http://www.pjbs.org

PJBS

ISSN 1028-8880

Pakistan Journal of Biological Sciences

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Pakistan Journal of Biological Sciences 18 (4): 149-165, 2015
ISSN 1028-8880

RESEARCH ARTICLE OPEN ACCESS

DOI: 10.3923/pjbs.2015.149.165

Identification and Bioinformatics Analyses of the Basic Helix-loop-helix Transcription Factors in *Xenopus laevis*

[1,2,3]Wuyi Liu and [2]Fengmei Li
[1]Department of Science and Technology, Fuyang Normal University, Qing He West Road No. 100, Fuyang, 236037, People's Republic of China
[2]Department of Biology Sciences, Faculty of Biological and Food Engineering, Fuyang Normal University, Fuyang City, China
[3]Key Laboratory of Embryo Development and Reproductive Regulation of Anhui Province, Fuyang Normal University, Fuyang City, China

ARTICLE INFO
Article History:
Received: June 03, 2015
Accepted: July 23, 2015

Corresponding Author:
Wuyi Liu
Department of Science and Technology,
Fuyang Normal University,
Qing He West, Road No. 100, Fuyang,
236037, People's Republic of China
Tel: +86-558-2596562
Fax: +86-558-2596561

ABSTRACT
Xenopus laevis is a long established model organism for developmental, behavioral and neurological studies. Herein, an updated genome-wide survey was conducted using the ongoing genome project of *Xenopus laevis* and 106 non-redundant Basic Helix-Loop-Helix (bHLH) genes were identified in the *Xenopus laevis* genome databases. Gene Ontology (GO) enrichment statistics showed 51 significant GO annotations of biological processes and molecular functions and 5 significant KEGG pathways and a number of *Xenopus laevis* bHLH genes play significant role in specific development or special physiology processes like the development processes of muscle and eye and other organs. Furthermore, each sub-group of the bHLH family has its special gene functions except for the common GO term categories. Molecular phylogenetic analyses revealed that among these identified bHLH proteins, 105 sequences could classified into 39 families with 46, 25, 10, 5, 16 and 3 members in the corresponding high-order groups A, B, C, D, E and F, respectively with an addition bHLH member categorized as an orphan. The present study provides much useful information for further researches on *Xenopus laevis*.

Key words: *Xenopus laevis*, BHLH transcription factor, genome project, genome database, molecular phylogeny

INTRODUCTION

Transcription Factors (TFs) are frequently identified and classified into families or subfamilies based on the sequence similarity or comparability of DNA-binding domains which are highly conserved among species. Some TF families are common to most eukaryotic organisms, while others TF families are specific to particular taxonomic groups (Luscombe *et al*., 2000; Riechmann *et al*., 2000). The Basic Helix-Loop-Helix (bHLH) transcription factors constitute one of the largest families of functionally important proteins and were believed to be key regulators in cell proliferation and differentiation, cell lineage determination, the formation of muscle, neurons, gut and blood, sex determination, as well as other essential developmental and genetic processes (Atchley and Fitch, 1997; Massari and Murre, 2000; Ledent and Vervoort, 2001; Jones, 2004). Due to the interaction networks and important functions that they display in various organisms, bHLH TFs have been the subject of many researches designed to identify and characterize their functions and interactions and classification informations. The first identification of bHLH TFs was reported by Murre *et al*. (1989) who focused on the murine factors E12 and E47 and then an increasing number of bHLH TFs has been characterized by scholars, particularly in recent years. The studies of animal bHLH proteins have led to the definition of six major functional and evolutionary lineages or groups (A-F) that can be further subdivided into smaller orthologous

families named after the first discovered or best-known member (Atchley and Fitch, 1997; Ledent and Vervoort, 2001; Ledent *et al*., 2002; Jones, 2004; Skinner *et al*., 2010). Groups A and B bHLH proteins bind to E boxes (CANNTG), in which group A binds to sequences CACCTG or CAGCTG and group B binds to sequences CACGTG or CATGTTG. Group C proteins are complex molecules with one or two PAS domains following the basic helix-loop-helix motif. They bind the core sequences of ACGTG and/or GCGTG. Group D proteins lack a basic domain and form inactive heterodimers with group A bHLH proteins. Group E proteins bind preferentially to core sequences typical of N boxes (CACGCG or CACGAG). Group F lack the basic domain part and are characterized by the presence of an additional domain for DNA binding and dimerization, the COE domain. The bHLH TFs share a common bHLH motif or domain of 60 amino acids or so which holds a basic region and two helices separated by a loop (HLH) region of variable length (Massari and Murre, 2000; Ledent and Vervoort, 2001). The basic region is a DNA-binding domain. The amphipathic α-helices of two bHLH proteins can interact can interact with each other and the HLH domain promotes dimerization, allowing the formation of homodimeric or heterodimeric protein complexes between different members (Ledent and Vervoort, 2001). Atchley *et al*. (1999) computed and deduced a predictive motif for the bHLH domains based on 242 bHLH proteins, in which 19 conserved sites were found within the bHLH domain (Atchley *et al*., 1999). Their works showed that a sequence identified with less than 8 mismatches to the predictive motif was possibly a bHLH protein (Atchley and Fitch, 1997; Atchley *et al*., 1999; Atchley *et al*., 2000) and other researchers found that a sequence with nine mismatches might be a potential bHLH protein as well (Toledo-Ortiz *et al*., 2003).

With the genome resources of interested organisms being available, it would be desirable to have a more refined classification scheme of various types of bHLH motifs, as well as a better understanding of their functions and evolutionary implications within and/or among species. Recently, more and more bHLH genes has been identified and bHLH TF families have been analyzed in many organisms whose genome drafts have been available (Ledent *et al*., 2002; Toledo-Ortiz *et al*., 2003; Buck and Atchley, 2003; Heim *et al*., 2003; Li *et al*., 2006a, b; Simionato *et al*., 2007; Stevens *et al*., 2008; Wang *et al*., 2007, 2008, 2009; Zheng *et al*., 2009; Pires and Dolan, 2010; Carretero-Paulet *et al*., 2010; Liu and Zhao, 2010, 2011; Liu *et al*., 2012, 2013; Liu and Chen, 2013). However, the family of bHLH TFs has not yet been studied and characterized in *Xenopus laevis*. The *Xenopus laevis* and its congener *Xenopus tropicalis* are established model organisms for biological researches in development, behavior and neurology (Carruthers and Stemple, 2006; Bowes *et al*., 2008). The draft of *Xenopus tropicalis* genome assembly was recently accomplished by American scientists at the Lawrence Berkeley National Laboratory (Hellsten *et al*., 2010) but the *Xenopus laevis* ongoing genome project has not been accomplished because of its genomic complex. Our interest focuses on the identification and functional and evolutionary analysis of the bHLH TFs in *Xenopus laevis*. In a previous work (Liu, 2011), the preliminary survey identified 98 bHLH TFs in *Xenopus laevis* but it was proved to be incomplete. Therefore, we conducted an updated genome-wide survey here using the ongoing genome project database of *Xenopus laevis* and identified 106 bHLH sequences in its genome. In the study, we used both the predictive motif developed by Atchley *et al*. (1999) and the 45 representative bHLH domains defined by Ledent *et al*. (2002) and Simionato *et al*. (2007) to do similarity searches using BLAST algorithm in the *Xenopus laevis* genomic database and finally identified 106 bHLH proteins. We also did similarity searches using BLAST algorithm with 105 *Xenopus tropicalis* bHLH domains (Liu and Chen, 2013) (Table 1). Next, we made phylogenetic analyses of the *Xenopus laevis* bHLH factors with 118 human bHLH proteins (Simionato *et al*., 2007) which allowed us to define the *Xenopus laevis* bHLH orthologous genes. We further reported the result of gene function enrichment of the *Xenopus laevis* bHLH proteins using GO annotations.

MATERIALS AND METHODS

Similarity searches using BLAST algorithm and retrieval of bHLH proteins: We initially followed the criteria developed by Atchley *et al*. (1999) to define a bHLH protein and retrieved 8 bHLH sequences in primary searches based on the protein domain consensus sequences predicted by Atchley *et al*. (1999). The predictive motif is:

$$\text{`++}X_{(3-6)}\text{E+XR}X_{(3)}\alpha\text{N}X_{(2)}\Phi X_{(2)}\text{L+}X_{(5-22)}\text{+}X_{(2)}$$
$$KX_{(2)}\delta LX_{(2)}\ A\delta XY\alpha X_{(2)}L\text{'}$$

where, + is K, R, α is I, L, V, Φ is F, I, L, δ is I, V, T, E, R, K, A and Y are as defined, X is any residue, $X_{(I)}$ is any i residues and $X_{(i-j)}$ is i to j of any residues.

Then, we used both these initially retrieved bHLH sequences and the 45 representative bHLH domains to make BLASTP and TBLASTN and PSI-BLAST searches of bHLH domains. Each sequence was then used to perform multiple searches against the non-redundant databases of *Xenopus laevis* built by NCBI (http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid = 8355) and the *Xenopus laevis* Genome Project databases (http://xenopus.lab.nig.ac.jp/). Stringency was set to E<10 in order to obtain all bHLH-related sequences for later examination. With TBLASTN against the *Xenopus laevis* database, we obtained all putative bHLH proteins that had more than 10 conserved amino acids among the 19 residues (Toledo-Ortiz *et al*., 2003). Each sequence was used to perform second and third TBLASTN and/or PSI-BLAST searches against the *Xenopus laevis* genomic database. We also checked the Xenbase Release v3.3 (http://www.xenbase.org) (Bowes *et al*., 2010; Karpinka *et al*., 2015) used the previous obtained 105

Table 1: Information of the 105 *Xenopus tropicalis* basic helix-loop-helix transcription factors used in the article

BHLH gene family	Gene name	Protein accession	Genome contig
ASCa	*Xsash*1	XP_002944648.1	NW_003169609.1
ASCa	*Xsash*2	XP_002940290.1	NW_003163913.1
ASCb	*Xsash*3	XP_002940370.1	NW_003163927.1
MyoD	*Myf*3	NP_988972.1	NW_003166075.1
MyoD	*Myf*4	NP_001016725.1	NW_003163495.1
MyoD	*Myf*5	NP_988932.1	NW_003163331.1
MyoD	*Myf*6	NP_001017160.1	NW_003163331.1
E12/E47	*E*2A	NP_001093743.1	NW_003163736.1
E12/E47	*TCF*3	XP_002940299.1	NW_003163915.1
E12/E47	*TCF*4	NP_001096226.2	NW_003163423.1
Ngn	*Xsath*4c	NP_001116895.1	NW_003163503.1
NeuroD	*NDF1* (*neurod*1)	NP_001090868.1	NW_003163341.1
NeuroD	*NDF*2	NP_001072486.1	NW_003163936.1
NeuroD	*Xsath*2	NP_001072273.1	NW_003163914.1
NeuroD	*Xsath*3	NP_001124513.1	NW_003163487.1
Mist1	*Mist*1	XP_002931994.1	NW_003163340.1
Beta3	*Beta*3a	XP_002944506.1	NW_003167409.1
Beta3	*Beta*3b	NP_001072933.1	NW_003163515.1
Oligo	*Oligo*1	XP_002938497.1	NW_003163700.1
Oligo	*Oligo*2	XP_002938491.1	NW_003163700.1
Oligo	*Oligo*3	NP_001008191.1	NW_003163713.1
Oligo	*Oligo*4	NP_001039180.1	NW_003163795.1
Net	*Xsath*6	XP_002937330.1	NW_003163606.1
Mesp	*Mesp*1	NP_001039184.1	NW_003163348.1
Mesp	*Mesp*2	NP_001016653.1	NW_003163348.1
Mesp	*pMespo*	NP_001039104.1	NW_003163426.1
Twist	*Twist*1	NP_989415.1	NW_003163378.1
Twist	*Twist*2	NP_001096679.1	NW_003163487.1
Paraxis	*Paraxis*	NP_001016506.1	NW_003165117.1
Paraxis	*Sclerax*1	XP_002942929.1	NW_003164455.1
Paraxis	*Sclerax*2	XP_002937913.1	NW_003163647.1
MyoRa	*MyoRa*1	NP_001096235.1	NW_003163586.1
MyoRa	*MyoRa*2	NP_001103518.1	NW_003163498.1
MyoRb	*MyoRb*1	GNOMON|93674.p (*Ab initio* protein)	NW_003164157.1
MyoRb	*MyoRb*2	GNOMON|522504.p (*Ab initio* protein)	NW_003163470.1
Hand	*Hand*1	NP_001016743.1	NW_003163350.1
Hand	*Hand*2	NP_001093695.1	NW_003163380.1
PTFa	*PTF*a	NP_001095279.1	NW_003163378.1
PTFb	*PTF*b	XP_002933181.1	NW_003163373.1
SCL	*Tal*1	NP_001135468.1	NW_003163327.1
SCL	*Tal*2	XP_002934026.1	NW_003163404.1
SCL	*Lyl*1	XP_002939165.1	NW_003163774.1
NSCL	*NSCL*1	XP_002937307.1	NW_003163605.1
SRC	*SRC*1	NP_001106383.1	NW_003163796.1
SRC	*SRC*2	NP_001135631.1	NW_003163586.1
SRC	*SRC*3	XP_002933204.1	NW_003163374.1
Figα	*Figα*	NP_001016342.1	NW_003163469.1
MYC	*l-Myc*	NP_001011144.1	NW_003164143.1
MYC	*n-Myc*	NP_989390.1	NW_003163721.1
MYC	*v-Myc*	NP_001006874.1	NW_003163866.1
Mad	*Mxi*1	NP_001008129	NW_003180496.1 NW_003163820.1
Mad	*Mad*1	NP_001072228.1	NW_003163469.1
Mad	*Mad*3	NP_001017299.1	NW_003163577.1
Mad	*Mad*4	NP_001096239.1	NW_003164437.1
Mnt	*Mnt*	NP_001135494.1	NW_003163468.1
MAX	*MAX*	NP_001008208.1	NW_003163599.1
USF	*USF*1	NP_001096236.1	NW_003168160.1
USF	*USF*2	NP_001007857.1	NW_003163677.1
USF	*USF*3	NP_001120597.1	NW_003164188.1
MITF	*MITF*	NP_001093747.1	NW_003163951.1
MITF	*TFEb*	NP_001072648.1	NW_003163367.1
MITF	*TFEc*	XP_002935013.1	NW_003163447.1
MITF	*TFE*3	XP_002944430.1	NW_003166883.1

Table 1: Continue

BHLH gene family	Gene name	Protein accession	Genome contig
SREBP	*SREBP1*a	XP_002935886.1	NW_003163500.1
SREBP	*SREBP1*b	XP_002935887.1	NW_003163500.1
SREBP	*SREBP1*c	XP_002944649.1	NW_003169615.1
			NW_003163500.1
SREBP	*SREBP*2	NP_001116910.1	NW_003163395.1
AP4	*AP*4	NP_001123841.1	NW_003163353.1
Mlx	*Mondo*A	NP_001090682.1	NW_003163637.1
TF4	*TF*4	GNOMON:712044.p	NW_003164277.1
		(*Ab initio* protein)	NW_003164157.1
Clock	*Clock*	NP_001122127.1	NW_003163433.1
ARNT	*ARNT*1	NP_001116925.1	NW_003163477.1
ARNT	*ARNT*2	NP_001093686.1	NW_003163348.1
Bmal	*Bmal*2	NP_001096298.1	NW_003164805.1
AHR	*AHR*1	XP_002933348.1	NW_003163378.1
AHR	*AHR*2	XP_002935182.1	NW_003163457.1
Sim	*Sim*1	XP_002932187.1	NW_003163345.1
Sim	*Sim*2	XP_002941575.1	NW_003164120.1
Trh	*NPAS*3	NP_001072647.1	NW_003163363.1
HIF	*Hif1*α	NP_001011165.1	NW_003163817.1
HIF	*EPAS*1	NP_001005647.1	NW_003163351.1
Emc	*Id*2	NP_988885.1	NW_003163451.1
Emc	*Id*3	NP_001016271.1	NW_003163432.1
Emc	*Id*4	NP_001004839.1	NW_003163385.1
Hey	*Herp*1	NP_001007911.1	NW_003163551.1
Hey	*Herp*2	XP_002936042.1	NW_003163507.1
Hey	*HEYL*	XP_002934312.1	NW_003163416.1
H/E(spl)	*Dec*2	NP_001027504.1	NW_003163993.1
H/E(spl)	*Hes*1a	NP_001011194.1	NW_003163571.1
H/E(spl)	*Hes*1b	NP_988870.1	NW_003163533.1
H/E(spl)	*Hes*5a	NP_001037880.1	NW_003163546.1
H/E(spl)	*Hes*5b	NP_001037974.1	NW_003163546.1
H/E(spl)	*Hes*5c	NP_001039178.1	NW_003163399.1
H/E(spl)	*Hes*5d	NP_001037951.1	NW_003163399.1
H/E(spl)	*Hes*5e	NP_001107462.1	No finding
H/E(spl)	*Esr*9	NP_001037989.1	NW_003163399.1
H/E(spl)	*Hes*6	NP_001072210.1	NW_003163381.1
H/E(spl)	*Hes*7a	NP_001039166.1	NW_003164377.1
H/E(spl)	*Hes*7b	NP_001107508.1	NW_003164377.1
Coe	*EBF*1	XP_002939654.1	NW_003163834.1
Coe	*EBF*2	NP_989200.1	NW_003163356.1
Coe	*EBF*3	XP_002932694.1	NW_003163358.1
Coe	*EBF*4	XP_002932695.1	NW_003163358.1
Orphan	*Orphan*1	XP_002938975.1	NW_003163749.1
Orphan	*Orphan*4	XP_002943245.1	NW_003164609.1

Xenopus tropicalis bHLH domains (Table 1) to make BLASTP and TBLASTN and PSI-BLAST searches (Liu and Chen, 2013). Then, similar BLAST searches were conducted against the Xenbase for putative bHLH proteins too. All of the TBLASTN, BLASTP and PSI-BLAST searches were conducted with the methods and similar parameter setting-ups in the previous works (Liu and Zhao, 2010; Liu, 2011; Liu and Chen, 2013). Subsequently, redundant sequences were removed accordingly.

EST searches: In order to find putative bHLH gene and/or existing Expressed Sequence Tags (ESTs) matching the obtained *Xenopus laevis* HLH sequences, TBLASTN searches were performed against *Xenopus* Genome EST database on NCBI and Xenbase TBLASTN websites using each bHLH as the query sequence. The stringency was set as E<0.0001. A 90% or higher identity was considered to be an EST corresponding to the bHLH sequence. The obtained EST was translated into protein sequence with the EditSeq program in DNAStar version 7.1 to obtain the absent amino acid residues. In case where a query sequence composed of two or three *Xenopus laevis* coding regions, intron splice sites were assessed with the online program NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/) (Hebsgaard *et al*., 1996) to find the locations of possible introns.

Protein ID, sequence alignment and motif comparison: Protein sequence ID (accession number) was obtained by BLASTP searches against the *Xenopus laevis* protein database with the amino acid sequence of each identified bHLH domain. All of the gained sequences were finally aligned using ClustalX 2.0 (Thompson *et al*., 1997). The aligned bHLH domains were shaded using GeneDoc 2.6.02

(Nicholas and Nicholas, 1997) and copied into a RTF file for further annotation. Sequences were compared according to conserved amino acids.

Gene Ontology (GO) enrichment analysis: The Gene Ontology (GO) hierarchy annotations were downloaded from the Gene Ontology database (http://www.geneontology.org/). Enrichment for GO term categories was analyzed using DAVID Functional Annotation Bioinformatics Tools (Dennis *et al*., 2003; Huang *et al*., 2008) which reports enrichment by scores with respect to GO term categories. DAVID calculates the functional enrichment score of the same gene set based on the GO categories including biological processes, molecular functions, cellular components, KEGG pathways and other key words. In addition, it also provides a hyper-geometric p-value and a Benjamini p-value for each enrichment score.

Phylogenetic analyses of bHLH orthologous genes: Phylogenetic analyses were conducted by MRBAYES 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) and PHYML 2.4.4 (Guindon and Gascuel, 2003), with the JTT substitution frequency matrix (Jones *et al*., 1992). The obtained bHLH sequences were used to construct phylogenetic trees of Bayesian inference and maximum likelihood estimation. Maximum Likelihood (ML) analyses were performed using the frequencies of amino acids estimated from the data set and rate heterogeneity across sites modeled by one constant rate and eight γ-rates. Statistical support for the different internal branches was assessed by bootstrap resampling with 100 replicates in PHYML (Guindon and Gascuel, 2003). Phylogenetic analyses of Bayesian inference was performed with two independent Markov chains, each containing from 1200-1500 million Monte Carlo steps until the standard deviation of split frequencies was below 0.01 with sample frequency saved every 1000 generations. Finally, the trees obtained in the two runs of Markov chains were meshed and the first 25% of the trees were discarded as 'burnin' and the 50% majority consensus trees were edited and displayed using MEGA 4.0 (Tamura *et al*., 2007).

RESULTS AND DISCUSSION

Retrieval and identification of bHLH proteins: BLASTP, TBLASTN and PSI-BLAST searches with the 45 representative bHLH domains identified 106 sequences with manual improvement and examination (Table 2, Fig. 1). All of the names and related information of 106 *Xenopus laevis* bHLH proteins are listed in Table 2, in which six members were identified by EST searches. Each bHLH protein was named according to its phylogenetic relationship with the corresponding human (*Homo sapiens*) and frog (*Xenopus tropicalis*) orthologs and paralogs. Where one human or frog bHLH sequence had two or more *Xenopus laevis* orthologous genes, we used 'a', 'b' and 'c' or '1', '2' and '3' and so on, to number them. For instance, two orthologous genes of the human Hath5 and Hath4a were found in *Xenopus laevis*. Thus, the *Xenopus laevis* genes were named Xath5a and Xath5b, Xath4a1 and Xath4a2, accordingly. In this study, it was found that 106 proteins belonged to 39 families with 46, 25, 10, 5, 16 and 3 members in each of the groups A, B, C, D, E and F respectively, while one member belonged to none of these groups and was classified as an 'orphan'. Members of six families, i.e., Delilah, Mist, Net, MyoRb, PTFa and Trh, were not found in this study. In addition, sixteen hypothetical or predicted proteins were novel bHLH members identified in this study, i.e., NP_001085994.1, NP_001088572.1, NP_001079668.1, NP_001085596.1, NP_001091211.1, NP_001088421.1, NP_001154867.1, NP_001088667.1, NP_001089471.1, NP_001088134.1, NP_001088700.1, NP_001089031.1, NP_001085564.1, NP_001085718.1, NP_001087639.1 and NP_001079757.1. Moreover, three unclearly predicted proteins, i.e. Paraxis (NP_001087941.1), Id3b (NP_001079757.1) and Hes5 (NP_001079464.1) were verified and two misnamed proteins (NP_001079050.1, protein thylacine1, renamed as Mesp2a; NP_001081641.1, protein thylacine2, renamed as Mesp2b) were corrected and re-annotated by TBLASTN searches and robust phylogenetic analyses. Meanwhile, we must caution that our analyses have probably been carried out with unannotated genome assemblies or even ESTs retrieved from the *Xenopus* Genome Databases built by NCBI and the *Xenopus Laevis* Genome Project Resources and Xenbase. Therefore, it is possible that we may have missed some bHLHs or have included a few bHLH domains from some pseudogenes. Nevertheless, these data are sufficient for the purpose of this study and alignment of the 106 *Xenopus laevis* bHLH domains was shown (Fig. 1).

Phylogenetic analyses and identification of orthologous genes: Classification of human bHLH protein members has been extensively studied (Ledent *et al*., 2002; Simionato *et al*., 2007; Simionato *et al*., 2008; Wang *et al*., 2009; Zheng *et al*., 2009). Therefore, human bHLH family can be used as a good reference for orthologous gene identification of bHLH members in other organisms. Phylogenetic analysis is still regarded as an effective measure for orthologue identification by constructing phylogenetic trees using robust methods and setting an adequate standard for bootstrap values (Simionato *et al*., 2007). Determining the phylogenetic relationships of the bHLH proteins is an important step for elucidating the evolutionary and functional divergence of this gene family as well. Herein, phylogenetic analyses of Bayesian Inference (BI) and Maximum Likelihood estimate (ML) were used to identify putative orthologous sequences in different phylogenetic trees with other known bHLH members. If the unknown sequence forms a monophyletic clade with a known bHLH member or family with bootstrap value >50 in phylogenetic trees, the known member will be regarded as orthologous of the putative unknown sequence. This criterion was relaxed for the Mesp, Myc and

Table 2: Information of 106 bHLH genes from *Xenopus laevis* genome database

		Homo sapiens orthologous gene			Xenopus (Silurana) tropicalis orthologous gene			
bHLH family	Gene name	Name	ML bootstrap value (%)[a]	BI posterior marginal probability (%)[b]	Name	ML bootstrap value (%)[a]	BI posterior marginal probability (%)[b]	Protein accession number[c]
ASCa	*Xlash*1 (ascl1-A)	Hash1 (ASCL1)	88	98	Xsash1	99	1100	NP_001079247.1
ASCa	*Xlash*2	Hash2	96	95	Xsash2	79	75	NP_001085994.1
ASCb	*Xlash*3a (XASH-3)	Hash3(ASCL3) Mash3(Sgn1)	n/m*	n/m*	Xsash3	96	100	NP_001079106.1
ASCb	*Xlash*3b (ash3b-A)	Hash3(ASCL3) Mash3(Sgn1)	n/m*	n/m*	Xsash3	96	100	NP_001079125.1
MyoD	*Myf*3a	Myf3	71	60	Myf3	86	75	NP_001079366.1
MyoD	*Myf*3b	Myf3	71	60	Myf3	86	75	NP_001081292.1
MyoD	*Myf*4a	Myf4	80	94	Myf4	81	92	NP_001079326.1
MyoD	*Myf*4b	Myf4	87	94	Myf4	86	92	NP_001079199.1
MyoD	*Myf*5	Myf5	62	61	Myf5	n/m	84	NP_001095249.1
MyoD	*Myf*6a	Myf6	n/m	93	Myf6	n/m	92	NP_001081477.1
MyoD	*Myf*6b	Myf6	87	93	Myf6	n/m	92	NP_001088572.1
E12/E47	*E2*A	E2A	n/m*	59	E2A	n/m*	n/m*	NP_001080409.1
E12/E47	*TCF*3	TCF3	93	95	TCF3	87	96	NP_001079668.1
Ngn	*Xlath*4a1	Hath4a	98	100	Xsath4c	98	100	NP_001081802.1
Ngn	*Xlath*4a2	Hath4a	98	100	Xsath4c	98	100	NP_001081804.1
Ngn	*Xlath*4b	Hath4b	89	97	Xsath4c	n/m	80	NP_001128257.1
NeuroD	*NDF*1 (neurod1-A)	NDF1 (NEUROD1)	75	54	NDF1	n/m*	65	NP_001079263.1
NeuroD	*NDF*2	NDF2 (NEUROD2)	90	99	NDF2	n/m*	74	NP_001085596.1
NeuroD	*Xlath*2	Hath2	75	80	Xsath2	84	98	NP_001079218.1
NeuroD	*Xlath*3	Hath3	94	100	Xsath3	95	100	NP_001081213.1
Atonal	*Xlath*5a	Hath5	98	100	?	n/m*	n/m*	NP_001079289.1
Atonal	*Xlath*5b	Hath5	98	100	?	n/m*	n/m*	NP_001079290.1
Beta3	*Beta*3a	Beta3a	72	n/m	Beta3a	n/m*	62	BJ070553.1 EST
Oligo	*Oligo*4	Oligo2, Oligo3	n/m*	n/m*	Oligo4	99	98	Xl3.1-XL056o10.3 EST
Mesp	*Mesp*1a	Mesp1, Mesp2, pMesp1	n/m*	n/m*	Mesp1	76	81	NP_001079050.1
Mesp	*Mesp*1b	Mesp1, Mesp2, pMesp1	n/m*	n/m*	Mesp1	93	100	NP_001081641.1
Mesp	*Mesp*2a (mespa)	Mesp1, Mesp2, pMesp1	n/m*	n/m*	Mesp2	68	63	NP_001128698.1
Mesp	*Mesp*2b (mespa)	Mesp1, Mesp2, pMesp1	n/m*	n/m*	Mesp2	n/m	63	NP_001091431.1
Mesp	*pMeso*1	pMesp1	100	100	pMespo	79	90	NP_001083813.1
Mesp	*pMeso*2	pMesp1	100	100	pMespo	79	100	NP_001136111.1
Twist	*Twist*1	Twist1, Twist2	96	n/m	Twist1	89	94	NP_001079352.1
Twist	*Twist*2	Twist1, Twist2	n/m	n/m	Twist2	98	100	NP_001091211.1
Paraxis	*Paraxis*	Paraxis	73	86	Paraxis	63	86	NP_001087941.1
Paraxis	*Sclerax*	Sclerax	86	100	Sclerax1	95	99	NP_001092152.1
MyoRa	*MyoRa*1	MyoRa1	n/m	n/m	MyoRa1	n/m*	89	NP_001085957.1
MyoRa	*MyoRa*2	MyoRa2	n/m	n/m	MyoRa1	n/m*	89	gnl|ti|957022678 EST DT070535.1 EST
PTFb	*PTF*b	PTFb	87	100	PTFb	88	100	NP_001167491.1
Hand	*Hand*1	Hand1	94	100	Hand1	99	100	NP_001079128.1
Hand	*Hand*2a	Hand2	96	100	Hand2	n/m	n/m	NP_001079108.1
Hand	*Hand*2b	Hand2	96	100	Hand2	99	n/m	NP_001107665.1
SCL	*Tal*1	Tal1	65	68	Tal1	77	88	NP_001081746.1

Table 2: Continue

bHLH family	Gene name	Homo sapiens orthologous gene: Name	Homo sapiens orthologous gene: ML bootstrap value (%)[a]	Homo sapiens orthologous gene: BI posterior marginal probability (%)[b]	Xenopus (Silurana) tropicalis orthologous gene: Name	Xenopus (Silurana) tropicalis orthologous gene: ML bootstrap value (%)[a]	Xenopus (Silurana) tropicalis orthologous gene: BI posterior marginal probability (%)[b]	Protein accession number[c]
NSCL	*NSCL*1	NSCL1, NSCL2	n/m*	n/m*	NSCL	99	100	NP_001081852.1
NSCL	*NSCL*2	NSCL1, NSCL2	n/m*	n/m*	NSCL	99	100	NP_001088421.1
SRC	*SRC*2a	SRC2	97	100	SRC2	93	99	NP_001154867.1 (Uncharacterized protein LOC100301960)
SRC	*SRC*2b	SRC2	97	100	SRC2	93	99	NP_001081139.1
SRC	*SRC*3	SRC3	73	98	SRC3	n/m	100	NP_001081732.1
Figα	*Figα*	Figα	94	100	Figα	94	100	NP_001088667.1
MYC	*l-Myc*1	L-Myc1	62	100	l-Myc	100	100	NP_001081340.1
MYC	*l-Myc*2	L-Myc2	62	63	l-Myc	100	100	NP_001079460.1
MYC	*n-Myc*1	n-Myc	77	99	n-Myc	82	100	NP_001079365.1
MYC	*n-Myc*2	n-Myc	77	99	n-Myc	82	100	NP_001084122.1
MYC	*v-Myc*	v-Myc	89	94	v-Myc	81	100	NP_001080349.1
Mad	*Mxi*1	Mxi1	97	99	Mxi1	98	100	NP_001089170.1
Mad	*Mad*1	Mad1a	96	69	Mad1a	84	98	NP_001090200.1
Mad	*Mad*3	Mad3	100	100	Mad3	n/m	93	NP_001090188.1
Mad	*Mad*4a	Mad4	81	76	Mad4	n/m	100	NP_001079167.1
Mad	*Mad*4b	Mad4	81	99	Mad4	85	100	NP_001084456.1
Mnt	*Mnt*	Mnt	99	99	Mnt	n/m	93	NP_001089310.1
MAX	*MAX*1	MAX	95	100	MAX	88	100	NP_001079118.1
MAX	*MAX*2	MAX	95	100	MAX	88	100	NP_001089042.1
USF	*USF*1	USF1	97	100	USF1	68	88	NP_001089471.1
USF	*USF*2	USF2	59	100	USF2	n/m	95	NP_001088134.1
USF	*USF*3	USF3	100	100	USF3	n/m*	n/m*	NP_001088700.1
MITF	*TFE*3	TFE3	96	91	TFE3	n/m	100	NP_001088215.1
SREBP	*SREBP*2	SREBP2	89	96	SREBP2	90	100	NP_001085554.1
AP4	*AP*4	AP4	91	100	AP4	91	100	DC061935.1
Mlx	*Mondo*A	MondoA	88	98	MondoA	99	100	BQ731130.1
TF4	*TF*4	TF4	88	100	TF4	65	100	DT078575.1 EST, EE324674.1 EST, etc.
Clock	*Clock*	Clock	98	100	Clock	99	100	NP_001083854.1
ARNT	*ARNT*1	ARNT1, ARNT2	n/m*	n/m*	ARNT1	n/m	100	NP_001082130.1
ARNT	*ARNT*2a	ARNT1, ARNT2	n/m*	n/m*	ARNT2	93	n/m	NP_001080540.1
ARNT	*ARNT*2b	ARNT1, ARNT2	n/m*	n/m*	ARNT2	n/m*	100	NP_001083622.1
Bmal	*Baml*1a	Bmal1	n/m	90	Baml2	n/m	82	NP_001089024.1
Bmal	*Baml*1b	Bmal1	76	90	Baml2	n/m	82	NP_001089031.1 (Uncharacterized protein LOC503673)
AHR	*AHR*1	AHR1	94	66	AHR1	98	74	NP_001082693.1
AHR	*AHR*2	AHR2	95	71	AHR2	93	85	NP_001121349.1
Sim	*Sim*2	Sim2	86	96	Sim2	n/m	97	NP_001079101.1
HIF	*Hif1α*1	Hif1α	99	54	Hif1α	n/m	98	NP_001086426.1

Table 2: Continue

bHLH family	Gene name	Homo sapiens orthologous gene: Name	ML bootstrap value (%)[a]	BI posterior marginal probability (%)[b]	Xenopus (Silurana) tropicalis orthologous gene: Name	ML bootstrap value (%)[a]	BI posterior marginal probability (%)[b]	Protein accession number[c]
HIF	*Hif1*α2	Hif1α	99	54	Hif1α	82	98	NP_001080449.1
HIF	*EPAS*1a	EPAS1	85	89	EPAS1	n/m	98	NP_001085564.1
HIF	*EPAS*1b	EPAS1	85	89	EPAS1	84	98	NP_001085718.1
Emc	*Id*2a	Id2	79	80	Id2	67	72	NP_001087639.1
Emc	*Id*2b	Id2	79	80	Id2	n/m	56	NP_001081902.1
Emc	*Id*3a	Id3	85	100	Id3	100	100	NP_001079535.1
Emc	*Id*3b	Id3	85	100	Id3	100	100	NP_001079757.1
Emc	*Id*4	Id4	86	92	Id4	76	95	NP_001080704.1
Hey	*Herp*1	Herp1	86	79	Herp1	n/m	77	NP_001083926.1
H/E (spl)	*Hes*1a	Hes1	68	92	Hes1a, Hes1b	n/m*	n/m*	NP_001081396.1
H/E (spl)	*Hes*1b	Hes1	68	92	Hes1a, Hes1b	n/m*	n/m*	NP_001079386.1
H/E (spl)	*Hes*4a	Hes4	76	92	Hes1a, Hes1b	n/m*	n/m*	NP_001082574.1
H/E (spl)	*Hes*4b	Hes4	76	92	Hes1a, Hes1b	n/m*	n/m*	NP_001082161.1
H/E (spl)	*Hes*5a	Hes5	84	100	Hes5a, Hes5b	n/m*	n/m*	NP_001079464.1
H/E (spl)	*Hes*5b (*Hes*5.1, *Esr*1, *Esr*1b)	Hes5	84	100	Hes5a, Hes5b	n/m*	n/m*	NP_001089096.1
H/E (spl)	*Esr*2	Hes5	95	100	Hes5e	64	92	NP_001082163.1
H/E (spl)	*Esr*3	Hes5	84	100	Hes5b	94	98	NP_001089095.1
H/E (spl)	*Esr*6e	Hes5	95	100	Hes5c	89	100	NP_001081972.1
H/E (spl)	*Esr*7	Hes5	84	100	Hes5b	94	72	NP_001081974.1
H/E (spl)	*Esr*9a	Hes5	95	100	Esr9, Hes5d	n/m*	n/m*	NP_001081706.1
H/E (spl)	*Esr*9b	Hes5	95	100	Esr9, Hes5d	n/m*	n/m*	NP_001089097.1
H/E (spl)	*Esr*10a	Hes5	95	100	Esr9a, Esr9b	n/m*	n/m*	NP_001079236.1
H/E (spl)	*Hes*6	Hes2	94	97	Hes1a, Hes1b	n/m*	n/m*	NP_001116354.1
H/E (spl)	*Hes*7	Hes7	73	89	Hes7a	94	98	NP_001082175.1
Coe	*EBF*2a	EBF2	91	79	EBF2	87	81	NP_001079146.1
Coe	*EBF*2b	EBF2	91	79	EBF2	n/m	81	NP_001079147.1
Coe	*EBF*3	EBF3	n/m	n/m	EBF3	76	66	NP_001083801.1
Orphan	*Orphan*1	Orphan1	99	100	Orphan1	99	100	BG234872.1 EST

Xenopus laevis bHLH genes were named according to their human orthologous genes' names (or common abbreviations) and the reference nomenclature was mainly from the tables and additional tables provided by Ledent *et al.* (2002). Bootstrap values were inferred and transformed from phylogenetic analyses with human bHLH sequences using Bayesian inference and ML algorithm, respectively. ML bootstrap value (note a) refers the result from maximum likelihood estimate in phylogenetic analysis and BI posterior marginal probability (note b) refers the result from Bayesian inference in phylogenetic analysis. The numbers in the phylogenetic trees are converted into percentages. Note c: The accession numbers were retrieved from two resources. These numbered as 'NP' were from the RefSeq protein database and those numbered as 'XP' were from the Build protein database. All the bHLH members are organized in the order of bHLH families manifested in Table 2 of Ledent *et al*. (2002). Notes in the bracket are also gene symbols recorded in NCBI and Xenbase. The question mark means no matching; mark n/m* means none monophyletic group with single particular orthologous gene sequences but formed a monophyletic group with two or more orthologous gene sequences of the family; n/m denotes the case of lower bootstrap value estimated less than 50%

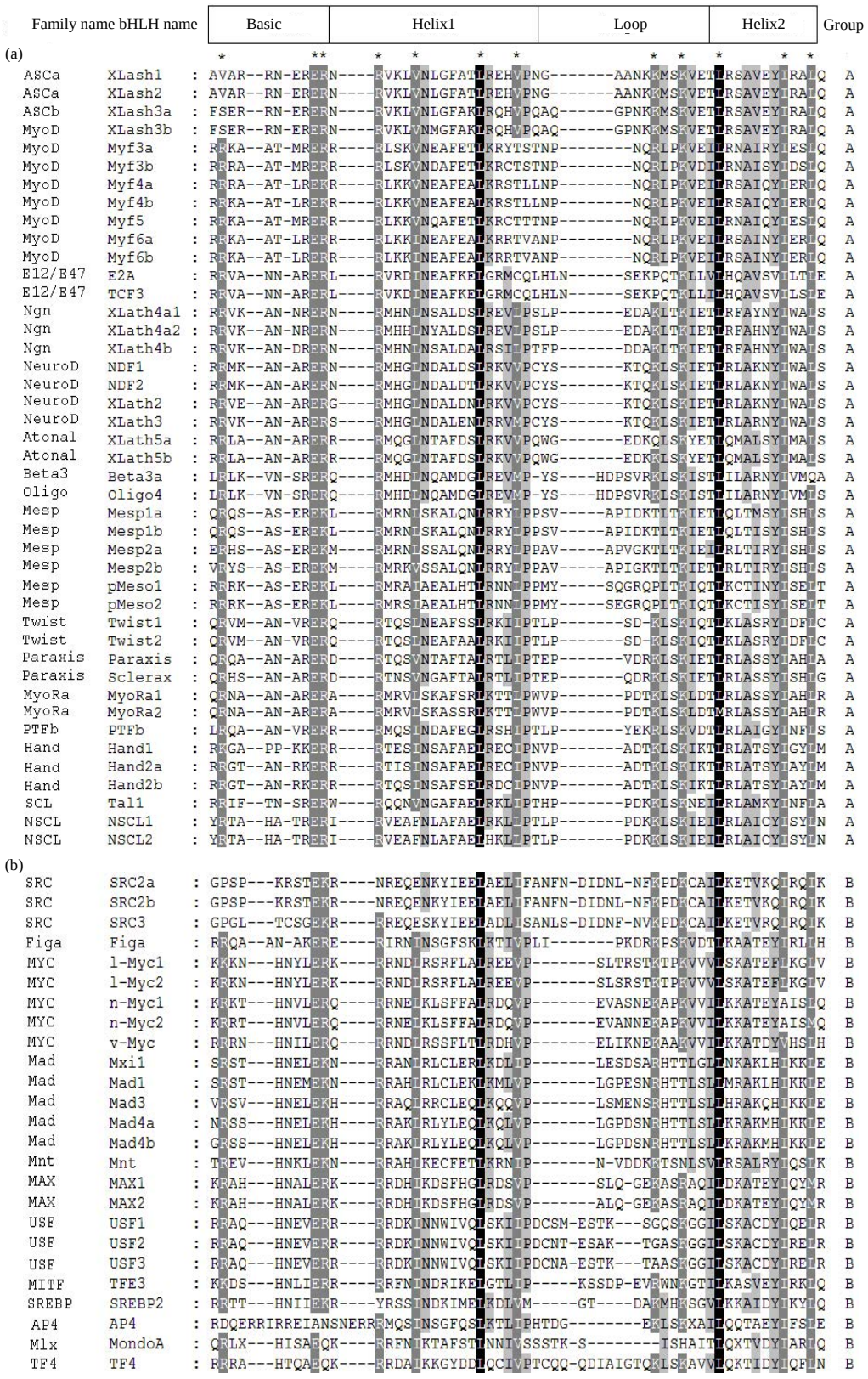

Fig. 1(a-c): Continue

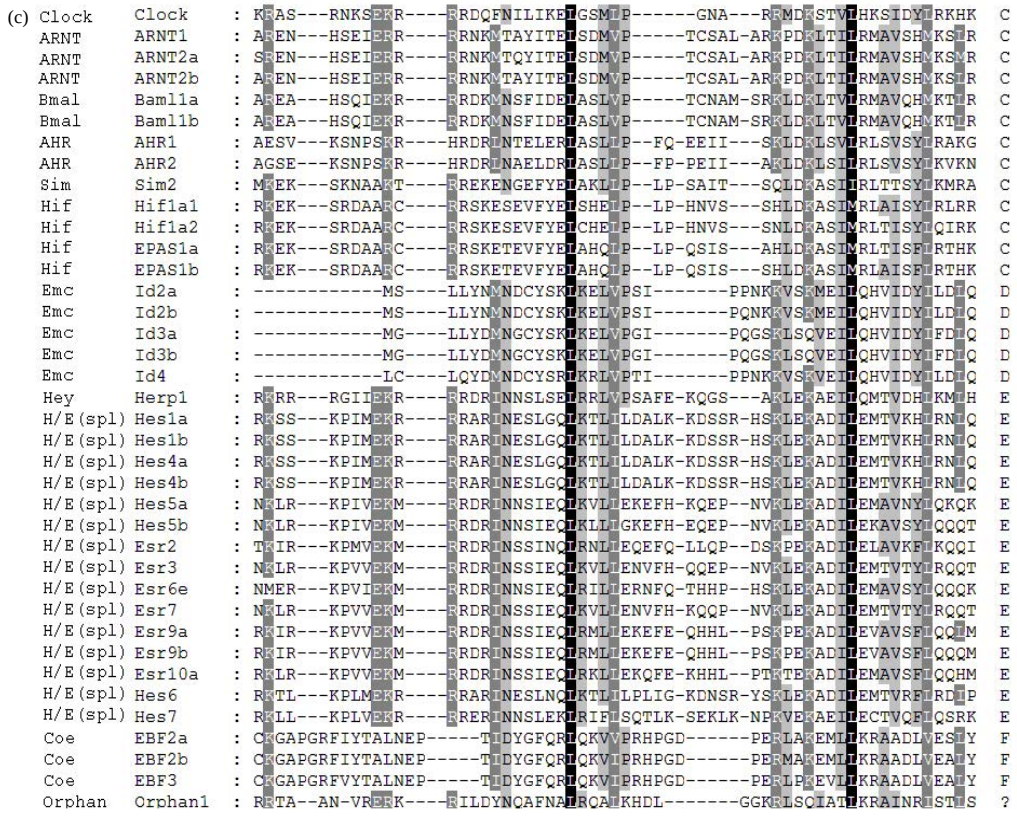

Fig. 1(a-c): Alignment of 106 *Xenopus laevis* bHLH domains shaded using Genedoc, Designation of basic, helix 1, loop and helix 2 follows (Ferre-D'Amare *et al*., 1993). Family and bHLH protein names and high-order groups were organized according to Table 2 in Ledent *et al*. (2002). Highly conserved sites are shaded and indicated with asterisks on the top

Hairy/E (spl) families (Morgenstern and Atchley, 1999; Ledent *et al*., 2002; Simionato *et al*., 2007, 2008).

In the study, the phylogenetic analyses with 118 human bHLH domains and 105 *Xenopus tropicalis* bHLH domains revealed that the 106 *Xenopus laevis* bHLH belong to 39 families in the trees of Bayesian inference and maximum likelihood estimate. The bootstrap values obtained that support the formation of a monophyletic clade with its human orthologous gene are listed in Table 2. The topologies of these BI and ML phylogenetic trees agreed well with each other, though the bootstrap values of maximum likelihood estimate was obtained through 100 replicates. Phylogenetic trees of Maximum Likelihood (ML) estimate and Bayesian inference showed the diversity of the *Xenopus laevis* bHLH family (Table 2). All the phylogenetic trees are available upon request. It was found that both human and amphibian proteomes have a number of lineage specific or unique bHLH members. For instance, from the *Xenopus laevis* proteomes, there were no findings of orthologous genes for human Mist, Hath6, MyoRb1 and MyoRb2, PTFa, NPAS3 and Id1 in this study. The *Xenopus laevis* also have multiple orthologous genes corresponding to one specific human bHLH sequence, such as Hes4a and Hes4b (orthologous genes of human Hes1); Hes5, Esr1a, Esr1b, Esr2, Esr3, Esr6e, Esr7, Esr9a and Esr9b (orthologous genes of human Hes5).

We must caution that our analysis has probably not been carried out on a comprehensive dataset of bHLHs sequences, since the *Xenopus laevis* lineage experienced a tetraploidization following divergence from the *Xenopus tropicalis* lineage and the genome project of *Xenopus laevis* is still under way. In some cases, our search for bHLH genes has been with unannotated genome assemblies or even ESTs retrieved from both the *Xenopus* Genome Resources built by NCBI and the Xenbase. Therefore, it is possible that we may have missed some bHLHs or that we may have included a few bHLH domains from some duplicates or pseudogenes (Table 2), such as gene duplicates named *Myf*3a and *Myf*3b, *Myf*4a and *Myf*4b, *Myf*5, *Myf*6a and *Myf*6b, etc. However, these data are sufficient for the purpose of this study. While a few genes showed detectable asymmetric evolution, the majority of genes analyzed in the *Xenopus laevis* genome were generally consistent with a hypothesis of symmetric change under purifying selection or strong constraint (Hellsten *et al*., 2007). In addition, there was only a subtle effect affecting on average only about 1-2 amino acids per peptide for any single gene and the

relaxed constraint experienced by retained gene duplicates was consistent with their overlapping/redundant biochemical functions (Hellsten *et al*., 2007).

Analysis of GO enrichment and Functions of bHLH proteins: DNA binding, protein dimerization and transcription coactivator activity are important functional activities of the bHLH domain. In the experiments of site-directed mutagenesis and crystal structures of bHLH proteins, it was shown that Glu-9 and Arg-12 pair constitutes the CANNTG recognition motif. The critical Glu-9 contacts the first CA of the DNA-binding motif and the role of Arg-12 is to fix and stabilize the position of Glu-9 (Fisher and Goding, 1992; Ferre-D'Amare *et al*., 1993; Ellenberger *et al*., 1994; Shimizu *et al*., 1997; Fujii *et al*., 2000). However, these experiments reported little information of the diverse functions and detailed mechanisms of bHLH proteins and their interactions. To further understand the functions of the *Xenopus laevis* bHLH gene family systematically, we analyzed the full information and experimental data currently available on the 106 *Xenopus laevis* bHLH genes by collecting the enrichment records and statistics of Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways with significant hypergeometric p-values and FDR values. Among all the GO terms retrieved, 51 were most significant GO terms ($p<0.05$) showing key cellular components, molecular functions, biological processes and 5 significant KEGG pathways for the 106 *Xenopus laevis* bHLH genes (Table 3). Furthermore, each sub-group of the bHLH family has its special gene functions ignoring

Table 3: GO enrichment of each group of *Xenopus laevis* bHLH genes

Group	Genes	GO term id	GO category	GO definition	Coherence (%)[a]	p-value
A	43	GO:0030528	MF	Transcription regulator activity	97.4	3.20E-36
		GO:0045449	BP	Regulation of transcription	97.4	4.80E-29
		GO:0007517	BP	Muscle organ development	17.9	3.60E-09
		GO:0003677	MF	DNA binding	56.4	1.20E-08
		GO:0006350	BP	Transcription	30.8	1.40E-03
B	25	GO:0030528	MF	Transcription regulator activity	100.0	8.80E-21
		GO:0045449	BP	Regulation of transcription	100.0	9.90E-17
		GO:0006350	BP	Transcription	50.0	2.60E-05
		GO:0051427	MF	Hormone receptor binding	13.6	6.40E-05
		GO:0035257	MF	Nuclear hormone receptor binding	13.6	6.40E-05
		GO:0003713	MF	Transcription coactivator activity	13.6	4.80E-04
		GO:0003677	MF	DNA binding	50.0	6.10E-04
		GO:0003712	MF	Transcription cofactor activity	13.6	1.30E-03
		KEGG Id:480088584	KEGG PATHWAY	MAPK signaling pathway	13.6	4.40E-03
		GO:0008134	MF	Transcription factor binding	13.6	5.00E-03
		GO:0006355	BP	Regulation of transcription, DNA-dependent	36.4	9.80E-03
		GO:0051252	BP	Regulation of RNA metabolic process	36.4	1.00E-02
		GO:0016563	MF	Transcription activator activity	13.6	2.50E-02
		KEGG Id:480088619	KEGG PATHWAY	Jak-STAT signaling pathway	9.1	4.50E-02
		KEGG Id:480088585	KEGG PATHWAY	ErbB signaling pathway	9.1	4.70E-02
C	13	GO:0006350	BP	Transcription	100.0	2.60E-12
		GO:0030528	MF	Transcription regulator activity	100.0	3.70E-12
		GO:0006355	BP	Regulation of transcription, DNA-dependent	100.0	1.00E-11
		GO:0003700	MF	Transcription factor activity	92.3	1.10E-11
		GO:0051252	BP	Regulation of RNA metabolic process	100.0	1.20E-11
		GO:0003677	MF	DNA binding	100.0	2.30E-10
		GO:0045449	BP	Regulation of transcription	100.0	7.50E-10
D	5	KEGG Id:480088606	KEGG PATHWAY	TGF-beta signaling pathway	100.0	6.30E-07
		GO:0030528	MF	Transcription regulator activity	100.0	1.60E-04
		GO:0045449	BP	Regulation of transcription	100.0	9.20E-04
E	16	GO:0030528	MF	Transcription regulator activity	100.0	5.00E-15
		GO:0045449	BP	Regulation of transcription	100.0	3.80E-12
		GO:0006350	BP	Transcription	87.5	2.40E-11
		GO:0006355	BP	Regulation of transcription, DNA-dependent	87.5	1.00E-10
		GO:0043425	MF	bHLH transcription factor binding	31.2	1.10E-10
		GO:0051252	BP	Regulation of RNA metabolic process	87.5	1.20E-10
		GO:0000122	BP	Negative regulation of transcription from RNA Polymerase II promoter	43.8	2.00E-10
		GO:0016564	MF	Transcription repressor activity	43.8	3.60E-10
		GO:0003677	MF	DNA binding	87.5	2.70E-09
		GO:0045892	BP	Negative regulation of transcription, DNA-dependent	43.8	8.70E-09
		GO:0051253	BP	Negative regulation of RNA metabolic process	43.8	1.30E-08
		GO:0016481	BP	Negative regulation of transcription	43.8	1.30E-08
		GO:0045934	BP	Negative regulation of nucleobase, nucleoside, Nucleotide and nucleic acid metabolic process	43.8	3.30E-08

Table 3: Continue

Group	Genes	GO term id	GO category	GO definition	Coherence (%)[a]	p-value
		GO:0010629	BP	Negative regulation of gene expression	43.8	3.30E-08
		GO:0051172	BP	Negative regulation of nitrogen compound metabolic process	43.8	3.30E-08
		GO:0010558	BP	Negative regulation of macromolecule biosynthetic process	43.8	3.50E-08
		GO:0031327	BP	Negative regulation of cellular biosynthetic process	43.8	3.80E-08
		GO:0009890	BP	Negative regulation of biosynthetic process	43.8	4.10E-08
		GO:0046982	MF	Protein heterodimerization activity	31.2	6.50E-08
		GO:0006357	BP	Regulation of transcription from RNA polymerase II promoter	43.8	1.20E-07
		GO:0010605	BP	Negative regulation of macromolecule metabolic process	43.8	1.20E-07
		GO:0008134	MF	Transcription factor binding	31.2	7.50E-07
		GO:0021915	BP	Neural tube development	25.0	3.90E-06
		GO:0007219	BP	Notch signaling pathway	25.0	7.90E-06
		GO:0033504	BP	Floor plate development	18.8	2.20E-05
		GO:0048635	BP	Negative regulation of muscle development	18.8	2.20E-05
		GO:0046983	MF	Protein dimerization activity	31.2	2.80E-05
		GO:0048634	BP	Regulation of muscle development	18.8	4.40E-05
		GO:0043010	BP	Camera-type eye development	18.8	7.30E-05
		GO:0001654	BP	Eye development	18.8	2.00E-04
		GO:0009792	BP	Embryonic development ending in birth or egg hatching	25.0	2.10E-04
		GO:0043009	BP	Chordate embryonic development	25.0	2.10E-04
		GO:0007423	BP	Sensory organ development	18.8	3.30E-04
		GO:0021501	BP	Prechordal plate formation	12.5	5.60E-03
		GO:0002088	BP	Lens development in camera-type eye	12.5	8.40E-03
		KEGG Id:480088604	KEGG PATHWAY	Notch signaling pathway	12.5	1.10E-02
		GO:0007166	BP	Cell surface receptor linked signal transduction	31.2	1.30E-02
		GO:0006916	BP	Anti-apoptosis	12.5	3.00E-02
		GO:0043066	BP	Negative regulation of apoptosis	12.5	4.40E-02
		GO:0060548	BP	Negative regulation of cell death	12.5	4.40E-02
		GO:0043069	BP	Negative regulation of programmed cell death	12.5	4.40E-02
		GO:0008283	BP	Cell proliferation	12.5	4.70E-02
		GO:0042803	MF	Protein homodimerization activity	12.5	4.90E-02
F	3	GO:0006350	BP	Transcription	100.0	1.20E-02
		GO:0030528	MF	Transcription regulator activity	100.0	1.30E-02
		GO:0006355	BP	Regulation of transcription, DNA-dependent	100.0	1.50E-02
		GO:0051252	BP	Regulation of RNA metabolic process	100.0	1.50E-02
		GO:0003677	MF	DNA binding	100.0	2.50E-02
		GO:0008270	MF	Zinc ion binding	100.0	2.60E-02
		GO:0045449	BP	Regulation of transcription	100.0	3.00E-02
		GO:0046914	MF	Transition metal ion binding	100.0	4.10E-02
Orphan	1		N/A	N/A	N/A	N/A

BP: Biological process; MF: Molecular function. The above table showed the GO annotations enriched significantly ($p<0.05$, Benjamini corrected Value) in each group. GO annotations included every layer of biological process, molecular function and cellular component category and KEGG pathway. When a GO term and its sub-layer GO both enriched in group significantly, only deeper layer GO annotation is shown in the table. All GO terms in the table were from Gene Ontology database (http://www.geneontology.org). Note a: GO coherence of each group, measured as the percentage of genes in group covered by the GO category

the common GO term categories (Table 3). Specifically, muscle organ development, neural tube development, chordate embryonic development, embryonic development ending in birth or egg hatching, floor plate development, (negative) regulation of muscle development, nuclear hormone receptor binding, hormone receptor binding, camera-type eye development, eye development, transcription coactivator activity, sensory organ development and Notch signaling pathway also have high frequencies, overlooking the frequent GO categories of transcriptional factors such as regulation of metabolism and biosynthetic processes.

It has been well known that the bHLH genes in various groups have special recognition motifs of DNA-binding sites such as E-box and G-box, etc. So what about the gene functions of each group? To explore this issue, we calculated the hyper-geometric distribution enrichment score of gene molecular functions from group A to group F based on GO annotations of GO categories including the key words of

biological process, molecular function, cellular component, KEGG pathways and so on. Only significantly enriched annotations ($p<0.05$) in deeper layers are shown (Table 3). GO statistics are listed and analyzed with a brief summary of subtypes describing each sub-group too (Table 3). In view of our analysis focused on significant GO terms for the whole frog bHLH gene family and the six bHLH sub-groups (Table 3), we found that each sub-group of bHLH TFs has its own specific GO categories when the common GO terms of transcription such as transcription regulator activity, regulation of transcription and DNA binding and protein dimerization activity are overlooked. Group A is very important in muscle organ development (Table 3). Group B is characterized with hormone receptor binding, nuclear hormone receptor binding, transcription coactivator activity, transcription cofactor activity and a few signaling pathways, i.e., MAPK pathway, Jak-STAT pathway, ErbB pathway and TGF-beta pathway. The members of group C and group D are mainly composed of common GO terms like transcription, transcription regulator activity, transcription factor activity and regulation of transcription, etc. However, group D is evolved in the TGF-beta pathway (Table 3). However, there were only a few GO terms found for members of group C, D and F identified in this study. Group E is composed of some diversified transcription regulators, the GO terms of which are functionally enriched in many aspects of transcription, such as neural tube development, floor plate development, regulation of muscle development, camera-type eye development, eye development, embryonic development ending in birth or egg hatching, chordate embryonic development, sensory organ development, prechordal plate formation, lens development in camera-type eye, cell surface receptor linked signal transduction, anti-apoptosis, cell proliferation, epithelial to mesenchymal transition, neural crest formation, mesenchymal cell development, cell morphogenesis involved in differentiation, identical protein binding, mesenchyme development, mesenchymal cell differentiation, cell morphogenesis, cellular component morphogenesis and Notch signaling pathway. There are also some special GO terms in group F, e.g. DNA binding, zinc ion binding, regulation of transcription, transition metal ion binding, metal ion binding, cation binding and ion binding (Table 3), when omitting the common GO categories of transcriptional factors.

Comparing and analyzing bHLH genes in vertebrate and invertebrate species: With genome sequence data for more and more species becoming available, it is now feasible to compare the bHLH gene family among different animal species at the genomic level. A comparison of bHLH genes in vertebrate and invertebrate species was made across six vertebrate and three invertebrate species (Table 4). Vertebrates have much more bHLH genes than invertebrates and many families in vertebrates have more members, such as E12/E47, NeuroD, Atonal, Mesp, Twist, Paraxis, SCL, SRC, Myc, Mad, MITF, HIF, Emc, Hey and Coe. Among the 45 bHLH families, only 10 families have a single member in human, zebrafish, chicken and rat and mouse respectively, while 33 and 24 families have a single member in lancelet and giant owl limpet (Table 4). It should be noted that, from our result, there are 16 families with one member detected in *Xenopus laevis*. The Delilah family is missing in vertebrate species and giant owl limpet but exists in *Drosophila* and Lancelet. It could be attributed to the birth-and-death process of bHLH gene evolution in different vertebrate and invertebrate species (Nei *et al*., 1997; Nei and Rooney, 2005).

A remarkable group is the H/E (spl) family, including diverse hairy/enhancer of split related genes (mainly Hes proteins and Hes-like factors). In the three invertebrate species, they have either 11 or 12 members, while the vertebrate species have 8-15 members in the H/E (spl) family, respectively. The phylogenetic tree of hairy/enhancer of split like orthologous genes from human, mouse, rat, zebrafish, *Xenopus laevis* and chicken was explored by a maximum likelihood method with bHLH protein sequences and the zebrafish HEYL being used as out-group (data not shown). It was found that Hes members from human, mouse, rat, zebrafish, *Xenopus laevis* and chicken form clear monophyletic groups (except human Hes4), indicating that each Hes lineage has its own ancestral sequence. Furthermore, a considerable number of bHLH genes besides the H/E (spl) family were found to have a multi-member distribution pattern in human, mouse, rat, zebrafish, chicken and *Xenopus laevis* bHLH gene families too. This case suggests that they should arise through gene duplication at least before the divergence of vertebrates from invertebrates. Among those bHLH members, many closely related genes, known as Hes, Her, or ESR and enhancer of split related genes have now been isolated from vertebrates. Like the Drosophila E (spl) genes, many of their vertebrate homologues are expressed in response to Notch activity (Campos-Ortega, 1993, 1994, 1995) and the products of these genes are essential to implement many of the cell fate decisions mediated by Notch signaling, such as the selection of cells to become neural precursors (Artavanis-Tsakonas *et al*., 1995; Greenwald, 1998). Phylogenetic analysis of the H/E (spl) family revealed that more than four gene duplication events had occurred at an early date. The objective of the next research should be designed to further determine whether accelerated rates of evolution in the H/E (spl) members or bHLH domains are due to increased positive selection or decreased constraint. Although our results provide little support for the positive selection hypothesis, this study provides us an evolutionary scenario in which the diversity of H/E (spl) gene family has been established through possibly relaxed selective constraint (Streisfeld and Rausher, 2007).

Table 4: Comparing the number of bHLH transcription factors found among vertebrate and invertebrate species

Family	Group	Fruit fly	Lancelet	*Lottie gigantea*	Human	*Xenopus tropicalis*	*Xenopus laevis*	Chicken	Zebrafish	Rat and mouse
ASCa	A	4	3	6	2	2	2	2	2	2
ASCb	A	nf	1	1	3	1	2	2	3	3
MyoD	A	1	4	1	4	4	7	4	4	4
E12/E47	A	1	1	4	4	3	2	5	5	4
Ngn	A	1	1	3	3	1	3	2	2	3
NeuroD	A	nf	1	1	4	4	4	3	5	4
Atonal	A	3	1	2	2	nf	2	3	4	2
Mist	A	1	1	1	1	1	nf	1	1	1
Beta3	A	1	1	2	2	2	1	2	3	2
Oligo	A	nf	2	3	3	4	1	2	4	3
Net	A	1	1	2	1	1	nf	1	1	1
Delilah	A	1	1	nf	nf	nf	nf	nf	nf	nf
Mesp	A	1	1	nf	3	3	6	4	5	3
Twist	A	1	1	2	2	2	2	4	3	2
Paraxis	A	1	2	1	2	3	2	3	4	2
MyoRa	A	1	4	1	2	2	1	2	2	2
MyoRb	A	nf	1	1	2	2	nf	1	2	2
Hand	A	1	1	1	2	2	3	2	1	2
PTFa	A	1	1	1	1	1	nf	1	1	1
PTFb	A	2	3	1	1	1	1	1	2	1
SCL	A	1	1	5	3	3	1	2	3	3
NSCL	A	1	1	1	2	1	2	2	1	2
SRC	B	1	1	nf	3	3	3	3	3	3
Figα	B	nf	1	nf	1	1	1	nf	1	1
Myc	B	1	1	1	3	3	5	3	6	4
Mad	B	nf	1	1	5	4	5	3	4	4
Mnt	B	1	1	1	1	1	1	1	2	1
Max	B	1	1	1	1	1	2	1	1	1
USF	B	1	1	2	3	2	3	1	2	2
MITF	B	1	1	1	5	4	1	3	5	4
SREBP	B	1	1	1	2	4	1	2	2	2
AP4	B	1	1	1	1	1	1	nf	1	1
MLX	B	1	1	7	2	1	1	3	1	2
TF4	B	1	nf	1	1	1	1	1	1	1
Clock	C	3	1	2	2	1	1	3	3	2
ARNT	C	1	1	nf	2	2	3	2	2	2
Bmal	C	1	1	nf	2	1	2	2	2	2
AHR	C	2	1	1	2	2	2	3	4	2
Sim	C	1	1	1	2	2	1	2	2	2
Trh	C	1	1	nf	1	1	nf	1	2	1
HIF	C	1	1	1	4	2	4	2	6	4
Emc	D	1	1	2	4	3	5	4	5	4
Hey	E	1	1	1	4	3	1	2	4	4
H/E(spl)	E	11	11	12	10	12	15	6	15	8
Coe	F	1	1	1	4	4	3	3	5	4
Orphan	?	nf	6	4	4	2	1	4	2	4
Total		59	78	82	118	105	106	104	139	114

The vertebrate and invertebrate species referred lancelet (B*ranchiostoma floridae*), Lottie gigantea (*Lottie gigantea*), fruit fly (*Drosophila melanogaster*), human (*Homo sapiens*), zebrafish (*Danilo rerio*), frog (*Xenopus laevis*), chicken (*Gallus gallus*), rat (*Rattus norvegicus*) and mouse (*Mus musculus*). Data on lancelet, fruit fly and human are from Simionato *et al*. (2007). Data on zebrafish, rat and mouse, chicken and *Lottie gigantea* and *Xenopus tropicalis* are from Wang *et al*. (2009), Zheng *et al*. (2009), Liu and Zhao (2010) and Liu and Chen (2013). Data on *Xenopus laevis* are from the new data mining in this study. Mark nf denotes the members of particular families were not found in the reported studies

CONCLUSION

In the study, we have identified 106 bHLH domains and their protein sequences in the *Xenopus laevis* genome database by TBLASTN and BLASTP and PSI-BLAST searches with the 45 representative bHLH domains as query sequences. Through phylogenetic analyses of the *Xenopus laevis* bHLH domains with human bHLH orthologous domains, we assigned the 106 *Xenopus laevis* bHLH genes into 39 families and an unclassified group termed "orphan". Members of six families, i.e., Delilah, Mist, Net, MyoRb, PTFa and Trh, were not found in the study. Among all the identified bHLH members, these 105 sequences could classified into 39 families with 46, 25, 10, 5, 16 and 3 members in the corresponding high-order groups A, B, C, D, E and F respectively, while one orphan member was found belonging to none of these groups. Furthermore, 16 hypothetical proteins were newly identified and annotated by computational analysis, three unclearly predicted proteins were verified and two misnamed proteins were corrected. Those uncharacterized putative bHLH proteins may be novel transcription factors needing further validation. GO enrichment statistics showed 51 significant GO

annotations and 5 significant KEGG categories counted in frequency. In addition, the GO enrichment group-analysis showed that different groups of proteins have their special gene functions when overlooking the common GO term categories.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their constructive comments and suggestions. It's jointly funded by the 2014 annual Anhui Provincial Project of Outstanding Young Talents Fund in Colleges and Universities (No.[2014]181), the projects of National Natural Science Foundation of China (No.31071310) and Anhui Provincial Natural Science Foundation (No.1308085QC63) and Anhui Provincial Educational Commission Natural Science Foundation (No. KJ2012A216).

REFERENCES

Artavanis-Tsakonas, S., K. Matsuno and M.E. Fortini, 1995. Notch signaling. Science, 268: 225-232.

Atchley, W.R. and W.M. Fitch, 1997. A natural classification of the basic helix-loop-helix class of transcription factors. Proc. Natl. Acad. Sci. USA., 94: 5172-5176.

Atchley, W.R., W. Terhalle and A. Dress, 1999. Positional dependence, cliques and predictive motifs in the bHLH protein domain. J. Mol. Evol., 48: 501-516.

Atchley, W.R., K.R. Wollenberg, W.M. Fitch, W. Terhalle and A.W. Dress, 2000. Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. Mol. Biol. Evol., 17: 164-178.

Bowes, J.B., K.A. Snyder, E. Segerdell, R. Gibb and C. Jarabek *et al*., 2008. Xenbase: A *Xenopus* biology and genomics resource. Nucl. Acids Res., 36: D761-D767.

Bowes, J.B., K.A. Snyder, E. Segerdell, C.J. Jarabek, K. Azam, A.M. Zorn and P.D. Vize, 2010. Xenbase: Gene expression and improved integration. Nucl. Acids Res., 38: D607-D612.

Buck, M.J. and W.R. Atchley, 2003. Phylogenetic analysis of plant basic helix-loop-helix proteins. J. Mol. Evol., 56: 742-750.

Campos-Ortega, J.A., 1993. Mechanisms of early neurogenesis in *Drosophila melanogaster*. J. Neurobiol., 24: 1305-1327.

Campos-Ortega, J.A., 1994. Genetic mechanisms of early neurogenesis in *Drosophila melanogaster*. J. Physiol. Paris, 88: 111-122.

Campos-Ortega, J.A., 1995. Genetic mechanisms of early neurogenesis in *Drosophila melanogaster*. Mol. Neurobiol., 10: 75-89.

Carretero-Paulet, L., A. Galstyan, I. Roig-Villanova, J.F. Martinez-Garcia, J.R. Bilbao-Castro and D.L. Robertson, 2010. Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in arabidopsis, poplar, rice, moss and Algae. Plant Physiol., 153: 1398-1412.

Carruthers, S. and D.L. Stemple, 2006. Genetic and genomic prospects for *Xenopustropicalis* research. Seminars Cell Dev. Biol., 17: 146-153.

Dennis, G., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane and R.A. Lempicki, 2003. DAVID: Database for annotation, visualization and integrated discovery. Genome Biol., 4: R60-R60.

Ellenberger, T., D. Fass, M. Arnaud and S.C. Harrison, 1994. Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev., 8: 970-980.

Ferre-D'Amare, A.R., G.C. Prendergast, E.B. Ziff and S.K. Burley, 1993. Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature, 363: 38-45.

Fisher, F. and C.R. Goding, 1992. Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. EMBO J., 11: 4103-4109.

Fujii, Y., T. Shimizu, T. Toda, M. Yanagida and T. Hakoshima, 2000. Structural basis for the diversity of DNA recognition by BZIP transcription factors. Nat. Struct. Biol., 7: 889-893.

Greenwald, I., 1998. LIN-12/Notch signaling: Lessons from worms and flies. Genes Dev., 12: 1751-1762.

Guindon, S. and O. Gascuel, 2003. A simple, fast and accurate algorithm to estimate larges phylogenies by maximum likelihood. Syst. Biol., 52: 696-704.

Hebsgaard, S.M., P.G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze and S. Brunak, 1996. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Res., 24: 3439-3452.

Heim, M.A., M. Jakoby, M. Werber, C. Martin, B. Weisshaar and P.C. Bailey, 2003. The basic helix-loop-helix transcription factor family in plants: A genome-wide study of protein structure and functional diversity. Mol. Biol. Evol., 20: 735-747.

Hellsten, U., M.K. Khokha, T.C. Grammer, R.M. Harland, P. Richardson and D.S. Rokhsar, 2007. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. BMC Biol., Vol. 5. 10.1186/1741-7007-5-31

Hellsten, U., R.M. Harland, M.J. Gilchrist, D. Hendrix and J. Jurka *et al*., 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science, 328: 633-636.

Huang, D.W., B.T. Sherman and R.A. Lempicki, 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protocols, 4: 44-57.

Huelsenbeck, J.P. and F. Ronquist, 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, 17: 754-755.

Jones, D.T., W.R. Taylor and J.M. Thornton, 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci., 3: 275-282.

Jones, S., 2004. An overview of the basic helix-loop-helix proteins. Genome Biol., Vol. 5, 10.1186/gb-2004-5-6-226

Karpinka J.B., J.D. Fortriede, K.A. Burns, C. James-Zorn, V.G. Ponferrada *et al*., 2015. Xenbase, the *Xenopus* model organism database; new virtualized system, data types and genomes. Nucl. Acids Res., 43: D756-D763.

Ledent, V. and M. Vervoort, 2001. The basic helix-loop-helix protein family: Comparative genomics and phylogenetic analysis. Genome Res., 5: 754-770.

Ledent, V., O. Paquet and M. Vervoort, 2002. Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biol., 10.1186/gb-2002-3-6-research0030

Li, J., Q. Liu, M. Qiu, Y. Pan, Y. Li and T. Shi, 2006a. Identification and analysis of the mouse basic/Helix-Loop-Helix transcription factor family. Biochem. Biophys. Res. Commun., 350: 648-656.

Li, X., X. Duan, H. Jiang, Y. Sun and Y. Tang *et al*., 2006b. Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and arabidopsis. Plant Physiol., 141: 1167-1184.

Liu, W.Y. and C.J. Zhao, 2010. Genome-wide identification and analysis of the chicken basic helix-loop-helix factors. Comp. Funct. Genomics. 10.1155/2010/682095

Liu, W.Y., 2011. Genome-wide survey, identification and preliminary analysis of *Xenopus Laevis* BHLH transcription factors. Hans J. Biomed., 1: 6-16.

Liu, W.Y. and C.J. Zhao, 2011. Molecular phylogenetic analysis of Zebra finch basic Helix-Loop-Helix transcription factors. Biochem. Genet., 49: 226-241.

Liu, A., Y. Wang, C. Dang, D. Zhang, H. Song, Q. Yao and K. Chen, 2012. A genome-wide identification and analysis of the basic helix-loop-helix transcription factors in the ponerine ant, Harpegnathos saltator. BMC Evol. Biol., Vol., 12. 10.1186/1471-2148-12-165

Liu, W.Y. and D.Y. Chen, 2013. Phylogeny, functional annotation and protein interaction network analyses of the *Xenopus tropicalis* basic helix-loop-helix transcription factors. Biomed. Res. Int. 10.1155/2013/145037

Liu, A., Y. Wang, D. Zhang, X. Wang and H. Song *et al*., 2013. Classification and evolutionary analysis of the basic helix-loop-helix gene family in the green anole lizard, Anolis carolinensis. Mol. Genet. Genomics., 288: 365-380.

Luscombe, N.M., S.E. Austin, H.M. Berman and J.M. Thornton, 2000. An overview of the structures of protein-DNA complexes. Genome Biol., Vol. 1, No. 1. 10.1186/gb-2000-1-1-reviews001

Massari, M.E. and C. Murre, 2000. Helix-loop-helix proteins: Regulators of transcription in eucaryotic organisms. Mol. Cell. Biol., 20: 429-440.

Morgenstern, B. and W.R. Atchley, 1999. Evolution of bHLH transcription factors: modular evolution by domain shuffling?. Mol. Biol. Evol., 16: 1654-1663.

Murre, C., P.S. McCaw and D. Baltimore, 1989. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD and myc proteins. Cell, 56: 777-783.

Nei, M. and A.P. Rooney, 2005. Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet., 39: 121-152.

Nei, M., X. Gu and T. Sitnikova, 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA., 15: 7799-7806.

Nicholas, K.B. and H.B. Nicholas, 1997. GeneDoc: A tool for editing and annotating multiple sequence alignments. http://65.54.113.239/Publication/3137864/genedoc-a-tool-for-editing-and-annotating-multiple-sequence-alignments.

Pires, N. and L. Dolan, 2010. Origin and diversification of basic-helix-loop-helix proteins in plants. Mol. Biol. Evol., 27: 862-874.

Riechmann, J.L., J. Heard, G. Martin, L. Reuber and C.Z. Jiang *et al*., 2000. *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. Science, 290: 2105-2110.

Ronquist, F. and J.P. Huelsenbeck, 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 12: 1572-1574.

Shimizu, T., A. Toumoto, K. Ihara, M. Shimizu and Y. Kyogoku *et al*., 1997. Crystal structure of PHO4 bHLH domain-DNA complex: Flanking base recognition. EMBO J., 16: 4689-4697.

Simionato, E., V. Ledent, G. Richards, M. Thomas-Chollier and P. Kerner *et al*., 2007. Origin and diversification of the basic helix-loop-helix gene family in metazoans: Insights from comparative genomics. BMC Evol. Biol., Vol. 7, 10.1186/1471-2148-7-33

Simionato, E., P. Kerner, N. Dray, M. Le Gouar, V. Ledent, D. Arendt and M. Vervoort, 2008. Atonal-and achaete-scute-related genes in the annelid Platynereis dumerilii: Insights into the evolution of neural basic-Helix-Loop-Helix genes. BMC Evol. Biol., 10.1186/1471-2148-8-170

Skinner, M.K., A. Rawls, J. Wilson-Rawls and E.H. Roalson, 2010. Basic helix-loop-helix transcription factor gene family phylogenetics and nomenclature. Differentiation, 80: 1-8.

Stevens, J.D., E.H. Roalson and M.K. Skinner, 2008. Phylogenetic and expression analysis of the basic helix-loop-helix transcription factor gene family: Genomic approach to cellular differentiation. Differentiation, 76: 1006-1022.

Streisfeld, M.A. and M.D. Rausher, 2007. Relaxed constraint and evolutionary rate variation between basic helix-loop-helix floral anthocyanin regulators in Ipomoea. Mol. Biol. Evol., 24: 2816-2826.

Tamura, K., J. Dudley, M. Nei and S. Kumar, 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol., 24: 1596-1599.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G. Higgins, 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res., 25: 4876-4882.

Toledo-Ortiz, G., E. Huq and P.H. Quail, 2003. The arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell, 15: 1749-1770.

Wang, Y., K. Chen, Q. Yao, W. Wang and Z. Zhu, 2007. The basic helix-loop-helix transcription factor family in *Bombyx mori*. Dev. Genes Evol., 217: 715-723.

Wang, Y., K. Chen, Q. Yao, W. Wang and Z. Zhu, 2008. The basic helix-loop-helix transcription factor family in the honey bee, *Apis mellifera*. J. Insect Sci., 8: 1-12.

Wang, Y., K. Chen, Q. Yao, X. Zheng and Z. Yang, 2009. Phylogenetic analysis of zebrafish basic helix-loop-helix transcription factors. J. Mol. Evol., 6: 629-640.

Zheng, X., Y. Wang, Q. Yao, Z. Yang and K. Chen, 2009. A genome-wide survey on basic helix-loop-helix transcription factors in rat and mouse. Mamm. Genome, 20: 236-246.