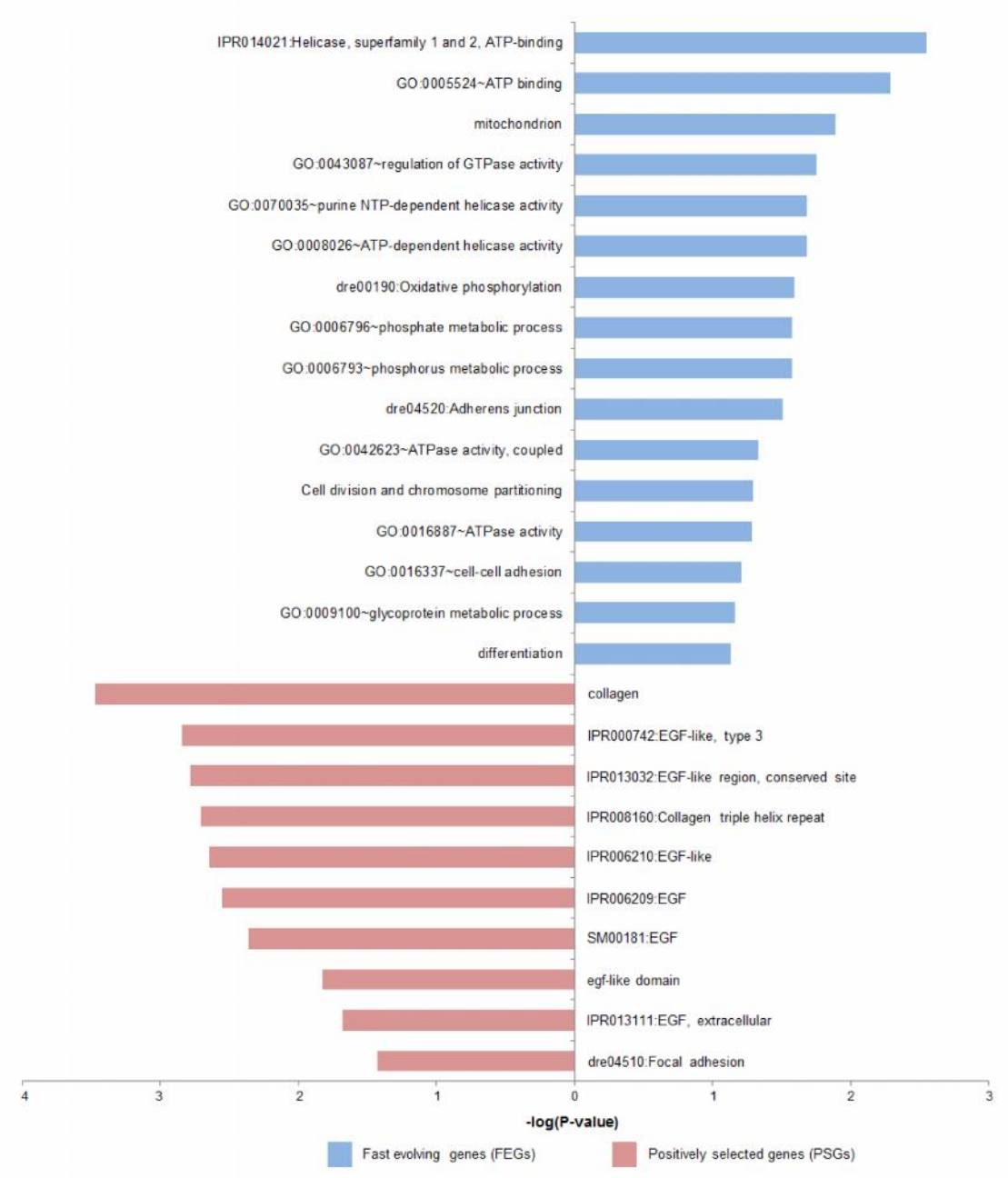

Fig. 4 Functional enrichment analyses of fast evolving genes and positively selected genes showing categories involved in energy metabolism and hypoxia response. (Blue) Fast evolving genes; (Red) Positively selected genes.