Genomic Basis of Adaptive Evolution: The Survival of Amur Ide (*Leuciscus waleckii*) in an Extremely Alkaline Environment

Jian Xu,[†,1] Jiong-Tang Li,[†,1] Yanliang Jiang,[1] Wenzhu Peng,[1,2] Zongli Yao,[3] Baohua Chen,[1] Likun Jiang,[1] Jingyan Feng[1], Peifeng Ji[1], Guiming Liu,[4] Zhanjiang Liu,[5] Ruyu Tai,[1] Chuanju Dong,[1] Xiaoqing Sun,[1] Zi-Xia Zhao,[1] Yan Zhang,[1] Jian Wang,[6] Shangqi Li,[1] Yunfeng Zhao,[1] Jiuhui Yang,[7] Xiaowen Sun,[1] and Peng Xu*[,1,2,8]

[1]Beijing Key Laboratory of Fishery Biotechnology, Centre for Applied Aquatic Genomics, Chinese Academy of Fishery Sciences, Beijing, China

[2]State Key Laboratory of Marine Environmental Science, College of Ocean & Earth Science, Xiamen University, Xiamen, China

[3]Engineering Research Centre for Saline-alkaline Fisheries, East China Sea Fisheries Research Institute, Chinese Academy of Fisheries Sciences, Shanghai, China

[4]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[5]The Fish Molecular Genetics and Biotechnology Laboratory, Aquatic Genomics Unit, School of Fisheries, Aquaculture and Aquatic Sciences, Auburn University, Auburn, AL

[6]Division of Animal and Nutritional Sciences, West Virginia University, Morgantown, WV

[7]Dalinor National Nature Reserve, Keshiketeng, Chifeng, China

[8]Fujian Collaborative Innovation Centre for Exploitation and Utilization of Marine Biological Resources, Xiamen University, Xiamen, China

[†]These authors contributed equally to this work.

***Corresponding author**: E-mail: xupeng77@xmu.edu.cn

Associate editor: Joel Dudley

Abstract

The Amur ide (*Leuciscus waleckii*) is a cyprinid fish that is widely distributed in Northeast Asia. The Lake Dali Nur population inhabits one of the most extreme aquatic environments on Earth, with an alkalinity up to 50 mmol/L (pH 9.6), thus providing an exceptional model with which to characterize the mechanisms of genomic evolution underlying adaptation to extreme environments. Here, we developed the reference genome assembly for *L. waleckii* from Lake Dali Nur. Intriguingly, we identified unusual expanded long terminal repeats (LTRs) with higher nucleotide substitution rates than in many other teleosts, suggesting their more recent insertion into the *L. waleckii* genome. We also identified expansions in genes encoding egg coat proteins and natriuretic peptide receptors, possibly underlying the adaptation to extreme environmental stress. We further sequenced the genomes of 10 additional individuals from freshwater and 18 from Lake Dali Nur populations, and we detected a total of 7.6 million SNPs from both populations. In a genome scan and comparison of these two populations, we identified a set of genomic regions under selective sweeps that harbor genes involved in ion homoeostasis, acid-base regulation, unfolded protein response, reactive oxygen species elimination, and urea excretion. Our findings provide comprehensive insight into the genomic mechanisms of teleost fish that underlie their adaptation to extreme alkaline environments.

Key words: **alkaline, adaptation, genome, acid–base regulation, urea excretion, *Leuciscus waleckii*.**

Article

Introduction

Amur ide (*Leuciscus waleckii*, Family Cyprinidae) ($2n = 50$) is broadly distributed throughout Northeast Asia. *L. waleckii* usually inhabits freshwater environments but can also survive in saline and alkaline lakes. As an extreme instance, *L. waleckii* can survive in the highly alkaline (up to pH 9.6) water of Lake Dali Nur (116°25′–116°45′E, 43°13′–43°23′N), a typical soda lake with an unusually high carbonate concentration. Lake Dali Nur is located in an endorheic basin on the eastern Inner Mongolia Plateau in North China. Lake Dali Nur is believed to have expanded to 1600 km^2 during the early Holocene (11,500–7,600 calibrated years BP) owing to a mass freshwater influx of glacier melt from The Greater Khingan Range. The lake level fluctuated dramatically during the middle Holocene and constantly shrank during the late Holocene (3,450 calibrated years BP to present). The evaporation exceeds the precipitation and inflows, which has caused the lake to shrink from 1,600 to less than 200 km^2. The alkalinity has, in turn, increased

steadily (Xiao et al. 2008). Currently, the pH value ranges from 8.25 to 9.6, with titratable alkalinity of over 50 mmol/L and salinity of approximately 6‰.

The *L. waleckii* population in Lake Dali Nur, which inhabited freshwater during the early Holocene, have been stressed and selected by the increasing alkalinity during the late Holocene. It gradually adapted the alkalinize environment and became the most dominant fish species in Lake Dali Nur (Geng and Zhang 1988; Chi et al. 2010).The unique geological and evolutionary history of the *L. waleckii* population in Lake Dali Nur provides an exceptional model with which to characterize the teleost mechanisms of adaptation to extreme environments under palaeoenvironmental changes.

L. waleckii is an important source of income for local Mongolians who live around Lake Dali Nur, and it is an important food source for birds migrating from Siberia to the south (Zhang et al. 2008). Despite the economic and ecological importance of *L. waleckii* in Lake Dali Nur, the mechanisms of its high tolerance to alkalinity are still largely unknown. Comparative transcriptome studies on the *L. waleckii* inhabiting Lake Dali Nur and its freshwater sister Lake Ganggeng Nur have unveiled significant expression differences in genes in comprehensive functional categories and pathways to cope with alkaline environmental stress (Wang et al. 2013; Xu, Ji et al. 2013; Chang et al. 2014). Selective pressure analysis of protein coding genes has also identified a number of genes under high positive selection, including carbonic anhydrase, superoxide dismutase, and glutathione S-transferase (Xu, Ji et al. 2013). These findings imply that unique genomic adaptations allow *L. waleckii* to cope with extreme alkaline environments. Therefore, sequencing the *L. waleckii* genome should facilitate the comprehensive discovery and understanding of the potential molecular mechanisms of high-alkaline adaptation in teleost species.

Here, we present the draft genome of the *L. waleckii* inhabiting the extremely alkaline waters of Lake Dali Nur. The genome consists of 752 megabases assembled using genome sequences from next-generation, massively parallel sequencing platforms. A total of 23,560 protein coding genes were annotated. We also re-sequenced individuals from populations inhabiting extremely alkaline water (ALK) and freshwater (FW) and compared their genetic diversity to explore the potential genetic basis for their adaptation to extreme alkalinity.

Results and Discussion

Shotgun Sequencing and *De Novo* Assembly

Genomic DNA was extracted from a female *L. waleckii* collected at Lake Dali Nur, Inner Mongolia (43°22′43″N, 116°39′24″E) and subjected to shotgun sequencing using the Illumina short paired-end sequencing platform. A total of 176.7 Gb of sequence data with 237-fold coverage was generated (supplementary table S1, Supplementary Material online). The final assembly was 752.3 Mb, with a contig N50 of 37.3 kb and a scaffold N50 of 447.7 kb (table 1). The 1,840 longest scaffolds (26.1% of all scaffolds) covered more than 90% of the assembly. We estimated the genome size to be 0.9 Gb based on K-mer analysis (supplementary fig. S1, Supplementary Material online). Thus, the scaffolds covered at least 84% of the genome.

To validate the genome assembly, we mapped the 213-fold coverage paired-end clean reads to the assembly and found that over 94% of the Illumina reads could be aligned with the assembly using BWA (Li and Durbin 2009) and SMALT (supplementary table S2, Supplementary Material online). The insert length distributions of the two mate-pair libraries were consistent with the expected sizes, demonstrating the accuracy of the genome assembly (supplementary fig. S2, Supplementary Material online). To assess the gene coverage of the assembly, we mapped the transcriptome sequences (Wang et al. 2013; Xu et al. 2014) to the assembly. The effort yielded ~87.8% coverage of these transcripts with a nucleotide sequence similarity threshold of 80% (supplementary table S3). These results indicated that the genome assembly of *L. waleckii* had high coverage and was of high quality.

The genetic linkage map of *L. waleckii* is still not available, which hinders to create chromosome framework for *L. waleckii* genome. In order to facilitate further genetic analysis such as genome-scale selective sweep, we created 24 pseudochromosomes by anchoring scaffolds to the medaka genome, which has same chromosome number with *L. waleckii* (supplementary fig. S3, Supplementary Material online).

Characterization of Repetitive Elements

The *L. waleckii* genome had a GC content of 37.4%, similar to other sequenced teleost genomes (supplementary fig. S4, Supplementary Material online). The genome contained 1.65 million repeat copies and occupied approximately 284.23 Mb (37.78%) of genome size, including 32.75% interspersed repeats (supplementary tables S4 and S5, Supplementary Material online) and 4.94% tandem repeats. Among the interspersed repeats, the most abundant transposable elements were DNA transposons (15.64% of the genome), with hATs constituting the most abundant transposable elements, representing 4.80% of the genome. Retrotransposons were the second most abundant repeat elements (5.66% of the genome), including three major families of long terminal repeats (LTRs), long interspersed elements (LINEs), and short interspersed elements (SINEs). The repetitive element contents and their proportions were similar to other teleost genomes except for the LTRs (Xu et al. 2014), which were significantly expanded to 3.02% of *L. waleckii* compared with teleost species other than zebrafish (supplementary table S6, Supplementary Material online). The average number of synonymous nucleotide substitutions per site for each fragmented repeat was estimated using the Jukes–Cantor model (Jukes and Cantor 1969). Interestingly, the LTRs in the *L. waleckii* genome had a significantly greater synonymous nucleotide substitution rate than did the LTRs in other teleost species, whereas the LINEs and SINEs displayed no significant shift in the substitution rates (fig. 1, supplementary figs. S5 and S6, Supplementary Material online). The insertion time of retroelements can be estimated based on their substitution rate. Therefore, the higher substitution rate implied that the majority of the LTRs in the *L. waleckii* genome have much younger insertion ages than do those in other surveyed teleost species.

Table 1. Summary of Genome Assembly.

Genome Assembly		N50 (size/number)	N90 (size/number)	Longest (kb)	Total Length
	Contigs	37.3 kb/5,701	9.6 kb/20,566	303.5	738 Mb
	Scaffolds	447.7 kb/477	95.1 kb/1,840	3277.6	752 Mb
	Chromosomes	24 pseudo-chromosomes (from 538 scaffolds)			510.2 Mb (74%)
Repetitive elements		**Number**	**Total length (Mbp)**	**Percentage of genome (%)**	
	Total	1,652,292	284	37.8%	
	DNA transposons	618,675	118	16.0%	
	Retroelements	139,335	43	5.7%	
Protein-coding genes	Total	23,560			
	Annotated	23,068			
	Un-annotated	492			

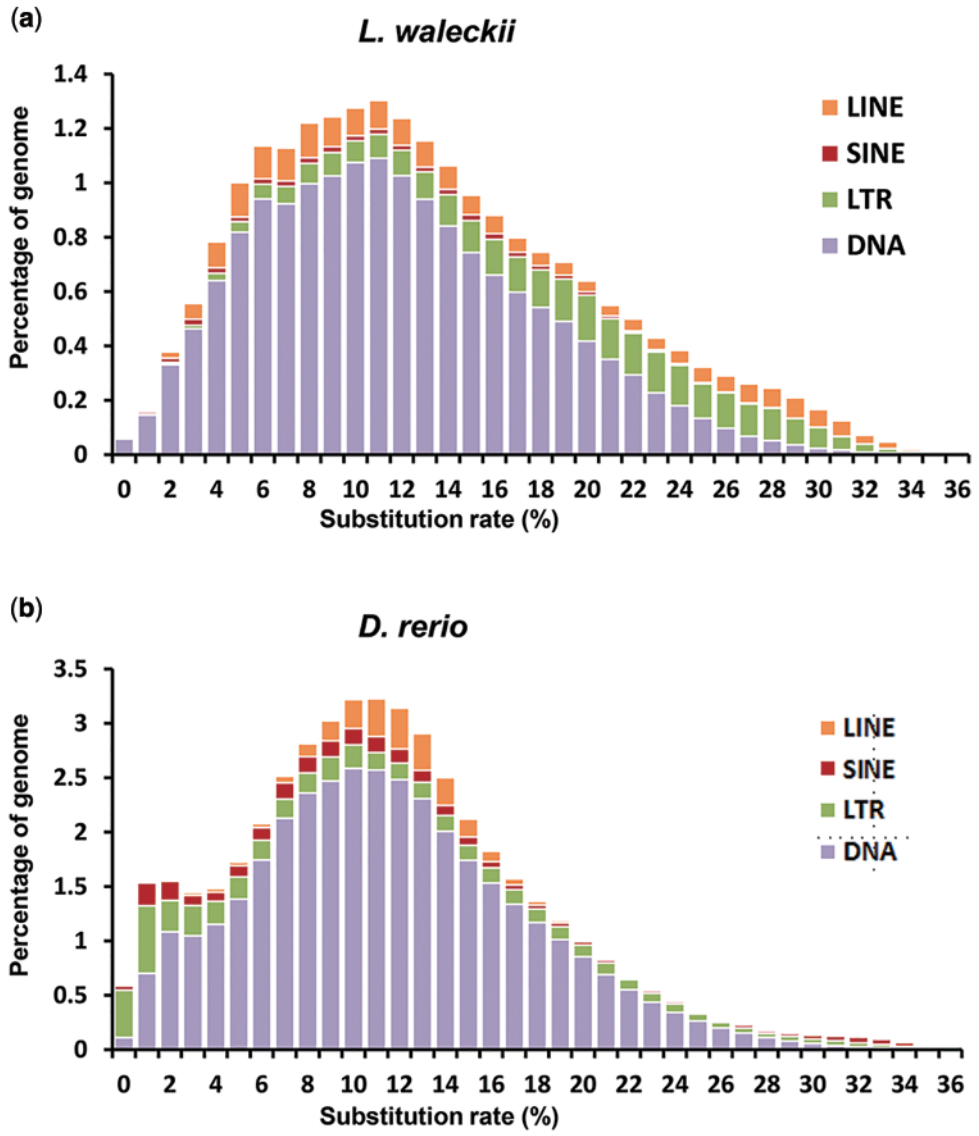

Fig. 1. Repetitive element contents in *L. waleckii*. Age distribution of repetitive element contents in (*a*) *L. waleckii* and (*b*) *D. rerio*. The average number of substitutions per site for each fragmented repeat was estimated using the Jukes-Cantor model. The substitution rates correlate with the ages of the repetitive elements. LINE, long interspersed elements; SINE, short interspersed elements; LTR, long terminal repeat.

Gene Annotation

We used a comprehensive strategy to annotate *L. waleckii* genes by combining FGENESH (Salamov and Solovyev 2000) *ab initio* gene prediction, protein-based homology and transcript-based evidence (supplementary table S7, Supplementary Material online). All predicted gene structures were integrated using EVidenceModeler (EVM) (Haas et al. 2008) to yield a consensus gene set containing a total of 23,560 protein-coding genes, of which 99.8% were proven to be expressed (table 1, supplementary fig. S7, supplementary tables S8 and S9, Supplementary Material online). The median gene and CDS lengths were 7,959 bp and 1,323 bp, respectively, with a median of 8 exons per gene (supplementary fig. S8 and table 10, Supplementary Material online). The gene structures of *L. waleckii*, including exon number, exon length and CDS length, were similar to the gene structures of other teleosts, indicating the high quality of the gene annotation (supplementary fig. S8, Supplementary Material online). Among the identified protein-coding genes, 17,099 genes were annotated with at least one Gene Ontology (GO) term (Qiu et al. 2012), and 4,671 genes were mapped to 331 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Moriya et al. 2007).

Genome Evolution

To determine the extent of genetic conservation among teleost fishes, we compared seven vertebrate genomes, including *L. waleckii*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Tetraodon nigroviridis*, *Danio rerio*, *Ctenopharyngodon idellus*, and *Homo sapiens*. Among them, *L. waleckii*, grass carp (*C. idellus*) and zebrafish (*D. rerio*) belong to the Cyprinidae family. We identified 11,006 orthologous gene families shared by at least two species, 1,372 of which were single-copy orthologues shared by all the studied species. Using these single-copy orthologues, we explored the phylogenetic relationships among the seven species (fig. 2*a* and supplementary table S11, Supplementary Material online). The phylogenetic tree revealed that *L. waleckii* was most closely related to grass carp, with an estimated divergence time of 158 My ago. We also studied the orthologue profiles of the five teleost fishes (including *L. waleckii*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Danio rerio*, and *Ctenopharyngodon idellus*) (fig. 2*b*). A total of 5,048 gene families were shared among the five fish species. Grass carp had 7,057 gene families overlapping with *L. waleckii*. Moreover, we uncovered 325 *L. waleckii* gene families with expansion and 4,420 families with contraction (fig. 2*a*).

Expanded Egg Coat Protein Genes for Protecting Embryos under Severe Stress

Almost all metazoan eggs are surrounded by extracellular egg coat (EC) glycoproteins, which are referred to as the zona pellucida domain-containing proteins (ZP proteins) or vitelline envelope proteins. These proteins play important roles in sperm recognition and EC polymerization during fertilization. In addition, they provide a critical protective barrier for the eggs of oviparous animals, which are directly exposed to

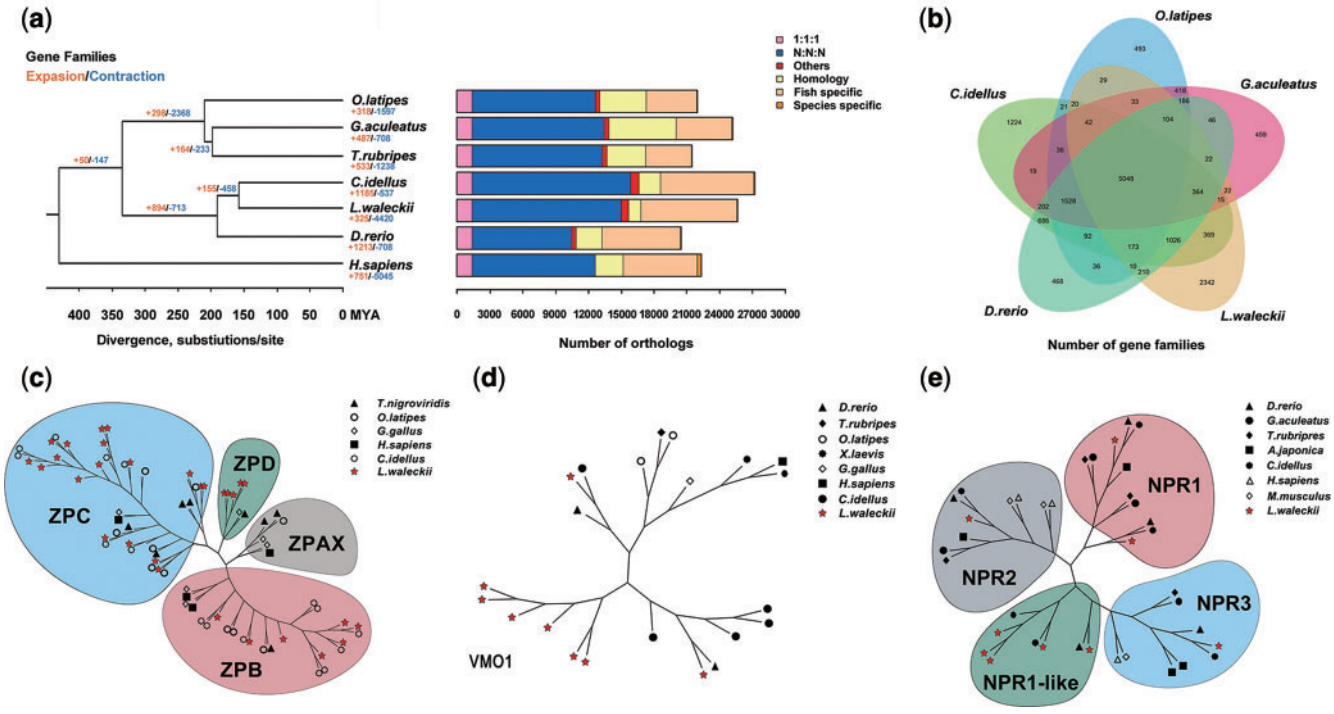

Fig. 2. Comparison of evolutionary features of *L. waleckii* and other vertebrates and expanded gene families. (*a*) Phylogenetic tree and numbers of gene families undergoing expansion (red)/contraction (green). My, million years ago. (*b*) Venn diagram showing unique and overlapping gene families in *L. waleckii*, *G. aculeatus*, *O. latipes*, *C. idellus*, and *D. rerio*. (*c*) Phylogenetic tree of ZP proteins in vertebrates showing gene expansion in the *L. waleckii* genome. (*d*) Phylogenetic tree of VMO1 proteins in vertebrates showing gene expansion in the *L. waleckii* genome. (*e*) Phylogenetic tree of NPR proteins in vertebrates showing gene expansion in the *L. waleckii* genome. The proteins of *L. waleckii* are marked with red stars in (c), (d) and (e).

external environments. In mammals, the ZP proteins are a family of 3–4 genes divided into three subtypes. However, in teleost fish, the ZP protein family typically contains a significantly greater number of ZP genes (Xu et al. 2012). For instance, we identified 10 ZP genes in *O. latipes* and 9 in *T. nigroviridis* based on the complete genome annotations. In the *L. waleckii* genome, 30 ZP genes were identified and annotated, representing a striking expansion compared with many other teleost fish. Phylogenetic analysis revealed significant expansions in several subfamilies, including ZPB, ZPC, and ZPD (fig. 2*c*). The expanded ZP genes in teleosts are considered to physically protect eggs against diverse aqueous environments. As an extreme example, remarkable ZP gene expansion has been confirmed in the Antarctic notothenioids. The enlarged ZP gene family undoubtedly contributes to the abundant ZP transcripts and the thick egg chorion, which is believed to act as a crucial physical barrier against ice penetration (Chen et al. 2008). In another case, the ZP gene family in viviparous platyfish is relatively contracted and changed under positive selection because those ZP proteins are no longer essential for protection but have evolved mechanisms to facilitate exchange of gas and substances (Schartl et al. 2013). In addition to ZP protein genes, we identified a considerable expansion of genes for vitelline membrane outer layer protein 1 (VMO1), a basic protein present in the outer layer of the vitelline membrane. A total of 8 VMO1 paralogs have been recognized in the *L. waleckii* genome, compared with one or two paralogs in vertebrates (fig. 2*d*, supplementary table S12, Supplementary Material online). The great expansions of both ZP and VMO1 genes in the *L. waleckii* genome may enhance the protective capacity for embryo development within the alkaline environment.

Expansion of Natriuretic Peptide Receptor Genes for Improving Sodium Transport

The natriuretic peptide (NP) system is a key endocrine system for osmoregulation and ion homoeostasis in vertebrates. It typically consists of three related hormones [atrial natriuretic peptide (ANP), ventricular natriuretic peptide (VNP), and C-type natriuretic peptide (CNP)] and three related receptors [natriuretic peptide receptors (NPRs) 1, 2, and 3]. ANP has been confirmed to be the primary sodium-regulating hormone in eels (Tsukada et al. 2005). We identified four copies of NP genes in the *L. waleckii* genome, which were similar to the ones in many other teleosts (supplementary fig. S9, Supplementary Material online). There are usually 4 or 5 copies of NPR genes in the model teleost fish. However, we identified nine copies of NPR genes in the *L. waleckii* genome, which were thus considerably expanded compared with other teleosts (fig. 2*e*). Of the nine NPR genes, NPR1 and NPR1-like genes, which encode NPRA, were expanded to four copies and three copies, respectively. The physiological role of the ANP/NPRA system in plasma sodium-proton (Na^+/H^+) exchange and excretion has been proven in various vertebrates (John et al. 1995; Shi et al. 2003; Tsukada et al. 2005). Na^+/H^+ exchange has been proposed to be associated with bicarbonate (HCO_3^-) transportation and acid-base regulation in mammals (Bobulescu and Moe 2006; Eladari and Kumai 2015). Furthermore, evidence in the Japanese eel has also demonstrated Na^+-dependent transepithelial HCO_3^- transport. Therefore, the expansion of NPR1 genes in the *L. waleckii* genome might be among the adaptive changes that enhance the capacity of sodium and HCO_3^- transport for acid–base regulation in response to alkaline stress.

Expansion of Solute Carrier Family 12 for Sustaining Ion and Acid-Base Balance

Solute carriers (SLCs) are the largest group of transporters, comprising transmembrane transporters for inorganic ions, amino acids, neurotransmitters, sugars, purines and fatty acids, among other solute substrates. There are approximately 400 genes organized into 52 SLC families in human and 338 genes organized into 50 families in the teleost fish, representing a major portion of the transporter-related genes in genomes (Fredriksson et al. 2008; Tiziano et al. 2012). In the *L. waleckii* genome, we identified an expanded SLC12 family, which contains 22 members in the *L. waleckii* genome compared with 9 in *H. sapiens*, 12 in *O. latipes*, 12 in *G. aculeatus*, 18 in *D. rerio* and 19 in *C. idellus* (supplementary fig. S10 and table S12, Supplementary Material online). SLC gene family 12 encodes a group of nine electroneutral cation-coupled chloride cotransporters, which includes three subgroups of sodium–potassium chloride cotransporters and sodium chloride cotransporters (NKCCs and NCCs) (including three members), potassium chloride cotransporters (KCCs) (including four members), and two orphans [cation–chloride cotransporter 9 (CCC9) and cotransporter interacting protein (CIP)], all of which are critical for cation-coupled chloride transport and chloride concentration modulation (Hebert et al. 2004; Arroyo et al. 2013). It is well documented that HCO_3^- and chloride can be exchanged across the membrane via a chloride shift mechanism, which is critical for carbon dioxide excretion and the maintenance of intracellular pH equilibrium (Crandall et al. 1981; Payne et al. 2003; Westen and Prange 2003). Therefore, the expansion of the SLC12 genes suggests that they may facilitate Cl^- uptake, thus play important roles on HCO_3^- transport and intracellular acid-base regulation for coping with severe alkaline stress in Lake Dali Nur.

Differential Gene Expression under Alkaline Stress

Previously, differential gene expression (DGE) analysis based on RNA-Seq data from ALK and FW populations using a *de novo* assembled transcriptome reference suggested that a large number of genes were differentially expressed in the two distinct water environments (Xu, Li et al. 2013). To provide experimental evidence linking the positively selected genes and expanded genes, we re-analyzed the DGE data based on the new draft genome and gene annotation. The results demonstrated that 2339, 1924, and 250 genes were differentially expressed in the liver, kidney, and gill, respectively (supplementary table S13, Supplementary Material online). Furthermore, the expression levels of 41 out of 50 selected candidate genes were validated by real-time

qRT-PCR analysis (supplementary fig. S11, Supplementary Material online). We found that a number of genes in the expanded gene families demonstrated significant differences in expression. For example, we found that at least two ZP protein genes (ZPB and ZPC) and one VMO1 gene were differentially expressed in the ALK samples. Two NPR genes (NPR1L, NPR3) showed significantly increased expression level in the kidney in ALK samples. We also found that four SLC12 genes (SLC12A2, SLC12A3, SLC12A5, and SLC12A8) were differentially expressed in the liver and three (SLC12A2, SLC12A3, and SLC12A8) in the kidney (supplementary tables S12 and S13, Supplementary Material online). The results provided functional evidence that these three expanded gene families are indeed linked to alkaline stress response and adaptation.

Genome Re-Sequencing and Genetic Diversity Analysis

L. waleckii is widely distributed in Northeast Asia and inhabits various aquatic environments. The Lake Dali Nur population shows an exceptional example of rapid adaptive evolution in response to its alkaline environment. We re-sequenced the genomes of 28 individuals from two distinct ecological environments, including 10 from River Ussuri and 18 from Lake Dali Nur (supplementary fig. S12 and table S14, Supplementary Material online), to investigate their genetic variation and the selective signatures of adaptive evolution. In contrast to the alkaline-tolerant population (ALK) from Lake Dali Nur, the River Ussuri population inhabits flowing freshwater with pH values ranging from 7.09 to 7.36 (FW) (Yong Liu et al. 2001). A total of 253.3 GB sequence data points were collected, generating 8.2 GB for each individual with 10.9-fold depth on average. The sequence data were mapped to the reference genome for SNP and small INDEL calling. After applying stringent quality control criteria, we identified a total of 6,477,849 candidate SNPs, of which 5,910,508 were intergenic, 285,442 were intronic, and 281,899 were exonic (supplementary fig. S13 and table S15, Supplementary Material online). Of the exonic SNPs, we identified 79,004 synonymous and 138,725 non-synonymous SNPs. Forty SNPs were randomly selected to assess SNP calling by Sanger sequencing, and 34 were validated as true SNPs, suggesting the high reliability of

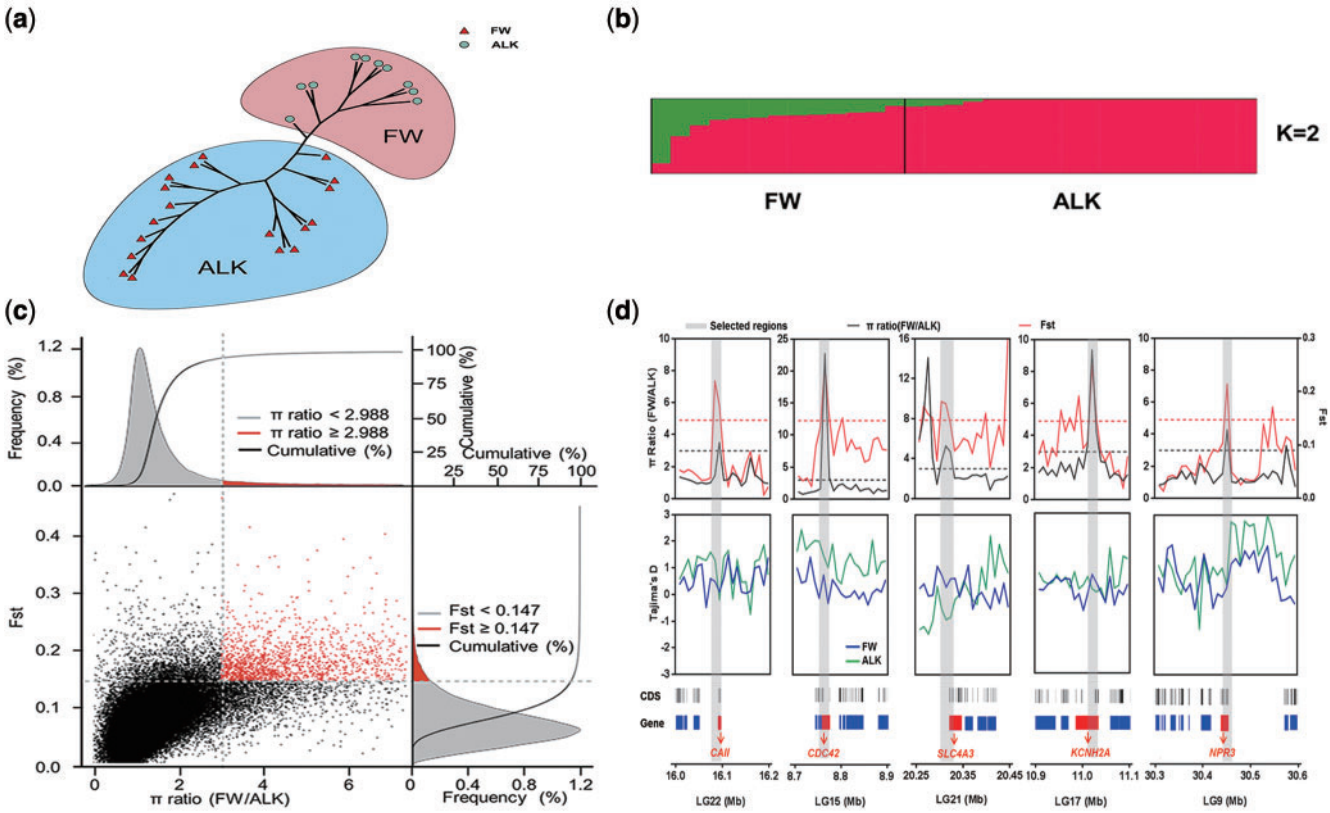

Fig. 3. Population genetics and genomic regions under selective sweeps. (*a*) Maximum-likelihood phylogenetic tree of FW ($n = 10$) and ALK ($n = 18$) populations. Blue dots represent FW samples, whereas red triangles represent ALK individuals. (*b*) Population structure. The length of each colored segment represents the proportion of the individual's genome from $K = 2$ ancestral populations. (*c*) Distribution of π ratios (FW/ALK) and *F*st values, calculated in 10 kb windows with 10 kb sliding steps. Data points located to the right of the vertical dashed line (corresponding to the 5% tail of the empirical π ratio distribution, where the π ratio is 2.988) and above the horizontal dashed line (the 5% right tail of the empirical F_{st} distribution, where F_{st} is 0.147) were identified as selected regions for the ALK population (red points). (*d*) Selective sweeps on five selected genes. The π ratios, F_{st} values and Tajima's *D* values were plotted using 10 kb sliding windows. Genomic regions located above the red dashed line (corresponding to the top 5% of F_{st} values, where F_{st} is 0.147) and above the black dashed line (corresponding to the top 5% of π ratios, where the π ratio is 2.988) were termed as regions under strong selective sweeps for the ALK population (grey regions). Genome annotations are shown at the bottom (black bar, coding sequences (CDS); blue bar, genes). The boundaries of CAII, CDC42, SLC4A3, KCNH2A, and NPR3 are marked in red.

the SNP calling pipelines (supplementary table S16, Supplementary Material online). In addition, we identified 1,147,384 small insertions/deletions (INDELs) (supplementary table S15, Supplementary Material online). A phylogenetic tree was constructed based on the sequence variations (fig. 3*a*), forming distinct clades in two groups that were consistent with the geographic divergence. There were significant differences in SNP diversity between the FW and ALK genomes. The SNPs identified from the FW population ranged from 182,409 to 406,202, with an average number of 300,529. The SNPs identified from the ALK population ranged from 93,730 to 242,580, with an average number of 159,543, showing significantly lower diversity than the FW population (*t*-test, $P = 7.36E{-}08$) (supplementary table S17, Supplementary Material online). Population structure was elucidated using the Bayesian clustering program STRUCTURE. Similar results were observed, with the FW population demonstrating more admixed and diversified genetic components than the ALK population (fig. 3*b*). Linkage disequilibrium (LD) decay analysis showed that the ALK population had longer LD blocks than the FW population (supplementary fig. S14, Supplementary Material online). The above observations support the idea that ALK populations have significantly lower genetic diversities than FW populations, which suggests that either genetic drift or a bottleneck effect may have occurred in the ALK population owing to climatic and geological changes over the past several thousand years, which reduced the genetic diversity both prominently and quickly. Intensive fishing may also have reduced the genetic diversity in recent decades.

Whole genome re-sequencing data also facilitated the gene copy number evaluation between FW and ALK population by comparing sequencing depth of each locus. The copy numbers of those genes in the expanded gene families in supplementary table S12, Supplementary Material online, were selected for depth ratio analysis. A total of 10 genes demonstrated read depth ratio (ALK/FW) greater than 1.4, suggesting their expansion in the ALK population. ALL ten candidate genes were selected for qPCR validation, and finally confirmed their copy number expansion in the ALK population (supplementary table S18, Supplementary Material online).

Genome-Wide Selection Pressure Analysis

We hypothesized that the ALK population would exhibit significant genetic differences in certain genes and genome regions compared with the FW population, underlying their adaptation and tolerance to the severe alkaline environment in Lake Dali Nur. Rapidly evolving genes in response to extreme environment challenges are usually under positive selection pressure and retain an elevated ratio of nonsynonymous to synonymous substitutions (K_a/K_s, or ω). The identification of genes that have undergone positive selection in the ALK population, in particular, can provide insight into the mechanism of adaptive evolution. To accurately detect the genomic footprints left by natural selection, we performed selection pressure analysis across the genome and compared the ALK and FW populations based on the genome-wide SNP loci and their genotypes. We calculated the K_a/K_s ratios of 10,549 genes in ALK and FW to detect positively selected genes (PSGs, K_a/K_s ratio ≥ 1) in both populations. We identified 2,843 PSGs in the ALK population (supplementary table S19, Supplementary Material online). We also identified 5,676 fast-evolving genes (FEGs) with higher ω values in ALK than in FW (supplementary table S20, Supplementary Material online). The high numbers of PSGs and FEGs in the genome of *L. waleckii* inhabiting Lake Dali Nur suggested that *L. waleckii* might have undergone adaptive evolution to cope with extremely inhospitable alkaline environments. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were performed on these PSGs and FEGs to identify enrichments for specific functional categories, revealing that significant portions of adaptive genes and enriched categories were under strong positive selection (supplementary tables S21 and S22 and fig. S16, Supplementary Material online). Interestingly, we found strong positive selection on many genes in the expanded gene families described above, suggesting that both effects have shaped these targeted genes on differential scales to ensure survivability under serious stresses (supplementary table S12, Supplementary Material online). Furthermore, we found that 416 PSGs and 917 FEGs were differentially expressed in ALK and FW compared with various tissues, thus providing more evidence that these genes are linked to alkaline acclimation.

Genome-Wide Selective Sweep Analysis

The genetic diversity in certain genome regions would be significantly decreased as a result of natural selection. To identify genome regions under selective sweep in the ALK genome, we scanned the genome-wide variations and allele frequency spectra based on the prospects of approximately 6.5 million SNPs. The π ratios ($\pi_{FW/ALK}$) were calculated using a 10 kb sliding-window approach with the VCFTOOLS software. In comparison to the FW population, the ALK population had a significantly lower level of diversity (median $\pi_{FW/ALK} = 1.28$) (supplementary fig. S17, Supplementary Material online), higher selective pressure (mean $\omega_{ALK}/\omega_{FW} = 1.30$), and higher linkage disequilibrium (LD) among SNP loci ($P < 0.0001$, Mann–Whitney *U* test) (supplementary fig. S15, Supplementary Material online), reflecting the higher selection pressure, fewer recombination events and skewed allele frequencies in the small and isolated ALK population in Lake Dali Nur. We identified a total of 3,760 significant windows corresponding to 37.6 Mb in size (top 5%, empirical π ratios $\geq$ **2.988**) with median $\pi_{FW/ALK} = 4.03$, which included 1,552 candidate genes (supplementary table S23, Supplementary Material online) based on the π ratio analysis. To further validate the genome regions under strong selective sweeps in the ALK population, the genome regions with Fst greater than 0.147 (top 5%) were also identified, corresponding to 37.6 Mb and 1,574 candidate genes (supplementary table S24, Supplementary Material online and fig. 3*c*). A total of 614 candidate genes shared by both the π ratio and Fst analysis were recognized as potentially affected genes under

selective sweeps (fig. 3*c*). The results based on Tajima's *D* analysis also provided supportive evidence identifying strong selective sweep signals (supplementary fig. S18, Supplementary Material online). The results suggested that the genomes of the ALK population in Lake Dali Nur have been significantly shaped by the extreme alkaline environment. To better understand these candidate genes and their potential functions, GO and KEGG analyses were performed on the candidate genes, offering clear insight into the genetic evolution and adaptive mechanisms of *L. waleckii* under extreme alkalinity (supplementary tables S25–S28, Supplementary Material online).

Genetic Mechanisms and Genes with Alkaline Acclimation

The consistent carbonate alkaline environment in the soda lake stimulates adaptive responses to maintain the intracellular acid–base balance and ion homoeostasis. Acid–base regulation in fish is linked to carbon dioxide (CO_2) and HCO_3^- excretion or uptake through the reversible hydration/dehydration reactions of CO_2 and the acid–base equivalents H^+ and HCO_3^- : $CO_2 + H_2O \Longleftrightarrow H^+ + HCO_3^-$. Carbonic anhydrases (CAs), the zinc metalloenzymes that catalyze these reversible reactions, are therefore of critical importance in CO_2 excretion and acid–base regulation (Gilmour and Perry 2009). We found that four CA genes (CAHZ, CAV, CAVIII, and CAXVI) are under higher selection pressures in the ALK population than in FW populations, suggesting that these CA genes are under positive selection in the alkaline environment. We also determined that two CA genes (CAII and CAXV) are located in selective sweep regions with significantly reduced genetic diversity (fig. 3*d* and supplementary tables S24 and S25, Supplementary Material online). The DGE results also showed that three CA genes (CAII, CAHZ, and CAXVI) have higher expression levels in the liver and kidney in ALK samples than in in FW samples (supplementary table S13, Supplementary Material online). This evidence implied that CA genes, as crucial acid–base regulation genes, evolved quickly to adapt to the extreme carbonate alkaline environment.

In fish, acid–base regulation is coupled to ionic regulation because acid–base compensation relies primarily on the direct transfer of H^+ and HCO_3^- in exchange for various anions and cations, especially Na^+ and Cl^- (Gilmour and Perry 2009). The regulation of ion transportation via the uptake, extrusion and sequestration of various ions is a critical mechanism of acid–base regulation. Because epithelial and endothelial membranes are impermeable to hydrophilic molecules, water and ion transportation uses alternative routes, including transcellular transportation (i.e., through the cell membrane and cytoplasm, mediated by various channels and vesicles) and paracellular transportation (i.e., through lateral intercellular spaces and compartments between cells). In addition to the expansions of certain gene families (NPR and SLC12), we also recognized substantial adaptive evidence regarding ion exchange and transportation, which are directly and indirectly linked to acid-base regulation and pH homoeostasis. We found a number of genes under positive selection and selective sweep in transcellular pathways. The genes embedded in the selective sweep regions were enriched in the categories "intracellular protein transport (GO: 0055085)", "phosphate ion transmembrane transporter activity (GO: 0015114)," and "voltage-gated calcium channel activity (GO: 0005245)" (supplementary tables S25–S28, Supplementary Material online). The genes involved in the ABC transporter (KEGG: K05416) and ion channel (KEGG: K04040) pathways had undergone selective sweep (supplementary fig. S18, Supplementary Material online). We identified a number of SLC genes that showed increased ω values in the ALK population and that were embedded in selective regions. These SLC genes encode various transmembrane transporters of anions and cations (e.g., SLC4A3, SLC4A4, SLC9A3, SLC12A1, SLC24A2, and SLC39A10). Similarly, we identified a considerable number of genes encoding energy-dependent ABC (ATP-binding cassette) transporters (e.g., ABCCs for transmembrane ion transportation) and ion channels (e.g., voltage-gated ion channel SCNs, CLCNs and KCNs; transient receptor potential cation channels TRPCs) in the same genome regions under selective sweeps (fig. 3*d* and supplementary table S29, Supplementary Material online). Many of these transporters and ion channels are critically important for acid-base regulation. For example, SLC4 family members encode Na^+-HCO_3^- cotranspoters (NBCs) that transport HCO_3^- (or related species, such as CO_3^{2-}) across the plasma membrane, which play important roles in acid-base homeostasis and intracellular pH regulation (Romero et al. 2013). The members of subfamily SLC9A encode Na^+/H^+ exchangers (NHE), which play a critical role in the fine tuning of intracellular pH and acid–base homeostasis (Kiela et al. 2006). Similarly, it is well known that those channels (SCNs, CLCNs and KCNs) are linked to acid-base regulation via ion and acid–base equivalent exchanges (Marshall and Grosell 2006; Holzer 2009). Therefore, the genes encoding transcellular ion transporters and channel proteins are reshaped by natural selection from the alkaline environment. These genetic changes facilitate the adaptation and survival of *L. waleckii* in Lake Dali Nur. Furthermore, we identified substantial amounts of differentially expressed genes as transporters and ion channels related to acid-base regulation (supplementary table S13, Supplementary Material online), such as SLC4 (SLC4A1 and SLC4A2), SLC9A (SLC9A3, SLC9A6, and SLC9A7), and various ion channels (SCN4, CLCN2, CLCKB, and KCNs), providing expression evidence for their functional connection with alkaline adaptation and tolerance.

Paracellular transportation is another key mechanism in the maintenance of ion homeostasis and acid–base balance. We found that both PSGs and FEGs were significantly enriched in various categories of paracellular permeability, including "BP: cell adhesion (GO: 0007155)", "CC: cell surface (GO: 0009986)," "CC: extracellular space (GO: 0005615)," "BP: extracellular matrix (ECM) organization (GO: 0030198)," "BP: cytoskeleton organization (GO: 0007010)," cell adhesion molecules (CAMs) (KEGG: 04514), ECM-receptor interaction (KEGG: 04512), and focal adhesion (KEGG: 04510) (supplementary figs. S15 and S19 and tables S21 and S22,

Supplementary Material online). These findings suggest that genes related to cell adhesion regulation and cytoskeletal organization are under positive selection. Furthermore, the genome-wide selective sweep analysis revealed that many genes embedded in selected genome regions belong predominantly to functional categories related to paracellular permeability, including "cell–cell adhesion (GO: 0016337)," "cadherin binding (GO: 0045296)," "cell junction assembly (GO: 0034329)," "cell–matrix adhesion (GO: 0007160)," "calcium ion binding (GO: 0005509)," "Ras GTPase binding (GO: 0017016)," and "actin binding (GO: 0003779)," in addition to cell adhesion molecule (CAM) pathways (KEGG: 04514), ECM–receptor interaction (KEGG: 04512), and the VEGF signaling pathway (KEGG: 04370) (supplementary fig. S18 and tables S25–S28, Supplementary Material online). We identified a number of important genes within the genome regions under selective sweeps, including integrins (6 genes: *ITGA4*, *ITGA8*, *ITGAM*, *ITGB2*, *ITGB3*, and *ITGB6*), tensins (2 genes: *TNS1* and *TNS3*), cadherins (13 genes: *CDH1*, *CDH2*, *CDH4*, *CDH7*, *CDH9*, *CDH11*, *CDH13*, *CDH16*, *CDH19*, *CDH23*, *CDH26*, *PCDH9*, and *PCDH15*), catenins (*CTNNA2*, *CTNNAL1*, *CTNND1*, and *CTNND2*), claudin (*CLDN7*), and occludin (*OCLN*), along with regulators such as vascular endothelial growth factor (*VEGF*), hypoxia-induced factors (*HIFs*), fibroblast growth factors (*FGFs*), platelet-derived growth factor receptor (*PDGFR*), and various small GTPases (*CDC42*, *RAS*, *RAB*, and *ARF*) (fig. 3d and supplementary table S30, Supplementary Material online). Integrins couple the ECM outside a cell to the cytoskeleton inside the cell, acting as a bridge for cell–cell and cell–ECM interactions (Marsden and DeSimone 2003). Tensins are critical proteins that stabilize integrin adhesion to the cytoskeleton (Torgler et al. 2004). Cadherin, catenin, claudin, and occludin are key components of tight junctions. VEGFs, HIFs and FGFs act as important modulators regulating paracellular permeability (Bazzoni and Dejana 2004; Komarova and Malik 2010; Maitre and Heisenberg 2013; Ganot, et al. 2015). Small GTPases play critical roles in regulating epithelial permeability and cytoskeletal rearrangement to maintain normal cell volume in response to external stress and stimuli (Citalan-Madrid et al. 2013). Thus, natural selection has substantially changed the genetic diversity of a vast number of genes related to paracellular transportation, which may facilitate efficient ion transportation and acid–base regulation for increasing the fitness of *L. waleckii* under alkaline stress in Lake Dali Nur.

Extreme environmental stresses can lead to the accumulation of misfolded or unfolded proteins in the endoplasmic reticulum (ER) and can trigger the unfolded protein response (UPR). The UPR then activates the protein folding machinery and triggers ubiquitin-mediated protein degradation, thus ensuring proper ER function (Liu and Howell 2010; Serohijos and Shakhnovich 2014). We observed significant gene enrichment of "proteolysis (GO: 0006508)" and "metallopeptidase activity (GO: 0008237)" within the selective sweep regions in the genome (supplementary tables S27 and S28, Supplementary Material online), which are critically important for removing misfolded and unfolded proteins and preventing irreversible cell damage and cell death (apoptosis). These genes encode various proteasome proteins (PSMs), peptidases, and proteases (supplementary table S31, Supplementary Material online). In addition to ER stress, extreme alkaline stress also generates other secondary stresses, particularly oxidative stress, which is caused by excessive reactive oxygen species (ROS). Scavenging ROS and protecting macromolecules such as DNA and proteins against oxidative damage are crucial under extreme environmental stress. The elimination of ROS relies primarily on antioxidant molecules such as glutathione (GSH) and scavenging enzymes such as superoxide dismutase (SOD), glutathione transferases (GSTs), glutathione peroxidase (GPX), and catalase (Xiong and Zhu 2002; Coleman et al. 2015). In our previous transcriptome study, we identified significant expression differences between the ALK and FW populations in the genes encoding SOD, GST, GPX, and catalase. We also found strong positive selection signals in the SOD and GST genes in the ALK population. As expected, the results based on whole-genome sequences were consistent with the previous findings and identified these genes as under positive selection. In addition, we also identified selective sweeps on genes encoding glutathione synthetase (GSS), S-formylglutathione hydrolase (ESD), and hydroxyacylglutathione hydrolase (HAGH), which play important roles in GSH accumulation and regulation. The evidence suggests that GSHs are key antioxidant compounds in *L. waleckii* for ameliorating ROS toxicity and adapting to severe alkaline stress as well as other environmental stresses. Differential gene expression analysis also revealed that many ROS scavenging genes, such as GSTs and SODs, were also differentially expressed in ALK compared with FW (supplementary table S13, Supplementary Material online), implying a connection between oxidative stress response-related genes and alkaline acclimation.

Ammonia is one of the main products of the nitrogen metabolism in fish, which is highly toxic and must be detoxified or excreted. Ammonotelism is the dominant strategy in teleost fish to remove nitrogenous waste by excreting ammonia directly into the aqueous environment via the gills. Ureotelism is an alternative strategy in teleost fish, in which excess nitrogen is excreted as urea (Wright and Anderson 2001). Urea rarely constitutes more than 20% of the total nitrogen excretion in teleost fish, except in certain species living in extreme habitats, such as the Lake Magadi tilapia (*Oreochromis alcalicus grahami*) (Randall et al. 1989; Wright and Land 1998; Walsh et al. 2001). *O. a. grahami* lives in water with titratable alkalinity around 300 mmol/L and pH 10, which hinders the passive diffusion of ammonia across the gills. Hence, ammonia is converted to urea, which is excreted across the gills instead (~80% of nitrogenous waste) (Wood et al. 1994). Similarly, *Chalcalburnus tarichi* in alkaline Lake Van (titratable alkalinity around 150 mmol/L and pH 9.8) at Eastern Anatolia in Turkey excretes approximate 37% of nitrogenous waste as urea (Danulat and Kempe 1992); whereas the Lahontan cutthroat trout (*Oncorhynchus clarki henshawi*) in the alkalized Pyramid Lake in Nevada (titratable alkalinity around 23 mmol/L and pH 9.4) excretes approximate 26% of nitrogenous waste as urea (Wilkie et al. 1993). The evidences suggested that ureotelism is commonly used alternative strategy in teleost fish for coping with elevated environmental

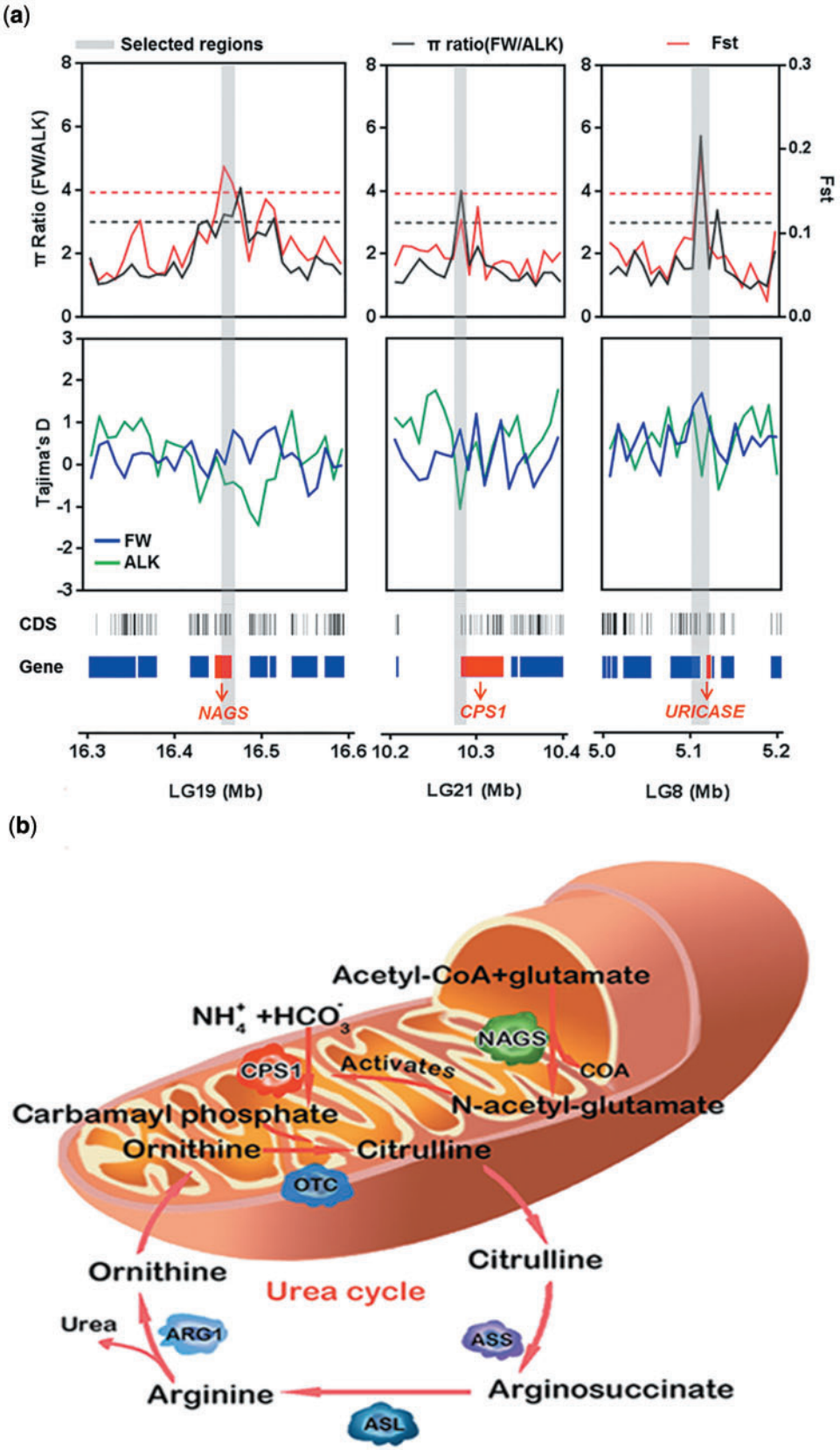

FIG. 4. Selective sweeps on urea excretion genes. (*a*) Selective sweeps on three selected genes. The π ratios, F_{st} values and Tajima's *D* values were plotted using 10 kb sliding windows. Genomic regions located above the red dashed line (corresponding to the top 5% of F_{st} values, where F_{st} is 0.147) and above the black dashed line (5% significance level of the π ratio, where the π ratio is 2.988) were termed as regions under strong selective sweeps for the ALK population (grey regions). Genome annotations are shown at the bottom [black bar, coding sequences (CDS); blue bar, genes]. The boundaries of *NAGS*, *CPS1* and *URICASE* are marked in red. (*b*) Schematic of the urea cycle. The components *CPS1* and *NAGS* under the selective sweeps are marked in red and green, respectively.

alkalinity. In most animals, the ornithine–urea cycle (OUC) is a crucial biochemical pathway that produces urea from ammonia, primarily in the liver. We have previously observed the elevated expression of two important genes in the OUC, argininosuccinate synthase (ASS) and arginase 1 (ARG1), in the liver of alkaline-acclimated *L. waleckii* compared with fish inhabiting freshwater (Xu, Li et al. 2013). We identified selective sweeps on the N-acetylglutamate synthase (NAGS) and carbamoyl-phosphate synthase I (CPS1) genes (fig. 4). NAGS encodes an enzyme that catalyzes the production of N-acetylglutamate (NAG). NAG is the essential cofactor of CPS1, the rate-limiting enzyme in ammonia fixation and urea generation in the OUC pathway. Therefore, both NAGS and CPS-1 act as upstream controllers of the OUC pathway and determine the urea excretion rate in vertebrates. We observed selective sweep on the gene encoding uricase (UOX) (fig. 4). We also observed elevated expression of UOX in the kidney of ALK compared with FW (supplementary table S13, Supplementary Material online). Uricase is the key enzyme in uricolysis, the biochemical pathway that catabolizes uric acid into urea for detoxification. Therefore, UOX is a key indicator for assessing the significance of uricolytic activity. The genes of both the OUC and uricolytic pathways were significantly selected for urea excretion in the *L. waleckii* genome. Previous reports found that *L. waleckii* had lower ammonia excretion rate in alkaline water in comparison with that in freshwater, whereas no significant difference on total ammonia concentration in plasma was observed (An et al. 2014). This evidence suggested that, similar to other teleost inhabiting various alkaline lakes, *L. waleckii* surviving in the Lake Dali Nur might also excrete nitrogenous waste with higher proportion of urea than those inhabiting freshwater environment.

Conclusion

L. waleckii exhibits a spectacular adaptation to extreme aquatic environments, providing a remarkable case for understanding evolutionary scenarios occurring under environmental changes. We developed the draft genome of *L. waleckii* inhabiting a soda environment, providing an important genomic resource for comprehensive comparative studies across teleost fish and facilitating the characterization of adaptive changes in gene families, transposable elements and genetic diversity. Re-sequencing multiple individuals from phenotypically divergent populations of the same species and genome scans further revealed important selective sweeps on a large set of genes. These data underscore broad biochemical and physiological mechanisms, including acid–base regulation and ion homoeostasis; the unfolded protein response and reactive oxygen species elimination; and ammonia fixation and urea excretion. Our results suggested that the extensive genomic changes occurred during the late Holocene under environmental stress accompanying climate changes, which drove *L. waleckii*'s adaptation to the extreme alkaline environment in the soda lake.

Materials and Methods

Genome Sequencing, Assembly, and Annotation

Genomic DNA was extracted from the blood cells of a female *L. waleckii* collected at Lake Dali Nur, Inner Mongolia (43°22′43″N, 116°39′24″E) using the DNeasy Blood and Tissue kit (Qiagen). We constructed one shotgun library and two mate-pair libraries according to Illumina standard operating procedures (supplementary table S1, Supplementary Material online). Each library was subjected to 2 × 100 bp read-length PE runs on an Illumina HiSeq 2000 instrument. We filtered out low-quality and short reads to obtain a set of usable reads and then assembled the genomes using ALLPATH-LG (Gnerre et al. 2011). To increase scaffold length, reads from mate-pairs of different insert sizes were added step-by-step for scaffolding using SSPACE (Boetzer et al. 2011), SOAP*denovo* (Luo et al. 2012), Opera (Gao et al. 2011), and SOPRA (Dayarian et al. 2010). Finally, we used the paired-end information from the short paired-end reads to fill the gaps between the scaffolds with Gapcloser in SOAP*denovo* and Gapfiller (Boetzer and Pirovano 2012).

We used three approaches for gene prediction: *ab initio* gene prediction, sequence homology-based prediction and RNA-Seq models. Briefly, FGENESH (Salamov and Solovyev 2000) was used to predict genes with fish gene model parameters in the repeat-masked genome sequences. Sequence homology-based gene predictions included both raw and precise alignments. First, the protein sequences of five model fish (*G. aculeatus*, *O. latipes*, *T. nigroviridis*, *D. rerio*, and *Takifugu rubripes*) downloaded from the Ensembl database (Cunningham et al. 2015) were aligned to their corresponding protein sequences in the database using BLAT (Kent 2002). If one alignment region covered at least 70% of the query proteins, the protein sequences were then aligned to those genome fragments using Genewise (Birney et al. 2004) to identify the splicing sites accurately. Transcriptome reads were generated using the Illumina platform (previously released data; available from the Sequence Read Archive (SRA) under accession numbers SRP018421 and SRP028616). Reads were mapped to genomic sequences with TopHat (Trapnell et al. 2009), and Cufflinks (Trapnell et al. 2012) was used to produce transcript assemblies. To represent a gene locus with several alternatively spliced transcripts generated by Cufflinks, the transcript with the longest exon length was chosen. All evidence was merged to form a comprehensive consensus gene set using EVM (Haas et al. 2008). To obtain homologous genes, BLAST searches were conducted against the NCBI nr, SwissProt and TrEMBL protein databases (UniProt_Consortium 2015), and homologues were identified with *E* values of $<1 \times 10^{-5}$. The functional classification of GO categories was performed using the InterProScan program (Quevillon et al. 2005) and homologue assignment. Pathway analysis was performed using the KEGG Automatic Annotation Server (KAAS) (Moriya et al. 2007).

The genome sequence reads and assembly have been deposited in the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena; last accessed February 8, 2016) under project PRJEB12292. The genome data can also be accessed

at http://www.fishbrowser.org/database/amur-ide/; last accessed February 8, 2016.

Repetitive Element Analysis and Construction of Pseudo-Chromosome

Both homology-based and *de novo* predictions were used to identify repeat contents in the Amur ide genome. For homology-based analysis, we used Repbase (version 20140131) to perform TE searches with RepeatMasker (version 4.0.5) using the WuBlast search engine. For *de novo* predictions, transposable repeats were identified in the Amur ide genome using RepeatMasker with a *de novo* repeat library constructed by RepeatModeler (version 1.0.8). RepeatModeler is a *de novo* repeat family identification and modeling package containing two *de novo* repeat-finding programs, RECON (Bao and Eddy 2002) and RepeatScout (Price et al. 2005). The repetitive elements were then classified using homologous searches with Repbase and a Support Vector Machine (SVM) method (TEClass). There is no linkage map available for *L. waleckii* so far. As the *O. latipes* genome maintains the teleost ancestral genome karyotypes, we selected the *O. latipes* genome (Kasahara et al. 2007) as a reference to construct *L. waleckii* pseudo-chromosomes to facilitate further population genetic analysis. We performed MCScanX (Wang et al. 2012) to identify syntenic blocks between *L. waleckii* and *O. latipes*, with the gap size set to 25 genes and at least 5 syntenic genes. Each *L. waleckii* scaffold was assigned to the best syntenic *O. latipes* genome region with the highest score. We then ordered and oriented the scaffolds following the locations of their syntenic regions on the *O. latipes* genome.

Gene Family Analysis

Orthologous gene sets were used for genome comparisons. Besides four diploid model fish (*G. aculeatus*, *O. latipes*, *T. rubripes*, and *D. rerio)* and *H. sapiens*, *C. idellus*, another diploid member of Cyprinidae family, was also used in the comparison. The protein sequences of *C. idellus* were downloaded from National Center for Gene Research website (http://www.ncgr.ac.cn/grasscarp/; last accessed September 1, 2016). We used TreeFam (Li et al. 2006) to define orthologous gene families among *L. waleckii* and the other six vertebrates. (1) The longest protein sequence in the seven sequenced genomes was chosen for each gene. (2) The proteins were blasted against themselves. Solar was used to conjoin the fragmental alignments for each gene-pair. (3) Hcluster_sg (hierarchical clustering) was used to define the clusters. The minimum edge weight (H score) and minimum edge density were set to 5 and 1/3, respectively. Expansion or contraction was defined by comparing the cluster size of the ancestor to the cluster size of each of the current species using the CAFÉ program (Hahn et al. 2007). Single-copy gene families were used to construct a phylogenetic tree for *L. waleckii* and other genomes. The protein alignment of each family with MUSCLE (Edgar 2004) was converted to codon alignment with PAL2NAL (Suyama et al. 2006). Gblocks (Talavera and Castresana 2007) was used to eliminate poorly aligned positions and divergent regions of the codon alignment. The refined codon alignments were concatenated to form one super-alignment. PhyML (Guindon et al. 2005) was used to reconstruct the phylogenetic tree using the best nucleotide substitution model (GTR + gamma + I) from the single-copy gene families that were present in all seven species. Human was selected as the outgroup of the phylogenetic tree. The divergence time for *L. waleckii* and other teleosts was estimated using the external calibration time of human and zebrafish (429 Myr ago) obtained from the TimeTree database (http://www.timetree.org/; last accessed February 8, 2016).

Genome Re-Sequencing, SNP Calling, Phylogenetics, and Genetic Structure

The two populations (FW and ALK) of *L. waleckii*, consisting of 28 individuals, were collected from Fuyuan County in Heilongjiang Province (10 samples) and Lake Dali Nur in Inner Mongolia (18 samples), respectively. Genome re-sequencing was performed using the Illumina HiSeq 2000 platform. Paired-end reads from each individual were aligned to the reference genome using the Burrows–Wheeler Aligner (BWA) (Li and Durbin 2009). After mapping, SNPs were identified based on the mpileup files generated by SAMtools (Li et al. 2011). The filtering threshold was set to require a read depth of $\geq$10 and a quality score of $\geq$20. Genotypes supported by at least two reads and with a minor allele frequency of $\geq$0.1 were assigned to each genomic position. A maximum-likelihood tree was constructed with RAxML (Stamatakis 2014) and displayed with TreeView (Page 1996), and all SNPs were used to investigate the population structure using STRUCTURE (Hubisz et al. 2009) with 2,000 iterations. The resulting structure matrix was plotted using the DISTRUCT software (Rosenberg 2004). Sequencing depth of each locus for was generated by SAMtools (Li et al. 2011) with the "–depth" parameter, and average depth of each gene was divided by depth of the sample, generating normalized depth ratios to evaluate the copy numbers. Real-time quantitative PCR was performed on selected genes for copy number validation following standard protocols of TransStart Top Green qPCR SuperMix (Transgene, Beijing, China).

K_a/K_s Ratio Calculation, PSGs, FEGs, and LD Analysis

The coding sequences of all annotated genes were used for the calculation of K_a/K_s (ω) ratios. Using SNP information from both the FW and ALK populations, "AXT" format files were prepared with individual scripts as input files for the KaKs_Calculator software (Wang et al. 2010). The K_a/K_s results were integrated and used to identify PSGs and FEGs. LD decays for the FW population, ALK population and all samples were calculated within a range of 50 kb using PLINK (Purcell et al. 2007). The SNP genotyping results were converted to PED/MAP format, and the average r^2 value of each 1 kb region was calculated. All r^2 values were then plotted against the physical distances of SNPs in units of kb.

Calculation of π Ratio, Fst, Tajima's D, and Identification of Selective Signatures

We calculated the π distribution for each linkage group using a sliding window method in VCFTOOLS (Danecek et al. 2011). The window width was set to 10 kb, and the stepwise

distance was 10 kb. The π values from the FW and ALK populations were compared, and the ratios were sorted. Fst and Tajima's *D* values were also calculated using VCFTOOLS with the parameters "–weir-fst-pop" and "–TajimaD", respectively. We identified the regions with the 5% highest π ratios and the regions with the 5% highest Fst values. Together with regions identified on the basis of the above two thresholds, genes within selective sweeps were annotated using GOEAST for Gene Ontology and the DAVID software for KEGG pathway analysis.

Differential Gene Expression Analysis

Differential gene expression analysis had been previously conducted based on RNA-Seq from the ALK and FW populations using the *de novo* assembled transcriptome reference as reference sequences (Xu, Li et al. 2013). Here, we mapped those transcriptome contigs to the new assembled *L. waleckii* genome and re-annotated the differentially expressed genes with the new gene identifiers. Real-time quantitative PCR was performed on selected genes to validate the DGE results, as in the previous study (Xu, Li et al. 2013).

URLs

Kyoto Encyclopedia of Genes and Genomes (KEGG), http://www.genome.jp/kegg/ (last accessed February 8, 2016); KEGG Automatic Annotation Server (KAAS), http://www.genome.jp/tools/kaas/; (last accessed February 8, 2016) DAVID, http://david.abcc.ncifcrf.gov/summary.jsp (last accessed February 8, 2016); SMALT, http://www.sanger.ac.uk/resources/software/smalt/ (last accessed February 8, 2016); SOAPdenovo, http://soap.genomics.org.cn/ (last accessed February 8, 2016); RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html (last accessed February 8, 2016); RepeatMasker, http://www.repeatmasker.org/ (last accessed February 8, 2016); Repbase, http://www.girinst.org/repbase/ (last accessed February 8, 2016); TimeTree, http://www.timetree.org/ (last accessed February 8, 2016); TreeFam, http://treefam.genomics.org.cn/ (last accessed February 8, 2016); Ensembl, http://ftp://ftp.ensembl.org/pub/ (last accessed February 8, 2016); GOEAST, http://omicslab.genetics.ac.cn/GOEAST/ (last accessed February 8, 2016); Amur ide genome database, http://www.fishbrowser.org/database/amur-ide/ (last accessed February 8, 2016).

Supplementary Material

Supplementary figures S1–S19 and tables S1–S30 are available at *Molecular Biology and Evolution* online.

Author Contributions

P.X. conceived the study and wrote the manuscript. P.X., J.L., and J.X. developed the sequencing strategy. J.X. performed the library construction and next-generation sequencing. J.L. conducted the genome assembly, assessment, annotation, and comparative genomic analysis. J.X. performed the genome re-sequencing study and genetic diversity analysis. Y.J. performed repeat annotation and analysis. P.X., J.X., Z.Z., Y.Z., P.J., J.W., and J.Y. collected fish samples. W.P., B.C., C.D., G.L., L.J., J.F., X.Q.S., S.L., and Y.Z. performed gene family analysis and gene validation. P.X., J.L., J.X., W.P., P.J., and Y.J. prepared the figures. Z.L. and X.W.S. participated in discussions and provided advice. Z.Y. participates in the manuscript revision.

Acknowledgments

We thank Mr Liezhi Dong, Mr Zuofa Jiang and Mr Changhai Zhou for their assistance with sample collection. We thank Dr Pung-pung Hwang and Dr Liqun Liang for their helpful discussions. This work was supported by the National Natural Science Foundation of China (Nos. 31422057, 31101893 and 31402353); Special Scientific Research Funds for Central Nonprofit Institutes, Chinese Academy of Fishery Sciences (2013A03YQ01 and 2015C005); National High-Technology Research and Development Program of China (2011AA100401); Fundamental Research Funds for the Central Universities, Xiamen University (20720160110) and National Infrastructure of Fishery Germplasm Resources of China (2016DKA30470). The authors declare no competing financial interests.

References

An X, Qi J, Luo X, Wu L, Wang N. 2014. Ammonia excretion rates and contents in blood and other tissues in *Leuciscus waleckii* inhabiting alkaline and freshwater lake (in Chinese). *Mod Agricul Sci Technol.* 244–247.

Arroyo JP, Kahle KT, Gamba G. 2013. The SLC12 family of electroneutral cation-coupled chloride cotransporters. *Mol Aspects Med.* 34:288–298.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12:1269–1276.

Bazzoni G, Dejana E. 2004. Endothelial cell-to-cell junctions: molecular organization and role in vascular homeostasis. *Physiol Rev.* 84:869–901.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res.* 14:988–995.

Bobulescu IA, Moe OW. 2006. Na+/H+ exchangers in renal regulation of acid-base balance. *Semin Nephrol.* 26:334–344.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27:578–579.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol.* 13:R56.

Chang YM, Tang R, Dou XJ, Tao R, Sun XW, Liang LQ. 2014. Transcriptome and expression profiling analysis of *Leuciscus waleckii*: an exploration of the alkali-adapted mechanisms of a freshwater teleost. *Mol Biosyst.* 10:491–504.

Chen Z, Cheng CH, Zhang J, Cao L, Chen L, Zhou L, Jin Y, Ye H, Deng C, Dai Z, et al. 2008. Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. *Proc Natl Acad Sci U S A.* 105:12944–12949.

Chi B, Chang Y, Yan X, Cao D, Li Y, Gao Y, Liu Y, Liang L. 2010. Genetic variability and genetic structure of *Leuciscus waleckii* Dybowski in Wusuli River and Dali Lake. *J Chin Fish.* 17:228–235.

Citalan-Madrid AF, Garcia-Ponce A, Vargas-Robles H, Betanzos A, Schnoor M. 2013. Small GTPases of the Ras superfamily regulate intestinal epithelial homeostasis and barrier function via common and unique mechanisms. *Tissue Barriers* 1:e26938.

Coleman BD, Marivin A, Parag-Sharma K, DiGiacomo V, Kim S, Pepper JS, Casler J, Nguyen LT, Koelle MR, Garcia-Marcos M. 2015. Evolutionary conservation of a GPCR-independent mechanism of trimeric G protein activation. *Mol Biol Evol.* 33:820–837.

Crandall ED, Mathew SJ, Fleischer RS, Winter HI, Bidani A. 1981. Effects of inhibition of RBC HCO3-/Cl- exchange on CO2 excretion and

downstream pH disequilibrium in isolated rat lungs. *J Clin Invest.* 68:853–862.
Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res.* 43:D662–D669.
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.
Danulat E, Kempe S. 1992. Nitrogenous waste excretion and accumulation of urea and ammonia in *Chalcalburnus tarichi* (Cyprinidae), endemic to the extremely alkaline Lake Van (Eastern Turkey). *Fish Physiol Biochem.* 9:377–386.
Dayarian A, Michael TP, Sengupta AM. 2010. SOPRA: scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics* 11:345.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
Eladari D, Kumai Y. 2015. Renal acid-base regulation: new insights from animal models. *Pflugers Arch.* 467:1623–1641.
Fredriksson R, Nordstrom KJ, Stephansson O, Hagglund MG, Schioth HB. 2008. The solute carrier (SLC) complement of the human genome: phylogenetic classification reveals four major families. *FEBS Lett.* 582:3811–3816.
Ganot P, Zoccola D, Tambutte E, Voolstra CR, Aranda M, Allemand D, Tambutte S. 2015. Structural molecular components of septate junctions in cnidarians point to the origin of epithelial junctions in eukaryotes. *Mol Biol Evol.* 32:44–62.
Gao S, Sung WK, Nagarajan N. 2011. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J Comput Biol.* 18:1681–1691.
Geng K, Zhang Z. 1988. Geomorphologic features and evolution of the Holocene lakes in Dali Nor Area, the Inner Mongolia (in Chinese). *J Beijing Normal Univ.* 4:94–100.
Gilmour KM, Perry SF. 2009. Carbonic anhydrase and acid-base regulation in fish. *J Exp Biol.* 212:1647–1661.
Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A.* 108:1513–1518.
Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* 33:W557–W559.
Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9:R7.
Hahn MW, Demuth JP, Han SG. 2007. Accelerated rate of gene gain and loss in primates. *Genetics* 177:1941–1949.
Hebert SC, Mount DB, Gamba G. 2004. Molecular physiology of cation-coupled Cl- cotransport: the SLC12 family. *Pflugers Arch.* 447:580–593.
Holzer P. 2009. Acid-sensitive ion channels and receptors. *Handb Exp Pharmacol.* 283–332.
Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour.* 9:1322–1332.
John SW, Krege JH, Oliver PM, Hagaman JR, Hodgin JB, Pang SC, Flynn TG, Smithies O. 1995. Genetic decreases in atrial natriuretic peptide and salt-sensitive hypertension. *Science* 267:679–681.
Jukes T, Cantor C. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammal Protein Metabolism. New York: Academic Press. p. 21–132.
Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447:714–719.
Kent WJ. 2002. BLAT–the BLAST-like alignment tool. *Genome Res.* 12:656–664.
Kiela PR, Xu H, Ghishan FK. 2006. Apical NA+/H+ exchangers in the mammalian gastrointestinal tract. *J Physiol Pharmacol.* 57 Suppl 7:51–79.
Komarova Y, Malik AB. 2010. Regulation of endothelial permeability via paracellular and transcellular transport pathways. *Annu Rev Physiol.* 72:463–493.
Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, et al. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 34:D572–D580.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
Li Y, Xu P, Zhao ZX, Wang J, Zhang Y, Sun XW. 2011. Construction and characterization of the BAC library for common carp *Cyprinus carpio* L. and establishment of microsynteny with zebrafish *Danio rerio*. *Mar Biotechnol.* 13:1183–1183.
Liu JX, Howell SH. 2010. Endoplasmic reticulum protein quality control and its relationship to environmental stress responses in plants. *Plant Cell* 22:2930–2942.
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18.
Maitre JL, Heisenberg CP. 2013. Three functions of cadherins in cell adhesion. *Curr Biol.* 23:R626–R633.
Marsden M, DeSimone DW. 2003. Integrin-ECM interactions regulate cadherin-dependent cell adhesion and are required for convergent extension in Xenopus. *Curr Biol.* 13:1182–1191.
Marshall W, Grosell M. 2006. Ion transport, osmoregulation, and acid-base balance. *Physiol Fishes.* 3:177–230.
Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35:W182–W185.
Page RD. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci.* 12:357–358.
Payne JA, Rivera C, Voipio J, Kaila K. 2003. Cation-chloride co-transporters in neuronal communication, development and trauma. *Trends Neurosci.* 26:199–206.
Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* 21 Suppl 1:i351–i358.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81:559–575.
Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, Cao C, Hu Q, Kim J, Larkin DM, et al. 2012. The yak genome and adaptation to life at high altitude. *Nat Genet.* 44:946–949.
Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res.* 33:W116–W120.
Randall D, Wood C, Perry S, Bergman H, Maloiy G, Mommsen T, Wright P. 1989. Urea excretion as a strategy for survival in a fish living in a very alkaline environment. *Nature* 337:165–166.
Romero MF, Chen AP, Parker MD, Boron WF. 2013. The SLC4 family of bicarbonate (HCO(3)(-)) transporters. *Mol Aspects Med.* 34:159–182.
Rosenberg NA. 2004. Distruct: a program for the graphical display of population structure. *Mol Ecol Notes.* 4:137–138.
Salamov AA, Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* 10:516–522.
Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff JN, Lesch KP, et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nat Genet.* 45:567–572.
Serohijos AW, Shakhnovich EI. 2014. Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. *Mol Biol Evol.* 31:165–176.
Shi SJ, Vellaichamy E, Chin SY, Smithies O, Navar LG, Pandey KN. 2003. Natriuretic peptide receptor A mediates renal sodium excretory responses to blood volume expansion. *Am J Physiol Renal Physiol.* 285:F694–F702.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56:564–577.
Tiziano V, Terova G, Romano A, Barca A, Pisani P, Storelli C, Saroglia M. 2012. The solute carrier (SLC) family series in teleost fish. In: Sargolia M, Liu A, editors. Functional genomics in aquaculture. Oxford: John Wiley, p. 219.
Torgler CN, Narasimha M, Knox AL, Zervas CG, Vernon MC, Brown NH. 2004. Tensin stabilizes integrin adhesive contacts in Drosophila. *Dev Cell.* 6:357–369.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111.
Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 7:562–578.
Tsukada T, Rankin JC, Takei Y. 2005. Involvement of drinking and intestinal sodium absorption in hyponatremic effect of atrial natriuretic peptide in seawater eels. *Zool Sci.* 22:77–85.
UniProt_Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Res.* 43:D204–D212.
Walsh PJ, Grosell M, Goss GG, Bergman HL, Bergman AN, Wilson P, Laurent P, Alper SL, Smith CP, Kamunde C, et al. 2001. Physiological and molecular characterization of urea transport by the gills of the Lake Magadi tilapia (*Alcolapia grahami*). *J. Exp. Biol.* 204:509–520.
Wang B, Ji P, Xu J, Sun J, Yang J, Xu P, Sun X. 2013. Complete mitochondrial genome of *Leuciscus waleckii* (Cypriniformes: Cyprinidae: Leuciscus). *Mitochondrial DNA* 24:126–128.
Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom Proteom Bioinformatics*. 8:77–80.
Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49.
Wright P, Anderson P. 2001. Fish physiology: nitrogen excretion. New York: Academic Press.
Westen EA, Prange HD. 2003. A reexamination of the mechanisms underlying the arteriovenous chloride shift. *Physiol Biochem Zool.* 76:603–614.
Wilkie MP, Wright PA, Iwama GK, Wood CM. 1993. The physiological responses of the Lahontan cutthroat trout (*Oncorhynchus clarki henshawi*), a resident of highly alkaline Pyramid Lake (pH 9.4), to challenge at pH 10. *J Exp Biol.* 175:173–194.
Wood C, Bergman H, Laurent P, Maina J, Narahara A, Walsh P. 1994. Urea production, acid-base regulation and their interactions in the Lake Magadi tilapia, a unique teleost adapted to a highly alkaline environment. *J Exp Biol.* 189:13–36.
Wright PA, Land MD. 1998. Urea production and transport in teleost fishes. *Comp Biochem Physiol A Mol Integr Physiol.* 119:47–54.
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. *J Paleolimnol.* 40:519–528.
Xiong L, Zhu JK. 2002. Molecular and genetic aspects of plant responses to osmotic stress. *Plant Cell Environ.* 25:131–139.
Xu J, Ji P, Wang B, Zhao L, Wang J, Zhao Z, Zhang Y, Li J, Xu P, Sun X. 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:e59703.
Xu J, Li Q, Xu L, Wang S, Jiang Y, Zhao Z, Zhang Y, Li J, Dong C, Xu P, et al. 2013. Gene expression changes leading extreme alkaline tolerance in Amur ide (*Leuciscus waleckii*) inhabiting soda lake. *BMC Genomics* 14:682.
Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, Xu J, Zheng X, Ren L, Wang G, et al. 2014. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nat Genet.* 46:1212–1219.
Xu Q, Li G, Cao L, Wang Z, Ye H, Chen X, Yang X, Wang Y, Chen L. 2012. Proteomic characterization and evolutionary analyses of zona pellucida domain-containing proteins in the egg coat of the cephalochordate, *Branchiostoma belcheri*. *BMC Evol Biol.* 12:239.
Yong L, Peirong Z, Ling L, Caixia Z, Xiangyang X, Li E. 2001. Assessment and status of water chemistry from Hutou to Raohe section of the Ussuri River (In Chinese). *J Chin Fisheries.* 14:54–57.
Zhang Y, Liang L, Jiang P, Li D, Lu C, Sun X. 2008. Genome evolution trend of common carp (*Cyprinus carpio* L.) as revealed by the analysis of microsatellite loci in a gynogentic family. *J Genet Genomics.* 35:97–103.