OPEN ACCESS Freely available online

PLOS ONE

Transcriptome Sequencing and Analysis of Wild Amur Ide (*Leuciscus waleckii*) Inhabiting an Extreme Alkaline-Saline Lake Reveals Insights into Stress Adaptation

Jian Xu[1], Peifeng Ji[1], Baosen Wang[1,2], Lan Zhao[1], Jian Wang[1], Zixia Zhao[1], Yan Zhang[1], Jiongtang Li[1], Peng Xu[1]*, Xiaowen Sun[1]*

1 Centre for Applied Aquatic Genomics, Chinese Academy of Fishery Sciences, Beijing, China, **2** College of Life Sciences, Tianjin Normal University, Tianjin, China

Abstract

Background: Amur ide (*Leuciscus waleckii*) is an economically and ecologically important species in Northern Asia. The Dali Nor population inhabiting Dali Nor Lake, a typical saline-alkaline lake in Inner Mongolia, is well-known for its adaptation to extremely high alkalinity. Genome information is needed for conservation and aquaculture purposes, as well as to gain further understanding into the genetics of stress tolerance. The objective of the study is to sequence the transcriptome and obtain a well-assembled transcriptome of Amur ide.

Results: The transcriptome of Amur ide was sequenced using the Illumina platform and assembled into 53,632 cDNA contigs, with an average length of 647 bp and a N50 length of 1,094 bp. A total of 19,338 unique proteins were identified, and gene ontology and KEGG (Kyoto Encyclopedia of Genes and Genomes) analyses classified all contigs into functional categories. Open Reading Frames (ORFs) were detected from 34,888 (65.1%) of contigs with an average length of 577 bp, while 9,638 full-length cDNAs were identified. Comparative analyses revealed that 31,790 (59.3%) contigs have a significant similarity to zebrafish proteins, and 27,096 (50.5%), 27,524 (51.3%) and 27,996 (52.2%) to teraodon, medaka and three-spined stickleback proteins, respectively. A total of 10,395 microsatellites and 34,299 SNPs were identified and classified. A dN/dS analysis on unigenes was performed, which identified that 61 of the genes were under strong positive selection. Most of the genes are associated with stress adaptation and immunity, suggesting that the extreme alkaline-saline environment resulted in fast evolution of certain genes.

Conclusions: The transcriptome of Amur ide had been deeply sequenced, assembled and characterized, providing a valuable resource for a better understanding of the Amur ide genome. The transcriptome data will facilitate future functional studies on the Amur ide genome, as well as provide insight into potential mechanisms for adaptation to an extreme alkaline-saline environment.

Citation: Xu J, Ji P, Wang B, Zhao L, Wang J, et al. (2013) Transcriptome Sequencing and Analysis of Wild Amur Ide (*Leuciscus waleckii*) Inhabiting an Extreme Alkaline-Saline Lake Reveals Insights into Stress Adaptation. PLoS ONE 8(4): e59703. doi:10.1371/journal.pone.0059703

Editor: Zhanjiang Liu, Auburn University, United States of America

Received December 20, 2012; **Accepted** February 17, 2013; **Published** April 1, 2013

Funding: This study was supported by the grants from National Natural Science Foundation of China (No. 31101893), National Department Public Benefit Research Foundation (No. 200903045), National High-tech R&D Program of China (2011AA100401). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: xupeng@cafs.ac.cn (PX); sunxw2002@163.com (XS)

Introduction

Amur ide (*Leuciscus waleckii*) is a species of cyprinid fish, inhabiting the Amur River basin in Russia, Mongolia, China and Korea (www.fishbase.org). There are many local populations inhibiting lakes and watersheds in Northern Asia. One of the most renowned populations is the Dali Nor population which inhabits Dali Nor Lake, Inner Mongolia (E 116°25′–116°45′,N43°13′–43°23′). Dali Nor Lake is a typical saline-alkaline lake with high concentrations of carbonate salts. It is located in a basin of the eastern Inner Mongolia Plateau where outflow is completely prevented (an endothecia basin), and the evaporation is greater than precipitation and inflow. Thus, the lake has been shrinking consistently since the early Holocene (11,500–7,600 cal yr BP), and the alkalinity and salinity are increasing steadily [1]. Currently pH values range from 8.25 to 9.6, the alkali content (ALK) is over 50 mg/L, and salinity is ~6 ‰. There are only two fish species, Amur ide and crucian carp, that have adapted to the extreme conditions and are able to survive in this harsh environment. Amur ide is economically important to the local Mongolian people who have lived around Dali Nor Lake for generations. Selective breeding program on Amur ide has been recently initiated on both growth and alkaline tolerance. The strain with better performance would be important for the rural region with large saline-alkaline open water in northern China. Besides, it is ecologically important fish to migrating birds that stop over at Dali Nor Lake and feed on Amur ide during their journeys from Siberia to the south [2].

In spite of its economic and ecological importance, the mechanism of Amur ide's tolerance of the high salinity and alkalinity of Dali Nor Lake is still unclear. Very limited physiological and genetic studies have been performed, and

information on its genetic resources has been only minimally developed. To date, only a few genetic markers have been developed [3]. The genetic diversity and population structure have been investigated by using microsatellite markers of Amur ide and grass carp. This study revealed that the Dali Nor population of Amur ide is genetically distinct from the Ussuri River population of Amur ide, suggesting that the Dali Nor population was isolated geologically a long time ago [4]. The mitochondrial genome of Amur ide from Dali Nor Lake has been completely sequenced and annotated, providing some basic molecular tools for ecological and genetic studies [5]. Amur ide has been recently developed as a potential aquaculture species in the widely distributed saline and alkaline waters of northern China because of its high tolerance to increased salinity and alkalinity. To better understand the physiological and genetic basis of its salinity tolerance, to explore adaptive evolution in its genome, and to support prospective genetic breeding and stock improvement, more knowledge about genetics and genome resources are needed.

EST (Expressed sequence tag) sequencing has been considered an efficient approach for genomic study and functional gene identification, especially for those species without a genome reference. In the past decade, tens of thousands ESTs have been developed for several important aquaculture species using traditional Sanger's methods (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html), including catfish (500,000 ESTs) [6], Atlantic salmon (498,212 ESTs), rainbow trout (287,967 ESTs), as well as aquatic parasite species such as *Ichthyophthirius multifiliis* (33,516 ESTs) [7]. These EST resources allow efficient gene discovery and transcriptome profiling in these species [8–12], as well as genetic marker development [13]. In the past half-decade, high-throughput next-generation sequencing technologies have been developed and successfully exploited to obtain a vast amount of transcriptome sequences at a lower cost, providing scientists the ability to collect sufficient genetic and genome resources for the many different species [14–16].

In the present study, we sequenced the transcriptome of Amur ide with the Illumina sequencing platform. The transcriptome sequences were assembled into contigs. Function annotation and gene ontology analyses were performed, and a large amount of microsatellite and SNP loci were identified. Synonymous and non-synonymous sites were analyzed to identify those genes that may be under strong positive selection in extreme environment. It is the first high throughput data for the genus *Leuciscus*, which provides a valuable resource for unveiling the mechanism of alkaline tolerance, as well as facilitating the genetic improvement and conservation of Amur ide.

Results and Discussion

Transcriptome Sequencing and Assembly of Amur Ide

To enable a comprehensive understanding and profiling of the transcritpome of Amur ide, mixed RNA originating from 12 tissues was sequenced by using Solexa HiSeq2000 sequencing technology. A total of 99,883,236 paired-end reads were generated with a read length of 101 bp. After the removal of ambiguous nucleotides and low-quality sequences (Phred quality scores <20), a total of 87,740,916 cleaned reads (87.8%) were obtained. The raw transcriptome sequences in this study have been deposited in the NCBI SRA database (Accession number SRR677015). The cleaned reads were then assembled by using Trinity assembler [17]. As shown in Table 1, the trancriptome was assembled, combining 73,668,103 reads into 53,632 contigs, ranging from 107 to 9,691 bp in length. The average length was 647 bp, the N50 length was 1,094 bp and the median length was 356 bp. The contig length distribution is shown in Figure 1.

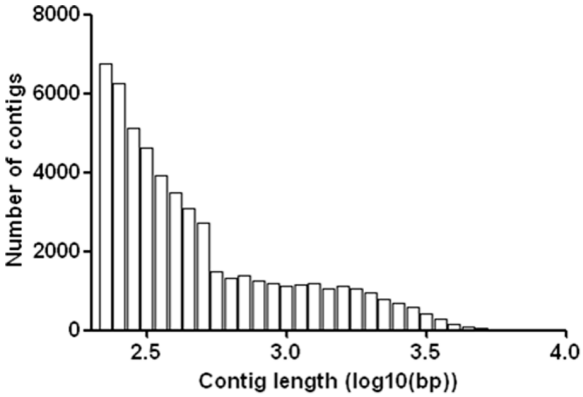

Figure 1. Length distribution of assembled contigs of Amur ide.
doi:10.1371/journal.pone.0059703.g001

To validate the assembly, 34 contigs were randomly selected. Primers were designed and transcripts were amplified from a cDNA template of pooled tissues of Amur ide. Thirty-one (91.2%) contigs were successfully amplified with the expected length (Table S1), which confirmed that the assembly using the Trinity algorithm was reliable.

Functional Annotation

All assembled contigs were first compared with the NCBI non-redundant (nr) protein database for functional annotation by using BLASTx with an e-value cutoff of 1e−10. A total of 30,866 contigs had a significant hit, corresponding to 19,338 unique protein accessions in the nr protein database (Table 2). The gene name of the top BLASTx hit was assigned to each of the contigs with significant hits.

We also calculated the "ortholog hit ratio" [18] by dividing the length of the putative coding region of a unigene by the total length of the ortholog found for that unigene (Figure 2). Each unigene and its best BLASTx hit were considered orthologs and the hit region in the unigene was considered to be a "putative

Table 1. Statistics of transcriptome sequencing, assembly and annotation of Amur ide.

Stage		
Sequencing	Number of reads (101-bp paired-end)	99,883,236
	Total bases	9.99 Gb
	Cleaned reads	87,740,916
Assembly	Number of contigs	53,632
	Maximum contig length	9,691 bp
	Minimum contig length	107 bp
	Average contig length	647 bp
	N50 length	1,094 bp
Annotation	Contigs with blast results	30,866
	Unigenes with blast results	19,338
	Contigs with GO terms	13,717
	Unigenes with GO terms	10,674

doi:10.1371/journal.pone.0059703.t001

Table 2. Summary of BLASTX search results of Amur ide transcriptome.

Database	Amur ide hits	Unique protein	% of total unique proteins
NR	30,866	19,338	
UTR	3,822	2,497	
Refseq/Ensembl			
Zebrafish	31,790	15,759	57.8% of 27,271
Medaka	27,524	13,419	54.4% of 24,661
Tetraodon	27,096	12,952	56.0% of 23,118
Three-spined stickleback	27,996	14,047	50.9% of 27,576

doi:10.1371/journal.pone.0059703.t002

coding region". Thus, the ortholog hit ratio gives an estimate of the amount of a transcript that is represented by each unigene [19]. Figure 2(a) shows that the completeness of the assembled transcripts only slightly decreases for longer genes, suggesting a high assembly quality in this study, even for long transcripts. The distribution of ortholog hit ratios is represented in Figure 2(b). Most unigenes with BLASTx results had high ortholog hit ratios, indicating a high completeness of these transcripts. Of the 30,866 transcripts with BLASTx results, 14,451 contigs had a ratio $\geq$0.9 and 26,074 had a ratio $\geq$0.5. In total only 827 (2.7%) contigs had a ratio greater than 1.0, indicating the possibility of insertions in relative unigenes. The evaluation also suggests that a higher proportion of full length transcripts were in the assembly.

Gene ontology (GO) analysis was conducted on those 19,338 unique proteins by using InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) and integrated protein databases with default parameters. A total of 10,674 unique proteins were assigned at least one GO term for describing biological processes, molecular functions and cellular components. The InterProScan output file was input into the BGI WEGO program and GO annotations were plotted (http://wego.genomics.org.cn) (Figure 3). Of these, the molecular function ontology made up the majority (9,429, 88.3%), followed by the biological process ontology (6,959, 65.2%) and the cellular component ontology (5,029, 47.1%). Briefly, for biological processes, genes involved in cellular processes (GO: 0009987) and metabolic processes (GO: 0008152) were highly represented; for molecular functions, binding (GO: 0005488) was the most represented GO term, followed by catalytic activity (GO: 0003824); cells (GO: 0005623) and organelles (GO: 0043226) were the most represented categories for the cellular component. Interestingly, within biological processes, a total of 303 unigenes were annotated to response to stimulus (GO: 0050896), including 190 unigenes to response to stress (GO: 0006950), 86 unigenes to cellular response to stimulus (GO: 0051716), 50 unigenes to response to external stimulus (GO: 0009605), 35 unigenes to response to chemical stimulus (GO: 0042221), and 4 unigenes to detection of stimulus (GO: 0051606). In previous studies, the GO result of common carp was reported, of which 250 unigenes were annotated to response to stimulus [20]. The expression of more stimulus-related genes in Amur ide from Dali Nor Lake is consistent with the extreme environmental stress of this habitat.

In addition, a KEGG pathway analysis was performed on all assembled contigs as an alternative approach for functional categorization and annotation. Enzyme commission (EC) numbers were assigned to 6,174 unique sequences, which categorized them into different functional groups (Table 3). Briefly, of these sequences with KEGG annotation, 1,975 (32.0%) were classified into metabolism, including major subgroups of carbohydrate metabolism (826 sequences), lipid metabolism (207 sequences) and amino acid metabolism (188 sequences). Sequences grouped into genetic information processing (GIP), accounted for 1,019 (16.5%), including translation (367 sequences), folding, sorting and degradation (366 sequences), transcription (176 sequences), and replication and repair (110 sequences). Organismal systems, cellular processes and environmental information processing (EIP) groups con-

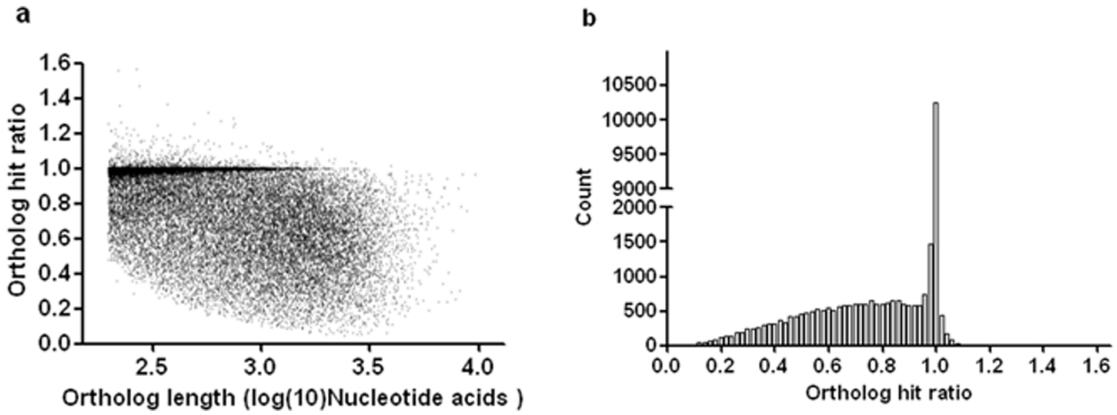

Figure 2. Distribution of ortholog hit ratio and its relationship with ortholog length. Ortholog hit ratios were calculated for contigs with BLASTx results. A ratio of 1.0 indicates the gene is likely fully assembled.
doi:10.1371/journal.pone.0059703.g002

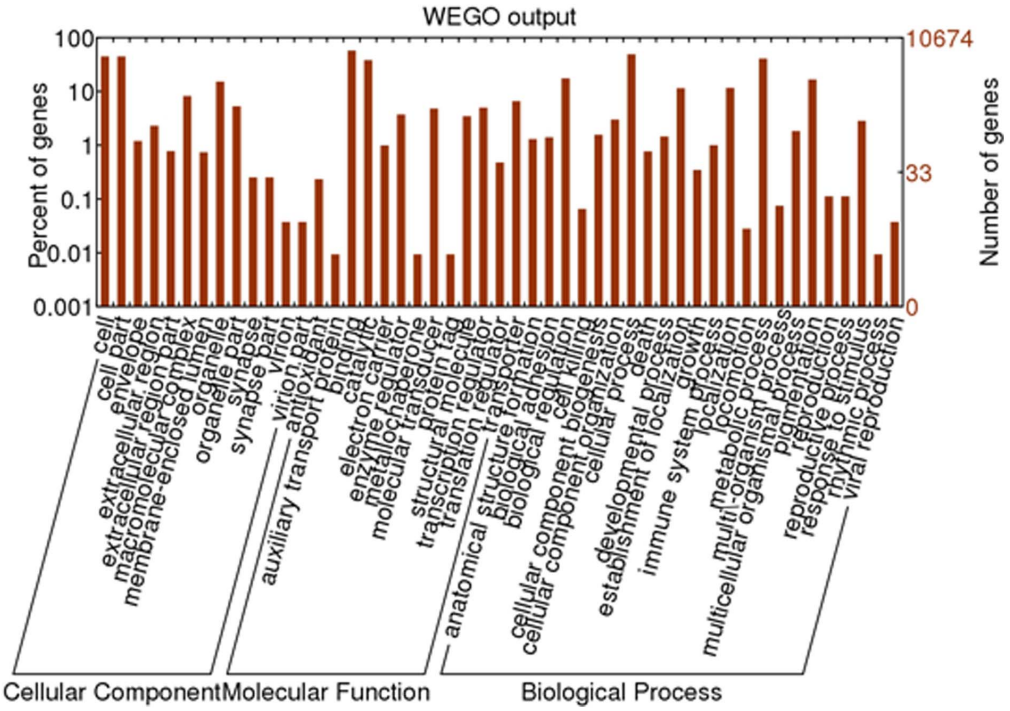

Figure 3. Gene Ontology (GO) categories of the unigenes. Distribution of the GO categories assigned to the Amur ide transcriptome. Unique transcripts (unigenes) were annotated in three categories: cellular components, molecular functions, and biological processes.
doi:10.1371/journal.pone.0059703.g003

tained 1,189(19.2%), 836(13.5%) and 538 (8.7%) KEGG-annotated sequences, respectively. A well-categorized and annotated transcriptome could serve as an important and valuable resource for gene identification and functional analysis of alkaline tolerance and osmotic regulation in Amur ide. For instance, 17 transcript contigs associated with osmotic regulation in KEGG analysis have been identified and collected (Table S2).

Of the 53,632 assembled contigs of the Amur ide transcriptome, Open Reading Frames (ORFs) were detected from 34,888 (65.0%) contigs, with an average ORF length of 577 bp and with a range from 51 bp to 9,339 bp (Figure 4). The remaining 18,744 contigs contained no ORFs, indicating they are non-coding sequences and likely come from untranslated regions (UTR). The assembled transcriptome contigs served as a reference for cSNPs identification from RNA-seq data. ORF analysis allows the discrimination of synonymous and non-synonymous SNPs and the identification of nonsense mutations in Amur ide.

Assessment of Transcriptome Assembly

The assembled contigs of Amur ide transcriptome were compared with Refseq proteins of zebrafish, fugu (*Takifugu rubripes*), medaka *(Oryzias latipes*) and three-spined stickleback (*Gasterosteus aculeatus*) by using the BLASTx program with an e-value cutoff of 1E−10. There were 31,790 contigs (59.3%) with 15,759 unique protein hits, 27,096 contigs (50.5%) with 12,952 unique protein hits, 27,524 contigs (51.3%) with 13,419 unique protein hits and 27,996 contigs (52.2%) with 17,047 unique protein hits identified with significant hits on Refseq proteins of zebrafish, tetraodon, medaka and three-spined stickleback, respectively. The contigs of the Amur ide transcriptome had hits with 50.9% to 57.8% of the unique proteins of zebrafish, tetraodon, medaka and three-spined stickleback (Table 2). Obviously, the closely related zebrafish showed the highest similarity to Amur ide at the gene expression level. However, transcriptome similarity was still relatively lower than expected. Both zebrafish and Amur ide belong to Cyprinidae but are found in different distant subfamilies according to phylogenetic research [21]. It may be that they only share a limited level of similarity. The current research however does not cover the whole transcriptome as the genetic material in this study was only collected from adult fish; transcripts from early development stages were therefore not included. In addition, some rare transcripts may be missed or were only collected as singletons during the assembly, even though a high sequencing depth was applied. For better understanding and characterization of Amur ide transcriptome, we would need a complete set of transcriptome data from virtually every tissue across every life stage and every circumstance, or the whole-genome sequencing and assembly.

Full-length cDNA Prediction

Full-length cDNAs are important resources for many genetic and genomic studies, including gene duplication analysis, alternative splicing and whole genome sequencing and assembly. To identify potential full-length cDNAs with complete ORF in the assembled transcriptome of Amur ide, all contigs were analyzed by the online tool of TargetIdentifier. A total of 9,638 full-length and ORF completely-sequenced sequences were identified from the assembly with a cutoff E-value of 1E−5, with sequence lengths from 201 bp to 8,204 bp (Figure 5). Most of the identified full-length cDNA sequences were shorter than 1.5 kb, suggesting those long full-length cDNA sequences were not easily assembled using only the current set of transcriptome data. The current data may need to be combined with a traditional full-length cDNA library

Table 3. KEGG biochemical mappings for Amur ide.

KEGG categories represented	Unique sequences* (Number of KO)
Metabolism	**1,975 (1,415)**
Carbohydrate Metabolism	826 (597)
Amino Acid Metabolism	188 (140)
Energy Metabolism	174 (122)
Nucleotide Metabolism	142 (96)
Metabolism of Cofactors and Vitamins	114 (82)
Lipid Metabolism	207 (152)
Glycan Biosynthesis and Metabolism	132 (101)
Metabolism of Other Amino Acids	78 (50)
Xenobiotics Biodegradation and Metabolism	78 (45)
Biosynthesis of Secondary Metabolites	15 (12)
Biosynthesis of Polyketides and Nonribosomal Peptides	21 (18)
Genetic Information Processing	**1,019 (812)**
Replication and Repair	110 (88)
Folding, Sorting and Degradation	366 (294)
Transcription	176 (146)
Translation	367 (284)
Environmental Information Processing	**538 (398)**
Signal Transduction	378 (275)
Signaling Molecules and Interaction	143 (107)
Membrane Transport	17 (16)
Cellular Processes	**836 (600)**
Cell Motility	104 (64)
Cell Growth and Death	178 (141)
Transport and Catabolism	348 (258)
Cell Communication	206 (137)
ORGANISMAL SYSTEMS	**1,189 (848)**
Immune System	318 (240)
Endocrine System	202 (149)
Development	80 (48)
Circulatory System	139 (98)
Digestive System	78 (56)
Excretory System	200 (134)
Nervous System	22 (15)
Sensory System	117 (85)
Environmental Adaptation	33 (23)
Total	**6,174 (4,073)**

*Unique sequences indicate non-redundant sequences involving particular KEGG category.

doi:10.1371/journal.pone.0059703.t003

and Sanger's sequencing method to collect more full-length cDNA sequences and to build a database.

Repetitive Element Analysis and Microsatellite Identification

A total of 10,395 microsatellites were initially identified from 8,447 contigs, including di-, tri-, tetra-, penta- and hexa-nucleotide repeats. After removing the microsatellites without enough flanking sequences (50 bp) for the design of primers, there were 4,120 unique sequences with microsatellites that can be used to design primers for genotyping (Table 4).

The proportion of the repetitive elements in the Amur ide genome was assessed by using Repeatmasker with Vertebrates Repeat Database. Repeatmasking of the 34,750,752 bp of the Amur ide contig sequences resulted in the detection of 890,553 bp (2.56%) of repeated sequences. The classification and respective proportion of the identified repetitive elements are shown in Table S1. The most abundant type of repetitive elements in the sequences were DNA transposons (0.99%), mostly hobo-Activator (0.40%), followed by retroelements (0.61%) including LINEs (0.27%), LTR elements (0.31%), and SINEs (0. 03%). Various satellite sequences, low complexity and simple sequence repeats

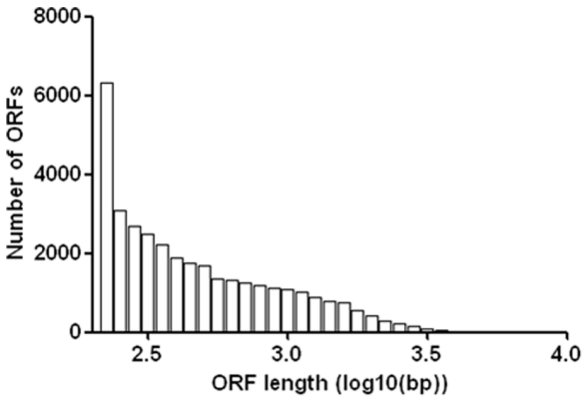

Figure 4. Length distribution of identified ORF.
doi:10.1371/journal.pone.0059703.g004

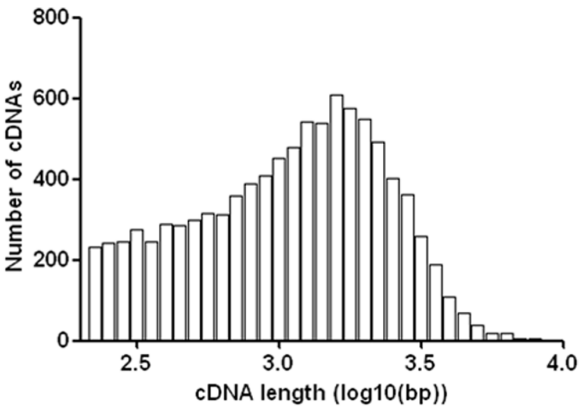

Figure 5. Length distribution of putative full-length cDNAs.
doi:10.1371/journal.pone.0059703.g005

accounted for 0.10%, 0.48% and 0.33% of the base pairs, respectively.

SNP Identification

For further application of the RNA-Seq data, SNPs were discovered using the assembled transcriptome. The short reads of RNA-Seq data were aligned onto the reference transcriptome of Amur ide and a total of 79,475,676 (79.6%) reads were mapped on the transcriptome, generating 34,299 SNPs after quality control and filtration (See Methods). The proportions of transition substitutions were 28.5% for A/G and 30.5% for C/T, compared with smaller proportions of transversion for A/C (11.2%), G/T (10.5%), A/T (11.8%) and C/G (7.5%). Among all SNPs detected, 10,408 were in the putative ORF region, of which 4, 335 were synonymous and 6,073 were non-synonymous. The mean number of SNPs per kilobase in the ORF region was 2.64. Further analysis was done to classify identified SNPs (Table 5).

Analysis of Synonymous and Non-synonymous Sites

The Amur ide population in Dali Nor Lake has been isolated from other populations for over 10,000 years since the early Holocene. The lake is consistently shrinking, and the alkalinity and salinity are also consistently increasing during this time. How could the Amur ide survive in such an extreme environment? Did the environmental changes speed up gene evolution to adapt to the changing environment? It is commonly known that positive selection likely plays a major role in shaping genetic architecture when populations are placed into new or changing environments [22–25]. Thus, we hypothesize that Amur ide experienced strong positive selection on a group of genes and pathways in response to extreme alkalinity and salinity, as well as concentrated heavy metal ions. Non-synonymous (dN) and synonymous (dS) substitution rates have been widely used to measure the intensity of gene evolution [26,27]. To identify genes undergoing strong positive selection, we estimated dN and dS rates of the assembled genes of Amur ide. A total of 2,646 unigenes which contained at least 1 SNP were used to calculate dN, dS and the dN/dS ratio. The results showed that the overall dN, dS and dN/dS were 0.002, 0.008 and 0.428, respectively, which indicated that most of the genes were not under positive selection. However, the dN/dS ratios of 61 unigenes were greater than 1, indicating strong positive selection did occur on them (Figure 6). The functions of these genes were further investigated either by a pathway analysis or a literature search. Interestingly, a suite of genes were clearly associated with stress adaptation and immunity [28–32], including carbonic anhydrase 4, superoxide dismutase, and glutathione S-transferase A (Table 6 and Table S3).

We looked into several typical genes which respond to stress. Carbonic anhydrase (CA) is a zinc metalloenzyme that catalyzes the hydration of CO_2 to provide H^+ and HCO_3^- for ion transport processes. CA is widely recognized as a response to environmental stress in eukaryotes; for instance, CA was induced in the gills of the euryhaline green crab, *Carcinus maenas*, when the crab was transferred between different salinities [33]; Increasing salinities or an alkaline shift could induce CA in green alga *Dunaliella salina* [34]; CA even plays a role in cold adaptation among Antarctic fish [35]. Glutathione S-transferases (GSTs) are a large family of detoxification enzymes contributing to the biotransformation of a wide variety of environmental xenobiotics. They are thought to play major roles in oxidative stress, being responsible for various stress tolerance in plants [36–38] and animals [39–41]. GST from wild soybean (*Glycine soja*) was transferred into tobacco, enhancing drought and salt tolerance in transgenic tobacco [42]. GST is even used as a stress biomarker in mollusc species to monitor fuel oil spills [43].

Superoxide dismutases (SOD) are enzymes that catalyze the disputation of superoxide into oxygen and hydrogen peroxide and are well known to respond to various environmental stresses in eukaryotes. Overexpressing chloroplastic Cu/Zn SOD in plants may increase resistance to oxidative stress [44]. In animals, high levels of Cu/Zn SOD were detected in spermatogonia, protecting from oxidative stress [45]. Similar to GST, SODs are also used as an important biomarker in aquatic organisms for monitoring toxic environmental pollutants [46,47].

Table 4. Statistics of microsatellites identified from Amur ide transcriptome.

Total number of contigs	**53,632**
Microsatellites identified	10,395
Di-nucleotide repeats	4,316
Tri-nucleotide repeats	749
Tetra-nucleotide repeats	40
Penta-nucleotide repeats	12
Number of contigs containing microsatellites	8,447
Number of microsatellites with sufficient flanking sequences	4,120

doi:10.1371/journal.pone.0059703.t004

Table 5. Classification of SNPs identified from Amur ide transcriptome.

SNP classification	Number of SNPs
5′ UTR	646
3′ UTR	5,159
Coding region	10,408
synonymous	4,335
non-synonymous	6,073
pre-terminated	265
skip-stop-codon	214
mis-sense	5,594
Undefined	18,086
Total	34,299

doi:10.1371/journal.pone.0059703.t005

The results of the dN/dS analysis revealed fast genome evolution in some genes probably in order to adapt the extreme environmental stress in Dali Nor Lake and confirmed our hypothesis that increasing salinity, alkalinity and heavy metal concentration likely resulted in powerful selective pressures on certain genes for new genotypes that were better suited these stressful conditions.

Conclusions

In this study, the transcriptome of Amur ide was sequenced using the HiSeq2000 platform with high coverage, and then *de novo* assembled and functionally annotated by comparing with exiting protein databases of closely related species. An ORF analysis was conducted and a large number of full length cDNA sequences have been identified. In addition, repetitive element analysis was conducted, and cDNA SSR and SNP loci were identified for future marker development and genetic analysis. Synonymous and non-synonymous sites were analyzed on unigenes, which revealed that the Amur ide population in Dali Nor Lake has experienced fast evolution to adapt the extreme alkaline-saline environment.

Methods

Ethics Statement

This study was approved by the Animal Care and Use committee of the Centre for Applied Aquatic Genomics at Chinese Academy of Fishery Sciences.

Biological Samples

Ten wild Amur ide were sampled at the north shore of Dali Nor Lake, Inner Mongolia, China on May 12, 2012. Twelve tissues, including brain, muscle, liver, intestine, blood, head kidney, trunk kidney, skin, gill, spleen, gonad and heart, were dissected and collected. Tissue samples were stored in RNAlater (Qiagen, Hilden, Germany), transported to the laboratory at room temperature, and then stored at −20°C prior to RNA extraction.

RNA Extraction

Total RNA was extracted from 12 tissues using the TRIZOL Kit (Invitrogen, Carlsbad, CA, USA) following manufacturer's instructions. RNA samples were then digested by DNase I to remove potential genomic DNA. Integrity and size distribution were checked with Bioanalyzer 2100 (Agilent technologies, Santa Clara, CA, USA).

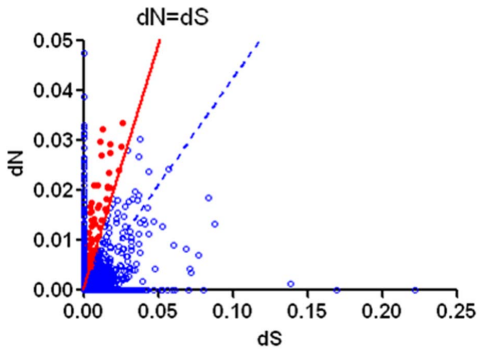

Figure 6. Distribution of SNP non-synonymous (dN) and synonymous (dS) substitution. The solid red line is the null expectation dN = dS. The filled red circles represent unigenes with dN/dS>1. The dashed blue line shows the slope (=0.428) of the overall average dN for all contigs/overall average dS for all contigs.
doi:10.1371/journal.pone.0059703.g006

Equal amounts of the high quality RNA samples from each tissue were then pooled for cDNA synthesis and sequencing.

cDNA Library Construction, Sequencing and Assembly

RNA-seq library preparation and sequencing was carried out by HudsonAlpha Genomic Services Lab (Huntsville, AL, USA) as previously described [13]. cDNA libraries were prepared with ~2.5 μg of starting total RNA following the protocols of the Illumina TruSeq RNA Sample Preparation Kit (Illumina). The final library had an average fragment size of ~270 bp and final yields of ~400 ng. After KAPA quantitation and dilution, the library was sequenced on an Illumina HiSeq 2000 with 101 bp paired-end reads. All sequenced reads in SRA format have been uploaded to the NCBI Short Read Archive with the accession number of SRR677015. Adaptor sequences were trimmed and reads with low quality or length less than 10 were further removed by SolexaQA software [48]. Cleaned reads were used for *de novo* assembly by TRINITY with default parameters.

Functional Annotation

The assembled transcriptome contigs were subjected to a similarity search against NCBI non-redundant (nr) protein database using BLASTx with an e-value cutoff of 1E−10. Gene names and descriptions were assigned to each contig based on the top BLASTx hit with the highest score. Gene ontology (GO) analysis was then conducted on the assembled transcriptome by using InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) and integrated protein databases with the default parameters. The GO terms associated with transcriptome contigs were then obtained to describe the genes in the areas of biological processes, molecular functions and cellular components. The InterProScan output file was input into BGI WEGO program and the GO annotations were plotted (http://wego.genomics.org.cn). All assembled contigs were analyzed by ESTScan to search for ORFs, which could distinguish between coding and non-coding sequences.

KEGG pathways were assigned to assembled contigs using the online KEGG Automatic Annotation Server (KAAS) (http://www.genome.jp/tools/kaas/) [49]. The Bi-directional Best Hit (BBH) method was used to obtain KEGG Orthology (KO) assignment.

Assembly Assessment

To compare the similarity to other teleost species, the transcriptome contigs were compared to Refseq and Ensemble proteins of zebrafish, fugu (*Takifugu rubripes*), medaka *(Oryzias latipes)*

Table 6. Unigenes showing positive selection (dN/dS>1) corresponding to stress adaption or immune response.

Unigenes	Uniprot ID	Description	dN/dS
contig7468	sp|P20702|ITAX_HUMAN	Integrin alpha-X	4.18
contig256	sp|P11364|TCB_FLV	T-cell receptor beta chain T17T-22	3.57
contig9830	sp|Q95118|IL2RG_BOVIN	Cytokine receptor common subunit gamma	3.46
contig344	sp|P49616|UPAR_RAT	Urokinase plasminogen activator surface receptor	3.30
contig1256	sp|P40189|IL6RB_HUMAN	Interleukin-6 receptor subunit beta	2.78
contig3025	sp|P11911|CD79A_MOUSE	B-cell antigen receptor complex-associated protein alpha chain	2.64
contig1148	sp|P48284|CAH4_RAT	Carbonic anhydrase 4	2.59
contig3210	sp|P08317|IL8_CHICK	Interleukin-8	2.59
contig6695	sp|P08294|SODE_HUMAN	Extracellular superoxide dismutase [Cu-Zn]	2.30
contig11187	sp|P01873|MUCM_MOUSE	Ig mu chain C region membrane-bound form	2.22
contig12382	sp|Q9NVE5|UBP40_HUMAN	Ubiquitin carboxyl-terminal hydrolase 40	1.96
contig13125	sp|P13387|EGFR_CHICK	Epidermal growth factor receptor	1.86
contig21350	sp|Q66S61|MBL2_CALJA	Mannose-binding protein C	1.77
contig941	sp|P10820|PERF_MOUSE	Perforin-1	1.74
contig59	sp|P06314|KV404_HUMAN	Ig kappa chain V-IV region B17	1.60
contig21332	sp|A7M9B2|YCF1_CUSRE	Putative membrane protein ycf1	1.55
contig21463	sp|Q6PIU2|NCEH1_HUMAN	Neutral cholesterol ester hydrolase 1	1.54
contig421	sp|P04114|APOB_HUMAN	Apolipoprotein B-100	1.54
contig11654	sp|P20759|IGHG1_RAT	Ig gamma-1 chain C region	1.41
contig255	sp|Q28085|CFAH_BOVIN	Complement factor H	1.32
contig895	sp|P30568|GSTA_PLEPL	Glutathione S-transferase A	1.31
contig2382	sp|Q8BK26|FBX44_MOUSE	F-box only protein 44	1.31
contig15120	sp|Q91009|NTRK1_CHICK	High affinity nerve growth factor receptor	1.27
contig7962	sp|Q9MZV7|CASP1_CANFA	Caspase-1	1.25
contig1236	sp|O95415|BRI3_HUMAN	Brain protein I3	1.18
contig10013	sp|Q80SU7|GVIN1_MOUSE	Interferon-induced very large GTPase 1	1.17
contig1492	sp|P19181|HV05_CARAU	Ig heavy chain V region 5A	1.11
contig1317	sp|P15684|AMPN_RAT	Aminopeptidase N	1.06
contig11745	sp|Q96G23|CERS2_HUMAN	Ceramide synthase 2	1.02
contig13871	sp|P29533|VCAM1_MOUSE	Vascular cell adhesion protein 1	1.02
contig3775	sp|P50283|CD7_MOUSE	T-cell antigen CD7	1.01
contig17067	sp|Q16787|LAMA3_HUMAN	Laminin subunit alpha-3	1.01

doi:10.1371/journal.pone.0059703.t006

and three-spined stickleback (*Gasterosteus aculeatus*), as well as the transcriptome of Amur ide by using the BLAST program with default parameters.

Full-length cDNA Identification

Putative full-length cDNAs were identified by using the online tool TargetIdentifier [6,50] and comparing to non-redundant protein databases with a cutoff e-value of 10^{-5}. The cDNA sequence was recognized as a full-length cDNA only if the start codon (ATG) and poly (A) tail were identified.

Repetitive Element Analysis and Microsatellite Identification

To identify all repetitive elements in the assembled transcriptome, RepeatMasker was used with Repbase for all vertebrates and zebrafish. A perl-based script Msatfinder V 2.0.9 [51] was used for microsatellite identification from assembled cDNA contigs. The mononucleotide repeats were ignored by modifying the configure file. The repeat thresholds for di-, tri-, tetra-, penta-, hexa-nucleotide motifs were set as 8, 5, 5, 5 and 5 respectively. Only microsatellite sequences with flanking sequences longer than 50 bp on both sides were identified for future marker development.

SNP Identification and dN/dS Analysis

To identify putative single nucleotide polymorphism (SNP) loci in the transcriptome of Amur ide, all RNA-Seq reads were mapped onto the assembled transcriptome using BWA and SAMtools. The filtering threshold was set as bellowing, the read depth to no less than 10, and the quality score to no less than 20. dN and dS were calculated by KaKs_Caculator 1.2 [52,53]. An input file (*.axt) for the KaKs_Caculator was generated from the ORF sequence and SNP list, which contains ID, reference sequence and sequence with SNPs for each gene. Then the output file was further extracted for useful information.

Supporting Information

Table S1 Validation of assembled contigs by PCR.
(XLS)

Table S2 Osmotic regulation related genes by KEGG analysis.
(XLS)

Table S3 All 61 unigenes putatively under positive selection.
(XLS)

Author Contributions

Conceived and designed the experiments: PX XS. Performed the experiments: JX PJ BW LZ. Analyzed the data: JX PJ PX JL. Contributed reagents/materials/analysis tools: JX PJ BW PX YZ ZZ JW. Wrote the paper: JX PX.

References

1. Xiao J, Si B, Zhai D, Itoh S, Lomtatidze Z (2008) Hydrology of Dali Lake in central-eastern Inner Mongolia and Holocene East Asian monsoon variability. Journal of Paleolimnology 40: 519–528.
2. Zhang J, Zhang Y, He J, Liu H, Zhu C (2008) Biology research of economical fish in Inner Mongolia Dalinor lake. Journal of Inner Mongolia Agricultural University 29: 197–200.
3. Liu J, Chang Y, Xu L, Liu C, Liang L, et al. (2011) Isolation and characterization of microsatellite from genome of *Leuciscus waleckii* Dybowski. Acta Agriculturae Boreali-Sinic 26: 87–93.
4. Chi B, Chang Y, Yan X, Cao D, Li Y, et al. (2010) Genetic variability and genetic structure of Leuciscus waleckii Dybowski in Wusuli River and Dali Lake. Journal of Fishery Sciences of China 17: 228–235.
5. Wang B, Ji P, Xu J, Sun J, Yang J, et al. (2012) Complete mitochondrial genome of Leuciscus waleckii (Cypriniformes: Cyprinidae: Leuciscus). Mitochondrial DNA.
6. Wang S, Peatman E, Abernathy J, Waldbieser G, Lindquist E, et al. (2010) Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. Genome Biology 11: R8.
7. Abernathy J, Xu P, Li P, Xu D-H, Kucuktas H, et al. (2007) Generation and analysis of expressed sequence tags from the ciliate protozoan parasite Ichthyophthirius multifiliis. BMC Genomics 8: 176.
8. Wang Q, Wang Y, Xu P, Liu Z (2006) NK-lysin of channel catfish: Gene triplication, sequence variation, and expression analysis. Molecular immunology 43: 1676–1686.
9. Hale M, Xu P, Scardina J, Wheeler P, Thorgaard G, et al. (2011) Differential gene expression in male and female rainbow trout embryos prior to the onset of gross morphological differentiation of the gonads. BMC Genomics 12: 404.
10. Xu P, McIntyre L, Scardina J, Wheeler P, Thorgaard G, et al. (2011) Transcriptome Profiling of Embryonic Development Rate in Rainbow Trout Advanced Backcross Introgression Lines. Marine Biotechnology 13: 215–231.
11. Nandi S, Peatman E, Xu P, Wang S, Li P, et al. (2007) Repeat structure of the catfish genome: a genomic and transcriptomic assessment of Tc1-like transposon elements in channel catfish (Ictalurus punctatus). Genetica 131: 81–90.
12. Sha Z, Xu P, Takano T, Liu H, Terhune J, et al. (2008) The warm temperature acclimation protein Wap65 as an immune response gene: its duplicates are differentially regulated by temperature and bacterial infections. Mol Immunol 45: 1458–1469.
13. Xu J, Ji P, Zhao Z, Zhang Y, Feng J, et al. (2012) Genome-wide SNP discovery from transcriptome of four common carp strains. PLoS One 7: e48140.
14. Hampton M, Melvin RG, Kendall AH, Kirkpatrick BR, Peterson N, et al. (2011) Deep Sequencing the Transcriptome Reveals Seasonal Adaptive Mechanisms in a Hibernating Mammal. PLoS ONE 6: e27021.
15. Hou R, Bao Z, Wang S, Su H, Li Y, et al. (2011) Transcriptome Sequencing and *De Novo* Analysis for Yesso Scallop (*Patinopecten yessoensis*) Using 454 GS FLX. PLoS ONE 6: e21560.
16. Jung H, Lyons RE, Dinh H, Hurwood DA, McWilliam S, et al. (2011) Transcriptomics of a Giant Freshwater Prawn (*Macrobrachium rosenbergii*): *De Novo* Assembly, Annotation and Marker Discovery. PLoS ONE 6: e27938.
17. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652.
18. O'Neil ST, Dzurisin JD, Carmichael RD, Lobo NF, Emrich SJ, et al. (2010) Population-level transcriptome sequencing of nonmodel organisms Erynnis propertius and Papilio zelicaon. BMC Genomics 11: 310.
19. Van Belleghem SM, Roelofs D, Van Houdt J, Hendrickx F (2012) De novo transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle Pogonus chalceus (Coleoptera, Carabidae). PLOS ONE 7: e42605.
20. Ji P, Liu G, Xu J, Wang X, Li J, et al. (2012) Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics. PLoS One 7: e35152.
21. He S, Mayden RL, Wang X, Wang W, Tang KL, et al. (2008) Molecular phylogenetics of the family Cyprinidae (Actinopterygii: Cypriniformes) as evidenced by sequence variation in the first intron of S7 ribosomal protein-coding gene: further evidence from a nuclear gene of the systematic chaos in the family. Mol Phylogenet Evol 46: 818–829.
22. Corona E, Dudley JT, Butte AJ (2010) Extreme evolutionary disparities seen in positive selection across seven complex diseases. PLOS ONE 5: e12236.
23. Fisher M, Gokhman I, Pick U, Zamir A (1996) A salt-resistant plasma membrane carbonic anhydrase is induced by salt in Dunaliella salina. J Biol Chem 271: 17718–17723.
24. Towle DW, Henry RP, Terwilliger NB (2011) Microarray-detected changes in gene expression in gills of green crabs (Carcinus maenas) upon dilution of environmental salinity. Comp Biochem Physiol Part D Genomics Proteomics 6: 115–125.
25. Tan YY, Zhou HY, Wang ZQ, Chen SD (2008) Endoplasmic reticulum stress contributes to the cell death induced by UCH-L1 inhibitor. Mol Cell Biochem 318: 109–115.
26. Li J, Zhang Z, Vang S, Yu J, Wong GK, et al. (2009) Correlation between Ka/Ks and Ks is related to substitution model and evolutionary lineage. J Mol Evol 68: 414–423.
27. Qiu Q, Zhang G, Ma T, Qian W, Wang J, et al. (2012) The yak genome and adaptation to life at high altitude. Nat Genet 44: 946–949.
28. Cirulli F, Alleva E (2009) The NGF saga: from animal models of psychosocial stress to stress-related psychopathology. Front Neuroendocrinol 30: 379–395.
29. Zhang Y, Xu W, Li Z, Deng XW, Wu W, et al. (2008) F-box protein DOR functions as a novel inhibitory factor for abscisic acid-induced stomatal closure under drought stress in Arabidopsis. Plant Physiol 148: 2121–2133.
30. Wu Z, Lauer TW, Sick A, Hackett SF, Campochiaro PA (2007) Oxidative stress modulates complement factor H expression in retinal pigmented epithelial cells by acetylation of FOXO3. J Biol Chem 282: 22414–22425.
31. Tsuji G, Takahara M, Uchi H, Takeuchi S, Mitoma C, et al. (2011) An environmental contaminant, benzo(a)pyrene, induces oxidative stress-mediated interleukin-8 production in human keratinocytes via the aryl hydrocarbon receptor signaling pathway. J Dermatol Sci 62: 42–49.
32. Ota T, Gayet C, Ginsberg HN (2008) Inhibition of apolipoprotein B100 secretion by lipid-induced hepatic endoplasmic reticulum stress in rodents. J Clin Invest 118: 316–332.
33. Henry RP, Gehnrich S, Weihrauch D, Towle DW (2003) Salinity-mediated carbonic anhydrase induction in the gills of the euryhaline green crab, Carcinus maenas. Comp Biochem Physiol A Mol Integr Physiol 136: 243–258.
34. Fisher M, Gokhman I, Pick U, Zamir A (1996) A salt-resistant plasma membrane carbonic anhydrase is induced by salt in Dunaliella salina. Journal of Biological Chemistry 271: 17718–17723.
35. Santovito G, Marino S, Sattin G, Cappellini R, Bubacco L, et al. (2012) Cloning and characterization of cytoplasmic carbonic anhydrase from gills of four Antarctic fish: insights into the evolution of fish carbonic anhydrase and cold adaptation. Polar Biology 35: 1587–1600.
36. Jain M, Ghanashyam C, Bhattacharjee A (2010) Comprehensive expression analysis suggests overlapping and specific roles of rice glutathione S-transferase genes during development and stress responses. BMC Genomics 11: 73.
37. Chen JH, Jiang HW, Hsieh EJ, Chen HY, Chien CT, et al. (2012) Drought and salt stress tolerance of an Arabidopsis glutathione S-transferase U17 knockout mutant are attributed to the combined effect of glutathione and abscisic acid. Plant Physiol 158: 340–351.
38. Qi Y, Liu W, Qiu L, Zhang S, Ma L, et al. (2010) Overexpression of glutathione S-transferase gene increases salt tolerance of Arabidopsis. Russian Journal of Plant Physiology 57: 233–240.
39. Wan Q, Whang I, Lee JS, Lee J (2009) Novel omega glutathione S-transferases in disk abalone: Characterization and protective roles against environmental stress. Comp Biochem Physiol C Toxicol Pharmacol 150: 558–568.
40. Zhou J, Wang WN, Wang AL, He WY, Zhou QT, et al. (2009) Glutathione S-transferase in the white shrimp Litopenaeus vannamei: Characterization and regulation under pH stress. Comp Biochem Physiol C Toxicol Pharmacol 150: 224–230.
41. Gomes SI, Novais SC, Gravato C, Guilhermino L, Scott-Fordsmand JJ, et al. (2012) Effect of Cu-nanoparticles versus one Cu-salt: analysis of stress biomarkers response in Enchytraeus albidus (Oligochaeta). Nanotoxicology 6: 134–143.
42. Ji W, Zhu Y, Li Y, Yang L, Zhao X, et al. (2010) Over-expression of a glutathione S-transferase gene, GsGST, from wild soybean (Glycine soja) enhances drought and salt tolerance in transgenic tobacco. Biotechnol Lett 32: 1173–1179.
43. Tim-Tim ALS, Morgado F, Moreira S, Rangel R, Nogueira AJA, et al. (2009) Cholinesterase and glutathione S-transferase activities of three mollusc species

from the NW Portuguese coast in relation to the 'Prestige' oil spill. Chemosphere 77: 1465–1475.
44. Gupta AS, Heinen JL, Holaday AS, Burke JJ, Allen RD (1993) Increased resistance to oxidative stress in transgenic plants that overexpress chloroplastic Cu/Zn superoxide dismutase. Proceedings of the National Academy of Sciences 90: 1629–1633.
45. Celino FT, Yamaguchi S, Miura C, Ohta T, Tozawa Y, et al. (2011) Tolerance of Spermatogonia to Oxidative Stress Is Due to High Levels of Zn and Cu/Zn Superoxide Dismutase. PLOS ONE 6: e16938.
46. Hepburn JJ, Arthington JD, Hansen SL, Spears JW, Knutson MD (2009) Technical note: copper chaperone for copper, zinc superoxide dismutase: a potential biomarker for copper status in cattle. J Anim Sci 87: 4161–4166.
47. Valavanidis A, Vlahogianni T, Dassenakis M, Scoullos M (2006) Molecular biomarkers of oxidative stress in aquatic organisms in relation to toxic environmental pollutants. Ecotoxicol Environ Saf 64: 178–189.
48. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11: 485.
49. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research 35: W182–W185.
50. Min XJ, Butler G, Storms R, Tsang A (2005) TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences. Nucleic Acids Res 33: W669–672.
51. MI T, D F (2005) Msatfinder: detection and characterisation of microsatellites. Available: http://www.genomics.ceh.ac.uk/msatfinder/. CEH Oxford, Mansfield Road, Oxford OX1 3SR.
52. Zhang Z, Li J, Yu J (2006) Computing Ka and Ks with a consideration of unequal transitional substitutions. BMC Evol Biol 6: 44.
53. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, et al. (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4: 259–263.