J Mol Evol (2009) 68:629–640
DOI 10.1007/s00239-009-9232-7

Phylogenetic Analysis of Zebrafish Basic Helix-Loop-Helix Transcription Factors

Yong Wang · Keping Chen · Qin Yao · Xiaodong Zheng · Zhe Yang

Received: 14 December 2008 / Accepted: 3 April 2009 / Published online: 16 May 2009

Abstract The basic helix-loop-helix (bHLH) proteins play important regulatory roles in eukaryotic developmental processes including neurogenesis, myogenesis, hematopoiesis, sex determination, and gut development. Zebrafish is a good model organism for developmental biology. In this study, we identified 139 *bHLH* genes encoded in the zebrafish genome. Phylogenetic analyses revealed that zebrafish has 58, 29, 21, 5, 19, and 5 bHLH members in groups A, B, C, D, E, and F, respectively, while 2 members were classified as "orphan." A comparison between zebrafish and human bHLH repertoires suggested that both organisms have a certain number of specific bHLH members. Eight zebrafish *bHLH* genes were found to have multiple coding regions in the genome. Two of these, *Bmal1* and *MITF*, are good anchor genes for identification of fish-specific whole-genome duplication events in comparison with mouse and chicken genomes. The present study provides useful information for future studies on gene family evolution and vertebrate development.

Keywords Basic helix-loop-helix · Phylogenesis · Transcription factor · Zebrafish

Introduction

Basic helix-loop-helix (bHLH) transcription factors have long been recognized as important regulators in various developmental processes including neurogenesis, myogenesis, hematopoiesis, sex determination, and gut development. bHLH transcription factors have a common bHLH structural motif containing a basic region and two helices separated by a loop (HLH) region of variable length (Massari and Murre 2000). The basic region acts as a DNA-binding domain, while the HLH region interacts with other bHLH sequences to form homodimers or heterodimers.

The bHLH motif has approximately 60 amino acids, among which 19 were found to be highly conserved in organisms ranging from yeast to mammals. Based on statistics of amino acid frequencies within the bHLH motif, a prediction motif for bHLH proteins was established (Atchley et al. 1999). Through examination of amino acids at the 19 highly conserved sites, more than 1000 bHLH sequences have been identified in organisms whose genome sequences are available. Among these organisms, rice and *Arabidopsis* were found to have 167 and 147 bHLH members, respectively (Li et al. 2006; Toledo-Ortiz et al. 2003). Apart from them, the human genome was found to encode 118 bHLH proteins (Simionato et al. 2007). Mouse had previously been reported to have 102 bHLH members (Ledent et al. 2002). However, our recent searches against

Electronic supplementary material The online version of this article (doi:10.1007/s00239-009-9232-7) contains supplementary material, which is available to authorized users.

Y. Wang
Department of Biotechnology, School of Food and Biological Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, People's Republic of China
e-mail: ywang@ujs.edu.cn

K. Chen (✉) · Q. Yao · X. Zheng · Z. Yang
Institute of Life Sciences, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, People's Republic of China
e-mail: kpchen@ujs.edu.cn

the latest version of the mouse genome sequence assembly have revised that figure to 114 (data to be published elsewhere). In addition, the Florida lancelet (*Branchiostoma florida*), cnidarian (*Nematodtella vectensis*), mollusk (*Lottia gigantea*), fruit fly (*Drosophila melanogaster*), and nematode (*Caenorhabditis elegans*) were found to possess 78, 68, 63, 59, and 33 bHLH members, respectively (Simionato et al. 2007).

Animal bHLH proteins have been classified into 45 orthologous families and six higher-order groups based on their phylogenetic relationships and different properties (Atchley et al. 1999; Ledent et al. 2002; Simionato et al. 2007). The 45 families were named according to the names (or common abbreviations) used when they were first reported or the names of the best-known members of the family. The six higher-order groups were named A, B, C, D, E, and F, each of which has different DNA-binding and functional properties. Briefly, groups A and B bHLH proteins bind to core DNA sequences called E boxes (CANNTG). Specifically, group A proteins bind to CACCTG or CAGCTG and group B proteins bind to CACGTG or CATGTTG. Group C proteins possess a PAS (*Drosophila*Period, human Arnt, and *Drosophila*Single-minded) domain in addition to the bHLH motif. Their target core sequence is ACGTG or GCGTG. Group D proteins do not have the basic domain. They interact with group A proteins to form inactive heterodimers. Group E proteins bind preferentially to core sequences called N boxes (CACGCG or CACGAG). They also have two additional domains, named 'Orange' and 'WRPW,' in their carboxyl termini. Group F proteins contain the COE domain, which has an additional domain functioning in both dimerization and DNA binding (Ledent and Vervoort 2001).

Zebrafish (*Danio rerio*) is a good model organism for studies on vertebrate development. Its developmental processes are similar to the embyrogenesis of higher vertebrates, including human. However, among the vast expanse of nonmammalian vertebrate species, which include fish, amphibian, reptile, and bird, only the Florida lancelet has been surveyed regarding its bHLH members (Simionato et al. 2007). Identification of bHLH members encoded in the zebrafish genome will greatly facilitate studies on vertebrate developmental biology and a variety of human congenital and genetic diseases. Although a great number of bHLH protein sequences have been deposited in NCBI (www.ncbi.nlm.nih.gov) databases of zebrafish (Adolf et al. 2004; Chong et al. 2005; Germanguz et al. 2007; Hinits et al. 2007), questions such as how many bHLH members are encoded by its genome and to which bHLH families they belong remain unanswered. Here we report the identification of zebrafish bHLH members and their phylogenetic relationships with human homologues. Moreover, their features such as distribution patterns on chromosomes and molecular phylogenesis in the evolutionary history are discussed.

Materials and Methods

tblastn Searches

Amino acid sequences of 45 representative bHLH motifs were prepared from the additional files of previous reports (Ledent and Vervoort 2001; Simionato et al. 2007). Each sequence was used to perform tblastn searches against genomic sequences of zebrafish (http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=7955). The expect value (*E*) was set as 10 in order to retrieve all bHLH-related sequences. The subject sequences obtained were manually examined to eliminate redundant ones, to add the missing amino acids on two ends of the bHLH motif, and to find introns within the bHLH motifs. Intron analysis was done using the NetGene2 application online (http://www.cbs.dtu.dk/services/NetGene2/).

Sequence Alignment

All sequences that had undergone the above examination were aligned using ClustalW online (http://www.ebi.ac.uk/clustalw/) with default settings. The aligned sequences were examined manually for their amino acid residues at the 19 conserved sites. Sequences with fewer than nine variations within the 19 sites were regarded as zebrafish bHLH members and subjected to further analyses.

Phylogenetic Analyses

Phylogenetic analyses were conducted using PAUP 4.0 Beta 10 (Swofford 1998) based on a step matrix constructed from the Dayhoff PAM250 distance matrix by R. K. Kuzoff (http://paup.csit.fsu.edu/nfiles.html). Each amino acid sequence of obtained zebrafish bHLH motifs was used to construct neighbor-joining (NJ), maximum parsimony (MP), and maximum likelihood (ML) trees with those of human bHLH motifs, respectively. NJ trees were bootstrapped with 1000 replicates to provide information about their statistical reliability. MP analysis was performed using heuristic searches and bootstrapped with 100 replicates. ML trees were constructed using TreePuzzle 5.2 (Schmidt et al. 2002). The number of puzzling steps was set to 25,000. Model of substitution was set to Jones-Taylor-Thornton (JTT; Jones et al. 1992). Other parameters were set to default values.

Identification of Protein Sequences, Genomic Contigs, Expressed Sequence Tags, and Chromosomal Locations

Protein sequence accession numbers were obtained by using the amino acid sequence of each identified zebrafish bHLH motif to conduct blastp searches against all zebrafish protein databases (including 'RefSeq protein,' 'Non-RefSeq protein,' 'Build protein,' and 'Ab initio protein'). Genomic contig numbers and number of ESTs (expressed sequence tags) were obtained using the amino acid sequences of each identified zebrafish bHLH motif to conduct tblastn search against zebrafish genome and EST sequences. All above searches used 0.01 as their *E* value and were without a filter. From all searches, only "hits" having 100% identity to the query sequence were accepted. This is because most bHLH family members are closely related. A 98% identity could very possibly refer to another bHLH member. The chromosomal location of each identified zebrafish bHLH sequence was obtained using the above-found protein sequence's accession number to search in the genome map view of zebrafish (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7955).

Results and Discussion

Zebrafish bHLH Members

The tblastn searches, sequence alignment, and examination of the 19 conserved amino acid sites revealed that 139 *bHLH* genes were encoded in the zebrafish genome, a number higher than that in human. The names of all 139 of these zebrafish bHLH members are listed in Table 1. Each zebrafish *bHLH* gene was named according to its phylogenetic relationship (explained below) with the corresponding human homologue. In the case where one human bHLH sequence has two or more zebrafish homologues, we used "a," "b," and "c" or "1," "2," and "3," etc., to number them. For instances, two homologues of the human *Tcf3* gene and the *PTFb* gene were found in zebrafish, respectively. Thus, the zebrafish genes were named *Tcf3a* and *Tcf3b*, and *PTFb1* and *PTFb2*, respectively. It was found that zebrafish has 58, 29, 21, 5, 19, and 5 bHLH members in groups A, B, C, D, E, and F, respectively. An additional two members could not be assigned to any known families and were thus regarded as "orphan." The existence of an EST sequence is a good indicator of an endogenous gene. Our EST searches revealed that 108 of the 139 zebrafish *bHLH* genes have corresponding EST sequences (data not shown), indicating a fairly high rate (77.7%) of genuine genes currently identified. ESTs of the other 31 bHLH members were not obtained, probably due to extremely low expression level in the tissues assayed. The amino acid sequences of the 139 zebrafish bHLH motifs together with their protein accession numbers are available in Supplementary File 1.

It was found that zebrafish and human each possess unique *bHLH* genes. For instance, zebrafish homologues were not found for human *Hash2*, *Hath4a*, *eHand*, *NSCL2*, *L-Myc*, *MondoA*, *NPAS1*, *Id1*, *Id3*, *Hes4*, *Orphan2*, *Orphan3*, and *Orphan4* genes. On the contrary, zebrafish either has extra members in certain bHLH families or has multiple homologues corresponding to one specific human bHLH sequence. The former includes *Beta3c*, *Oligo4*, *V-Myc1*, *V-Myc2*, *V-Myc3*, *Hif2a*, and *Hes8*. Human has only two *Beta3* genes (*Beta3a* and *Beta3b*) and three *Oligo* genes (*Oligo1*, *Oligo2* and *Oligo3*) and does not have *V-Myc*, *Hif2*, or *Hes8* genes. The latter include *Zath1a*, *Zath1b*, and *Zath1c* (homologues of human *Hath1*), *Sclerax1* and *Sclerax2* (homologues of human *Sclerax*), *PTFb1* and *PTFb2* (homologues of human PTFb), *C-Myc1* and *C-Myc2* (homologues of human *C-Myc*), *TFE3a* and *TFE3b* (homologues of human *TFE3*), *EPAS1a* and *EPAS1b* (homologues of human *EPAS1*), *Hes1a* and *Hes1b*, *Hes2a* and *Hes2b*, *Hes3a* and *Hes3b*, *Hes5a*, *Hes5b*, *Hes5c*, *Hes5d* and *Hes5e* (homologues of human *Hes1*, *Hes2*, *Hes3*, and *Hes5* respectively), *Orphan1a* and *Orphan1b* (homologues of human *Orphan1*), *pMeso1b* and *pMeso1c* (homologues of human *pMeso1*), and *Id2b* and *Id2c* (homologues of human *Id2*) (Table 1).

Homologue Identification of Zebrafish *bHLH* Genes

Classification of human bHLH family members has been extensively studied (Ledent et al. 2002; Simionato et al. 2007). Thus, human bHLH members can be used as a good reference for homologue identification of bHLH members in other organisms. Although orthologue identification has been accompanied by much uncertainty since there is no absolute criterion that can be used to decide whether two genes are orthologous (Ledent and Vervoort 2001), by constructing phylogenetic trees using various methods and setting an adequate standard for bootstrap values, phylogenetic analysis has remained an effective measure for homologue identification (Simionato et al. 2007). Furthermore, in our previous studies (Wang et al. 2007, 2008), in-group phylogenetic analysis was adopted to identify homologues for the unknown sequences that would form a monophyletic clade among themselves. An in-group phylogenetic analysis uses a single unknown bHLH sequence to construct different phylogenetic trees with other known bHLH members of the same group. If the unknown sequence forms a monophyletic clade with another known member and the bootstrap value is >50 in various phylogenetic trees, the known member will be regarded as a homologue of the unknown sequence.

Table 1 A complete list of bHLH genes from zebrafish (*Danio rerio*)

Gene name	bHLH Family	Human homolog	Bootstrap values			Protein accession Number(s)	Genomic contig Number(s)	Group
			NJ	MP	ML			
Zash1a	ASCa	*Hash1*	95	81	99	NP_571294.1	NW_001878907.1	A
Myf4	MyoD	*Myf4*	99	83	80	NP_571081.1	NW_001877150.1	A
Myf6	MyoD	*Myf6*	78	91	78	NP_001003982.1	NW_001878915.1	A
E2A	E12/E47	*E2A*	95	92	78	AAI00123.1	NW_001878252.1	A
Tcf3a	E12/E47	*Tcf3*	73	77	72	XP_694512.3	NW_001878342.1	A
Zath4c	Ngn	*Hath4c*	68	60	69	NP_571116.1	NW_001877445.1	A
Zath2b	NeuroD	*Hath2*	67	67	56	NP_571892.2	NW_001878739.1	A
Zath3	NeuroD	*Hath3*	99	95	94	NP_739568.1	NW_001878411.1	A
Zath1a	Atonal	*Hath1*	100	97	96	NP_571166.1	NW_001879350.1	A
Zath1b	Atonal	*Hath1*	96	98	94	XP_001335283.1	NW_001877486.1	A
Zath1c	Atonal	*Hath1*	65	67	91	hmm204194	NW_001877219.1	A
Zath5	Atonal	*Hath5*	99	100	94	NP_571707.1	NW_001877341.1	A
Mist1	Mist	*Mist1*	100	100	73	NP_001071120.1	NW_001877224.1	A
Beta3a	Beta3	*Beta3a*	91	74	79	NP_957249.1	NW_001878487.1	A
Beta3b	Beta3	*Beta3b*	61	61	53	NP_001025304.1	NW_001878407.1	A
Oligo1	Oligo	*Oligo1*	94	93	98	NP_001018632.1	NW_001879471.1	A
Oligo3	Oligo	*Oligo3*	91	69	98	NP_001103863.1	NW_001877318.1	A
Zath6	Net	*Hath6*	100	100	98	NP_001073460.1	NW_001877433.1	A
pMeso1a	Mesp	*pMeso1*	99	100	96	NP_878302.1	NW_001878881.1	A
Paraxis	Paraxis	*Paraxis*	79	53	61	NP_571047.1	NW_001879058.1	A
Sclerax1	Paraxis	*Sclerax*	97	100	100	XP_696212.3	NW_001877671.1	A
Sclerax2	Paraxis	*Sclerax*	96	96	100	NP_001076538.1	NW_001877936.1	A
MyoRa1	MyoRa	*MyoRa1*	78	83	80	XP_684279.3	NW_001878484.1	A
MyoRa2	MyoRa	*MyoRa2*	75	85	77	NP_001032770.1	NW_001878414.1	A
MyoRb1	MyoRb	*MyoRb1*	100	100	100	hmm488994	NW_001878487.1	A
MyoRb2	MyoRb	*MyoRb2*	93	82	99	XP_001342593.1	NW_001879128.1	A
PTFa	PTFa	*PTFa*	100	100	100	NP_997524.1	NW_001878670.1	A
PTFb1	PTFb	*PTFb*	100	100	100	hmm42634	NW_001877947.1	A
PTFb2	PTFb	*PTFb*	62	64	86	XP_686970.2	NW_001879481.1	A
Tal1	SCL	*Tal1*	100	59	94	NP_998402.1	NW_001878334.1	A
Lyl1	SCL	*Lyl1*	76	89	77	XP_001921822.1	NW_001880867.1	A
SRC1	SRC	*SRC1*	85	47	88	XP_691744.3	NW_001878142.1	B
SRC2	SRC	*SRC2*	99	97	100	NP_571852.1	NW_001878484.1	B
SRC3	SRC	*SRC3*	97	66	54	XP_692938.3	NW_001877149.1	B
Figa	Figα	*Figa*	98	70	56	NP_944601.2	NW_001880190.1	B
N-Myc	Myc	*N-Myc*	98	74	65	NP_997779.1	NW_001878142.1	B
C-Myc1	Myc	*C-Myc*	98	92	72	NP_571487.2	NW_001878479.1	B
C-Myc2	Myc	*C-Myc*	99	98	68	NP_956466.1	NW_001878672.1	B
Mad1	Mad	*Mad1*	96	80	72	XP_698607.2	NW_001878990.1	B
Mad4	Mad	*Mad4*	100	98	94	XP_687970.3	NW_001878040.1	B
Mnt1	Mnt	*Mnt*	100	100	78	NP_001096581.1	NW_001878236.1	B
Mnt2	Mnt	*Mnt*	100	100	73	XP_001919581.1	NW_001877558.1	B
Max	Max	*Max*	100	100	96	NP_571295.1	NW_001878142.1	B
USF1	USF	*USF1*	85	67	86	NP_956590.1	NW_001877557.1	B
USF2	USF	*USF2*	87	53	57	NP_001116257.1	NW_001877951.1	B
TFEb	MITF	*TFEb*	85	57	94	XP_690778.2	NW_001877150.1	B
TFEc	MITF	*TFEc*	99	99	98	NP_001025276.2	NW_001878892.1	B
TFE3a	MITF	*TFE3*	63	65	79	NP_571923.2	NW_001879334.1	B
TFE3b	MITF	*TFE3*	69	67	77	NP_001038531.1	NW_001877151.1	B
SREBP2	SREBP	*SREBP2*	97	99	96	NP_001082935.1	NW_001878860.1	B
AP4	AP4	*AP4*	100	100	97	XP_696700.3	NW_001881177.1	B
TF4	TF4	*TF4*	100	100	93	NP_001019394.1	NW_001878804.1	B
ARNT2	ARNT	*ARNT2*	96	87	96	NP_571749.1	NW_001879240.1	C

Table 1 continued

Bmal1	Bmal	*Bmal1*	100	97	85	NP_571652.1 NP_840085.1	NW_001878587.1 NW_001879283.1	C
Bmal2	Bmal	*Bmal2*	85	51	69	NP_571653.1	NW_001877867.1	C
AHR1a	AHR	*AHR1*	62	67	80	NP_001019987.1	NW_001878325.1	C
Hif3a	HIF	*Hif3a*	52	53	80	AAQ94179.1	NW_001877558.1	C
EPAS1a	HIF	*EPAS1*	95	86	69	XP_695262.3	NW_001877242.1	C
EPAS1b	HIF	*EPAS1*	97	86	70	NP_001034895.1	NW_001877330.1	C
Clock1a	Clock	*Clock1*	94	67	68	NP_571032.1	NW_001878137.1	C
Sim2	Sim	*Sim2*	96	79	95	NP_571911.1	NW_001877059.1	C
NPAS3a	Trh	*NPAS3*	99	82	68	hmm270314	NW_001877433.1	C
Id2a	Emc	*Id2*	98	92	94	NP_958448.1	NW_001877776.1	D
Id2d	Emc	*Id2*	81	78	82	NP_694499.1	NW_001877781.1	D
Herp1	Hey	*Herp1*	96	79	82	NP_997726.1	NW_001877966.1	E
Herp2	Hey	*Herp2*	96	52	93	NP_571697.1	NW_001878143.1	E
Hey4	Hey	*Hey4*	100	100	96	NP_996948.1	NW_001878026.1	E
Dec1	H/E(spl)	*Dec1*	58	54	96	NP_997844.2	NW_001877163.1	E
Hes1a	H/E(spl)	*Hes1*	99	89	94	NP_571948.1	NW_001878409.1	E
Hes1b	H/E(spl)	*Hes1*	99	92	94	NP_571154.1	NW_001879145.1	E
Hes2a	H/E(spl)	*Hes2*	85	85	91	XP_001919504.1	NW_001879376.1	E
Hes2b	H/E(spl)	*Hes2*	81	44	84	NP_001038818.1	NW_001879376.1	E
Hes3	H/E(spl)	*Hes3*	98	73	61	NP_571155.1	NW_001879374.1	E
Hes5a	H/E(spl)	*Hes5*	99	82	59	hmm1163593	NW_001880041.1	E
Hes5b	H/E(spl)	*Hes5*	98	82	71	Not found	NW_001877177.1 NW_001877177.1	E
Hes5c	H/E(spl)	*Hes5*	98	95	68	NP_991182.1	NW_001878409.1	E
Hes5d	H/E(spl)	*Hes5*	99	92	69	NP_571165.2	NW_001878409.1 NW_001878409.1	E
Hes5e	H/E(spl)	*Hes5*	99	90	69	NP_001096598.1	NW_001878409.1 NW_001878409.1 NW_001878409.1	E
Orphan1a	Orphan	*Orphan1*	100	100	100	XP_001921552.1	NW_001877070.1	?
Orphan1b	Orphan	*Orphan1*	100	100	100	hmm106414	NW_001877563.1	?
Zash1b	ASCa	*Hash1*	73	84	n/m*	NP_571306.1	NW_001879271.1	A
Zash3c	ASCb	*Hash3c*	65	71	n/m*	hmm21624	NW_001880340.1	A
TF12	E12/E47	*TF12*	89	n/m*	78	NP_999981.1	NW_001879275.1	A
Twist1a	Twist	*Twist1*	53	n/m*	54	NP_571059.1	NW_001877947.1	A
Tcf3b	E12/E47	*Tcf3*	73	n/m*	67	NP_571169.1	NW_001878252.1	A
Sclerax3	Paraxis	*Sclerax*	52	79	n/m*	XP_001340709.1	NW_001878331.1	A
Tal2	SCL	*Tal2*	81	64	n/m*	NP_958496.2	NW_001878287.1	A
SREBP1	SREBP	*SREBP1*	96	52	n/m*	NP_001098599.1	NW_001880610.1	B
Clock1b	Clock	*Clock1*	85	52	n/m*	NP_840080.1	NW_001878026.1	C
AHR1b	AHR	*AHR1*	n/m*	65	70	NP_571339.1	NW_001878325.1	C
Hif1a	HIF	*Hif1a*	84	n/m*	66	NP_956527.1	NW_001878137.1	C
Hes7	H/E(spl)	*Hes7*	90	n/m*	55	NP_571153.1	NW_001879050.1	E
Oligo2	Oligo	*Oligo2*	n/m*	51	n/m*	NP_835201.1	NW_001879471.1	A
pMeso1b	Mesp	*pMeso1*	n/m*	60	n/m*	XP_001920244.1 XP_001344193.1	NW_001878629.1 NW_001878629.1	A
pMeso1c	Mesp	*pMeso1*	n/m*	58	n/m*	AAI63806.1	NW_001879245.1	A
Zash3b	ASCb	*Hash3b*	50	n/m*	n/m*	hmm722874	NW_001878313.1	A
Tcf4	E12/E47	*Tcf4*	52	n/m*	n/m*	hmm191684	NW_001881747.1	A
Twist1b	Twist	*Twist1*	64	n/m*	n/m*	NP_001017820.1	NW_001877651.1	A
dHand	Hand	*dHand*	90	n/m*	n/m*	NP_571701.2	NW_001878058.1	A
Mad3	Mad	*Mad3*	51	n/m*	55	NP_957350.1	NW_001877435.1	B
MITF	MITF	*MITF*	91	n/m*	n/m*	NP_570998.1 NP_571922.1	NW_001879152.1 NW_001878385.1	B
NPAS2	Clock	*NPAS2*	n/m*	75	n/m*	NP_840084.1	NW_001878785.1	C

Table 1 continued

Sim1	Sim	*Sim1*	n/m*	79	n/m*	XP_695026.3	NW_001878884.1 NW_001878882.1	C
Dec2	H/E(spl)	*Dec2*	n/m*	n/m*	59	NP_001034196.1	NW_001877867.1	E
EBF1b	Coe	*EBF1*	63	n/m*	n/m*	XP_685990.3	NW_001878225.1	F
EBF3	Coe	*EBF3*	85	n/m*	n/m*	NP_001074071.1	NW_001877266.1	F
Myf3	MyoD	*Myf3*	47	61	61	NP_571337.2	NW_001878601.1	A
Zath4b	Ngn	*Hath4b*	46	64	n/m*	AAI62120.1	NW_001877341.1	A
NDF1	NeuroD	*NDF1*	35	n/m*	n/m	NP_571053.1	NW_001879479.1	A
NDF2	NeuroD	*NDF2*	48	34	70	AAI62129.1	NW_001877223.1	A
Zath2a	NeuroD	*Hath2*	46	44	66	NP_571891.1	NW_001878501.1	A
ARNT1	ARNT	*ARNT1*	50	33	n/m*	NP_001038736.1	NW_001877637.1	C
AHR1c	AHR	*AHR1*	39	62	n/m*	NP_001029092.1	NW_001878704.1	C
NPAS3b	Trh	*NPAS3*	51	n/m	n/m	XP_687851.3	NW_001877547.1	C
Hif1b	HIF	*Hif1a*	43	n/m*	51	XP_001337610.1	NW_001877343.1	C
Id2b	Emc	*Id2*	53	47	85	NP_955835.1	NW_001878142.1	D
Id2c	Emc	*Id2*	48	67	54	NP_571320.1	NW_001877149.1	D
Id4	Emc	*Id4*	43	55	62	NP_001035079.1	NW_001877642.1	D
HEYL	Hey	*HEYL*	42	n/m*	n/m*	NP_859425.1	NW_001877969.1	E
Hes6	H/E(spl)	*Hes6*	16	16	n/m*	Not found	NW_001881773.1	E
EBF1a	Coe	*EBF1*	93	41	n/m*	XP_693680.2	NW_001877460.1	F
EBF4	Coe	*EBF4*	n/m*	10	n/m*	XP_684384.1	NW_001877766.1	F
Zash3a	ASCb	*?*	n/m*	n/m*	n/m*	hmm720874	NW_001878313.1	A
Myf5	MyoD	*?*	n/m*	n/m*	n/m*	NP_571651.1	NW_001878915.1	A
Beta3c	Beta3	*?*	n/m*	n/m*	n/m*	hmm711444	NW_001879347.1	A
Oligo4	Oligo	*?*	n/m*	n/m	n/m*	NP_955808.1	NW_001877339.1	A
Mesp1	Mesp	*?*	n/m*	n/m*	n/m*	AAF72811.1	NW_001879245.1	A
Mesp2	Mesp	*?*	n/m*	n/m*	n/m*	XP_001344235.2	NW_001878629.1	A
						XP_001340039.2	NW_001878629.1	
Dermo1	Twist	*?*	n/m*	n/m*	n/m*	NP_571060.2	NW_001878384.1	A
NSCL1	NSCL	*?*	n/m*	n/m*	n/m*	NP_991232.1	NW_001879471.1	A
V-Myc1	Myc	*?*	n/m*	n/m*	n/m*	NP_001038607.1	NW_001877966.1	B
V-Myc2	Myc	*?*	n/m*	n/m*	n/m*	NP_998102.1	NW_001877350.1	B
V-Myc3	Myc	*?*	n/m*	n/m*	n/m*	XP_688836.2	NW_001879193.1	B
Mxi1	Mad	*?*	n/m*	n/m*	n/m*	NP_571312.1	NW_001878356.1	B
Mlx	Mlx	*?*	n/m*	n/m*	n/m*	XP_001338503.1	NW_001879047.1	B
AHR2	AHR	*?*	n/m*	n/m*	n/m*	NP_571103.1	NW_001877659.1	C
Hif2a	HIF	*?*	n/m	n/m*	n/m	NP_001012371.1	NW_001878223.1	C
Hes8	H/E(spl)	*?*	n/m*	n/m*	n/m*	XP_693641.2	NW_001877545.1	E
EBF2	Coe	*?*	n/m*	n/m*	n/m*	NP_571493.1	NW_001879047.1	F

Note: Zebrafish *bHLH* genes were named according to their human homologues. Bootstrap values were from in-group phylogenetic analyses with human bHLH sequences using NJ, MP, and ML algorithms, respectively. *n/m* none monophyletic. *n/m** an individual zebrafish bHLH sequence did not form a monophyletic group with another single bHLH sequence of a known family, but formed a monophyletic group with two or more bHLH sequences of *the same family*. Although multiple protein sequences were found for one individual zebrafish bHLH motif, only one protein accession number is listed for each *bHLH* gene. The accession numbers are from four different protein resources. Those labeled as "*NP*," "*XP*," and "*AA*" are from 'RefSeq protein,' 'Non-RefSeq protein,' and 'Build protein' databases, respectively. All sequences of these three types can be accessed at general NCBI search Web pages (www.ncbi.nlm.nih.gov). Those labeled as "*hmm*" are from the 'Ab initio protein' database, sequence of which may be accessed at the zebrafish genome map view Web page (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7955). All genomic contig numbers were obtained through tblastn searches against the assembly of 'reference' only. Group assignment is in accordance with Table 1 of Ledent et al. (2002) (Color Table online)

In this study, each identified zebrafish bHLH sequence was used to conduct in-group phylogenetic analyses with human bHLH members. The bootstrap values obtained that support the formation of a monophyletic clade with its human homologue are listed in Table 1. Table 1 indicates that the bootstrap support for identifying zebrafish bHLH sequences as homologues of specific human bHLH members varied greatly. First, among all the 139 zebrafish bHLH members, 80 have all NJ, MP, and ML bootstrap values >50 (ranging from 53 to 100), enabling us to confidently assign

corresponding human homologues for them (Table 1; very pale beige background). Second, 12 zebrafish bHLH members have two of the three bootstrap values >50 and one n/m* (see explanation in Table 1, Note) in the constructed phylogenetic trees (Table 1; pale aqua background), while 14 other members have only one of the three bootstrap values >50 and have two n/m* (pale mauve background). Although these 26 zebrafish bHLH members did not have sufficient bootstrap support, we assigned the corresponding homologues for them by considering that fish and human are relatively distant species and the above-set criterion can be relaxed in certain cases as Smionarto et al. did when they made phylogenetic analyses of Mesp, Myc, and H/E(spl) family members (Ledent and Vervoort 2001). Third, there are 16 zebrafish bHLH members that have one or two bootstrap values <50 and/or have one n/m* or, in a few cases, have bootstrap values as low as 10 and 16 (Table 1; aqua background). Yet we defined homologues for them because most of the values had supported the formation of a monophyletic clade with the same human counterpart. However, these assignments can be regarded as arbitrary and are subject to modification upon acquisition of new data. Finally, there was no bootstrap support information available for identifying human homologues for the other 17 zebrafish bHLH members, because none of them formed any monophyletic clade with known human bHLH members (Table 1; light-purple background). This is very possibly because human and fish diverged from their common ancestor very early in their evolutionary history. Therefore, these zebrafish bHLH members have quite low sequence similarity to human bHLH sequences and thus could not form a monophyletic clade in our phylogenetic analyses. It is anticipated that an increased number of identified bHLH sequences in higher animals such as amphibian, reptile, and bird will facilitate final homologue identification of these sequences, because some "missing pieces" can probably be found in those organisms and can thus establish clear phylogenetic relationships of fish bHLH sequences with human ones. Among these 17 bHLH members, *Beta3c*, *Oligo4*, *V-Myc1*, *V-Myc2*, *V-Myc3*, *Hif2a*, and *Hes8* are probably zebrafish specific sequences, because they are extra members of certain bHLH families that have not been found in other animals examined so far. For instance, only two *Beta3* genes have been found in other animals and were named *Beta3a* and *Beta3b*. In zebrafish, a third Beta3 member was found apart from the identification of *Beta3a* and *Beta3b* homologues. Therefore, the extra member was named *Beta3c*. The rest of these 17 bHLH members were temporarily named according to relevant human bHLH names. They were not extra bHLH members found in zebrafish. Their homology with bHLH members in other species awaits further analysis when new data are available.

Protein Sequences and Genomic Contigs of Zebrafish *bHLH* Genes

Protein sequence accession numbers and their genomic contig numbers for all 139 zebrafish bHLH motifs are also listed in Table 1. It should be noted that when the amino acid sequence of an individual zebrafish bHLH motif was used to conduct blastp searches against zebrafish various protein databases ('RefSeq protein,' 'Non-RefSeq protein,' 'Build protein,' and 'Ab initio protein'), generally a considerable number of "hits" with 100% identity in the bHLH motif could be obtained. These protein sequences often varied slightly in length. Yet most of them were not different protein sequences encoded in the zebrafish genome, because most tblastn searches using the amino acid sequence of each zebrafish bHLH motif against the zebrafish genome yielded only one coding region in the genome. One of the exceptions is seen in *Bmal1*, *pMeso1b*, *MITF*, and *Mesp2*, which codes for two different proteins and has separate genomic locations, respectively (Table 1; those with two protein accession numbers for one *bHLH* gene). The other exception is seen in *Hes5b*, *Hes5d*, *Hes5e*, and *Sim1*; it codes for only one protein sequence but has two or three separate genomic locations, respectively (Table 1; those with two or three genomic contig numbers for one *bHLH* gene). Considering this, it seems not very economical to use the zebrafish genome to code bHLH proteins because, of the 139 *bHLH* genes, these 8 genes were found to be multiple-copy genes. This figure is much higher than that in insects and mammals. In the silkworm (*Bombyx mori*) and honeybee (*Apis mellifera*), one and three *bHLH* genes were found as two-copy genes, respectively (Wang et al. 2007, 2008). In mouse, all 114 *bHLH* genes are single-copied, and in rat and human only one *bHLH* gene was found to be two-copied, respectively (data to be published elsewhere).

Chromosomal Locations of Zebrafish *bHLH* Genes

Chromosomal locations of all zebrafish *bHLH* genes are shown in Fig. 1. It can be seen that zebrafish *bHLH* genes are distributed in a rather uneven pattern. Chromosome 23 has 12 bHLH protein-coding regions, while each of chromosomes 11, 20, 21, and 22 has 8 and each of chromosomes 4, 7, 8, 13, 14, 15, and 19 has 6 or 7 coding regions, respectively. All other chromosomes were found to have two to five *bHLH* gene coding regions. In addition, chromosomal locations for eight zebrafish *bHLH* genes were not found, probably because the genomic sequences containing these genes have not been assembled into chromosomes (Fig. 1; 'not placed').

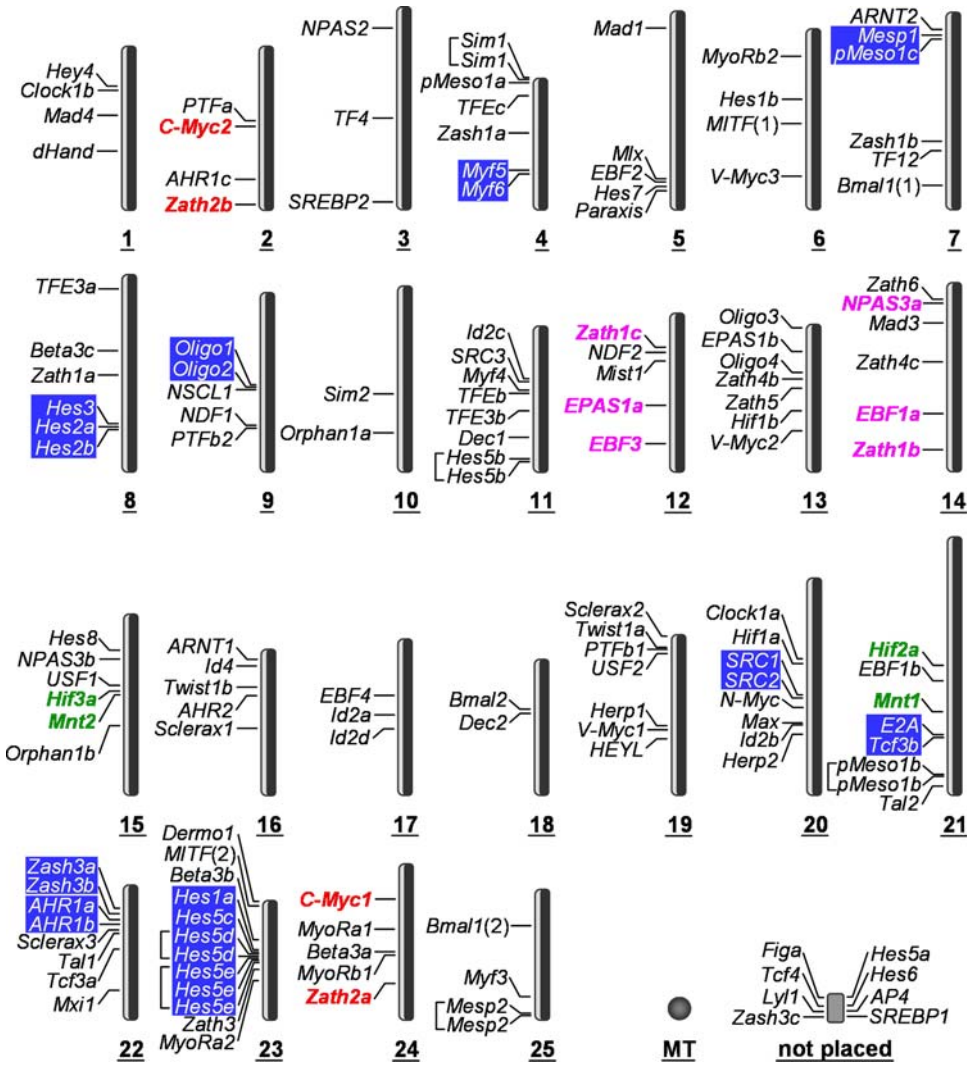

Fig. 1 Chromosomal locations of zebrafish bHLH genes. Gene names shaded in blue boxes are *bHLH* genes that belong to the same family and are clustered on the chromosome. A bracket ([) preceding a gene name means that the gene has multiple coding regions on the chromosome, except for *Bmal1* and *MITF*, which are not located on the same chromosome and are thus labeled *Bmal1*(1) and *Bmal1*(2), which are on chromosomes 7 and 25, and *MITF*(1) and *MITF*(2), which are on chromosomes 6 and 23, respectively. Gene names in red, light-purple, and green letters indicate closely related genes on separate chromosomes. Family information on each *bHLH* gene is listed in Table 1. *MT* mitochondrion (Color figure online)

It should be noted that two, three, or four zebrafish *bHLH* genes which belong to the same family are found to cluster on the chromosome. A total of 21 zebrafish *bHLH* genes fall into this category (Fig. 1; in blue boxes). For instances, *Myf5* and *Myf6* cluster on chromosome 4; *Hes2a*, *Hes2b*, and *Hes3* cluster on chromosome 8; and *Hes1a*, *Hes5c*, *Hes5d* (two copies), and *Hes5e* (three copies) cluster on chromosome 23. Figure 1 also shows the existence of the above-mentioned multiple coding regions for eight zebrafish *bHLH* genes, i.e., *Bmal1*, *pMeso1b*, *MITF*, *Mesp2*, *Hes5b*, *Hes5d*, *Hes5e*, and *Sim1*, all of which are marked with a square bracket before them except *Bmal1* and *MITF*, which have separate coding regions on different chromosomes and are thus labeled as *Bmal*(1) and *Bmal1*(2) and as *MITF*(1) and *MITF*(2).

Molecular Evolution of Zebrafish *bHLH* Genes

The above analyses revealed that the whole genome of zebrafish has coding regions for 139 *bHLH* genes, 8 of which have multiple copies. Given that human has only 118 *bHLH* genes (Simionato et al. 2007), how did this higher number of *bHLH* genes arise? It has been thought that two rounds of whole-genome duplication (WGD), i.e., the 2R hypothesis, have played an important role in the establishment of gene repertoires in vertebrates (Skrabanek and Wolfe 1998). In addition, a third round of fish-specific WGD (3R) was suggested according to observations of differences in *Hox* gene clusters and other duplicated genes between fish (tetraodon, fugu, medaka, and zebrafish) and birds/rodents (Amores et al. 1998, 2004; Naruse et al. 2004; Panopoulou and Poustka 2005; Woods et al. 2000). Evidence for 3R came from the comparative approach conducted on the pufferfish genome using the human genome as the unduplicated reference (Jaillon et al. 2004). In this approach, anchor genes which exist as a single copy in unduplicated genomes and as multiple copies on separate chromosomes in duplicated genomes were important indicators of potential duplication events (Kellis et al. 2004; Panopoulou and Poustka 2005). Among the 139

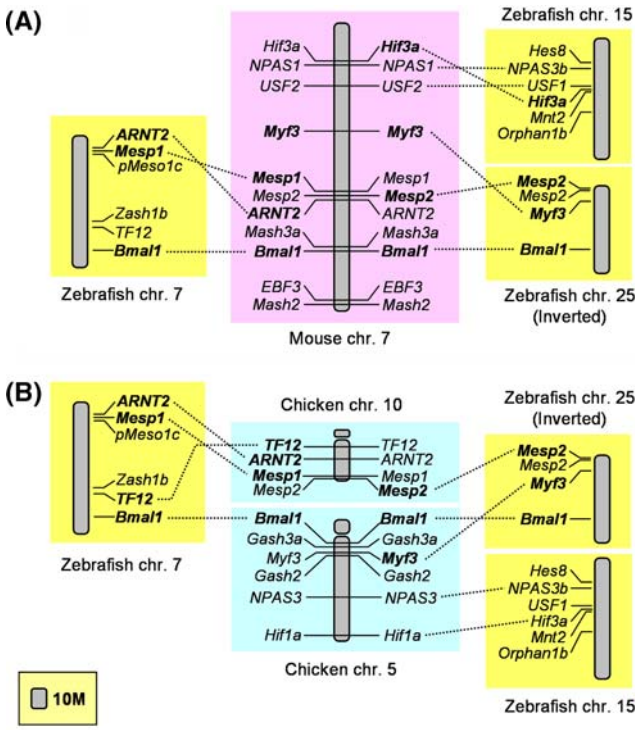

Fig. 2 *Bmal1* as anchor gene in explorations of a genome duplication event using mouse chromosomes (**a**) and chicken chromosomes (**b**) as the unduplicated reference. Zebrafish chromosomes are shown in yellow boxes. Those of mouse and chicken are in light-purple and light-aqua boxes, respectively. All chromosomes are drawn to scale. The very small gray box in the bottom-left corner represents 10 million base pairs. For clearness of labeling, genes on mouse and chicken chromosomes have been put on both sides of the chromosomes. Identical genes that appear on both zebrafish and mouse/chicken chromosomes are shown in *boldface* and are connected by *dotted lines*. Those that are not identical but are homologous genes are shown in regular typeface and are connected by *dotted lines* (Color figure online)

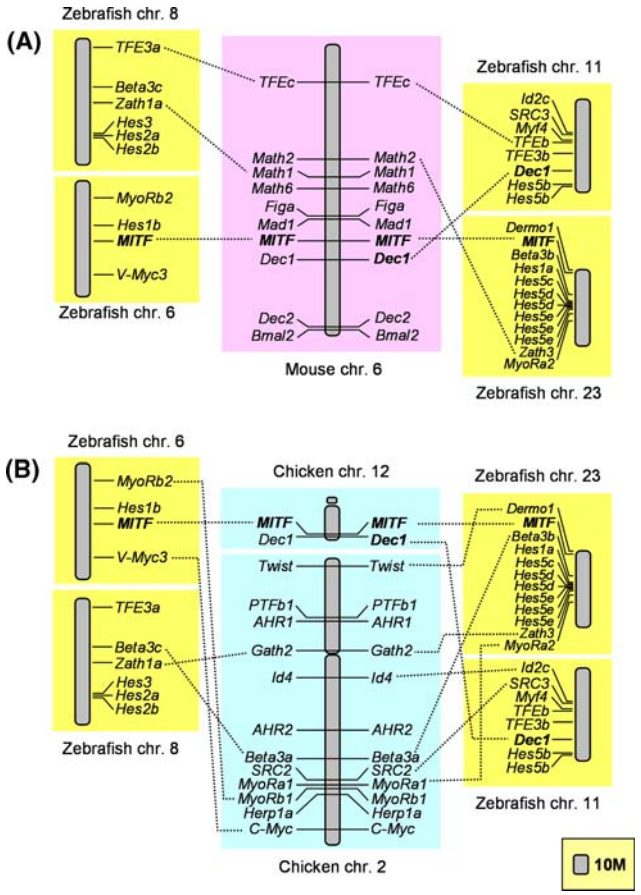

Fig. 3 *MITF* as anchor gene in explorations of a genome duplication event using mouse chromosomes (**a**) and chicken chromosomes (**b**) as the unduplicated reference. See the legend to Fig. 2 for details (Color figure online)

zebrafish *bHLH* genes identified in our study, *Bmal1* and *MITF* were found to exist on separate chromosomes (Fig. 1). Close examination enabled us to conclude that both are anchor genes from a duplication event.

First, *Bmal1* is the anchor gene between the zebrafish genome and the mouse/chicken genomes (Fig. 2). Figure 2 shows that zebrafish chromosomes 7 and 25 are the product of a duplication event, since both have *Bmal1* and *Mesp* genes on them. Their counterparts in an unduplicated reference genome are chromosome 7 in mouse and chromosomes 5 and 10 in chicken (Fig. 2a, b). Mouse chromosome 7 and zebrafish chromosome 7 have three anchor genes, i.e., *Bmal1*, *ARNT2*, and *Mesp1*. And mouse chromosome 7 and zebrafish chromosome 25 also have three anchor genes, i.e., *Bmal1*, *Myf3*, and *Mesp2*. In addition, zebrafish chromosome 15 was found to have *Hif3a*, *USF1*, and *NPAS3b* genes for which homologous genes were found on mouse chromosome 7, suggesting that zebrafish chromosome 15 was once connected with chromosome 7 or 25 and was separated later in the evolutionary process (Fig. 2a). The chicken *Bmal1* gene is located on chromosome 5. Looking at other *bHLH* genes on this chromosome, only the *Myf3* gene was found to exist on zebrafish chromosome 25. Other anchor genes such as *Mesp1* and *ARNT2* were found on chromosome 10. Therefore, both chicken chromosome 5 and chicken chromosome 10 were considered as the unduplicated reference. Four anchor genes, i.e., *Bmal1*, *Mesp1*, *ARNT2*, and *TF12*, were found on zebrafish chromosome 7. And three anchor genes, i.e., *Bmal1*, *Myf3*, and *Mesp2*, were found on zebrafish chromosome 25, while two pairs of homologous genes, i.e., *NPAS3*/*NPAS3b* and *Hif1a*/*Hif3a*, were found on chicken chromosome 5 and zebrafish chromosome 15, respectively (Fig. 2b).

Second, *MITF* is also an anchor gene between the zebrafish genome and the mouse/chicken genomes (Fig. 3). Figure 3 shows that zebrafish chromosomes 6, 8, 11, and 23 are the product of an ancient duplication event. Their counterparts in unduplicated reference genomes are

Table 2 A comparison of the number of bHLH family members in lancelet, zebrafish, chicken, and mouse

Family	Group	Lancelet	Zebrafish	Chicken	Mouse
ASCa	A	3	2	2	2
ASCb	A	1	3	2	3
MyoD	A	4	4	4	4
E12/E47	A	1	5	3	4
Ngn	A	1	2	2	3
NeuroD	A	1	5	4	4
Atonal	A	1	4	2	2
Mist	A	1	1	1	1
Beta3	A	1	3	2	2
Oligo	A	2	4	2	3
Net	A	1	1	1	1
Delilah	A	1	0	0	0
Mesp	A	1	5	3	3
Twist	A	1	3	3	2
Paraxis	A	2	4	2	2
MyoRa	A	4	2	2	2
MyoRb	A	1	2	1	2
Hand	A	1	1	1	2
PTFa	A	1	1	1	1
PTFb	A	3	2	2	1
SCL	A	1	3	2	3
NSCL	A	1	1	2	2
SRC	B	1	3	3	3
Figa	B	1	1	n/f	1
Myc	B	1	6	3	4
Mad	B	1	4	3	4
Mnt	B	1	2	1	1
Max	B	1	1	1	1
USF	B	1	2	2	2
MITF	B	1	5	3	4
SREBP	B	1	2	2	2
AP4	B	1	1	n/f	1
MLX	B	1	1	2	2
TF4	B	0	1	1	1
Clock	C	1	3	3	2
ARNT	C	1	2	2	2
Bmal	C	1	2	2	2
AHR	C	1	4	4	2
Sim	C	1 (2?)	2	3	2
Trh	C	1 (2?)	2	1	1
HIF	C	1 (2?)	6	2	4
Emc	D	1	5	4	4
Hey	E	1 (6?)	4	3	4
H/E(spl)	E	11 (16?)	15	11	8
Coe	F	1	5	3	4
Orphan	–	6	2	1	4
Total		78	139	104	114

Note: Data on lancelet are from Simionato et al. (2007). Data on zebrafish and chicken are from this study. Data on mouse are from our unpublished findings. Family names and group assignment followed Table 1 of Ledent et al. (2002)

chromosome 6 in mouse and chromosomes 2 and 12 in chicken, because anchor genes such as *MITF* and *Dec1* and a number of homologous genes exist among these chromosomes (Fig. 3a, b; gene names connected by dotted lines).

The distribution pattern of all anchor genes shown in Figs. 2 and 3 can be regarded as clear evidence for 3R occurring after zebrafish diverged from its common ancestor with chicken and mouse. This result is consistent with other observations in fish (Jaillon et al. 2004).

Apart from the existence of the above anchor genes, the distribution pattern of other closely related zebrafish *bHLH* genes also suggests an origination through chromosomal duplication. For example, the C-*Myc2* and *Zath2b* genes are on chromosome 2, and their closely related genes *C-Myc1* and *Zath2a* genes are on chromosome 24 (Fig. 1; gene names in red). More examples are seen on chromosomes 12 and 14 and on chromosomes 15 and 21 (Fig. 1; gene names in light purple and green, respectively). These distribution patterns of *bHLH* genes should not be regarded as random, and can be considered as additional evidence for the WGD hypothesis.

Molecular Evolution of Chicken and Mouse *bHLH* Genes

As discussed above, the chicken and mouse genomes can be used as an unduplicated reference for determining the duplication event in zebrafish. However, this nonduplication is only relative to a fish-specific duplication event. As the 2R hypothesis suggests, chicken and mouse *bHLH* genes should also be the products of ancient duplication events. This has been largely evident because, among the 45 bHLH families, only 11 and 10 families have a single member in chicken and mouse, respectively, while 33 families have a single member in lancelet, an ancestor of vertebrates (Table 2). Therefore, a WGD event should have occurred during the evolutionary stage of lancelet into jawless fish or cartilaginous fish. To prove this, data from two aspects are desirable. One is the distribution pattern of *bHLH* genes on lancelet chromosomes, because a specific lancelet chromosome may have suitable anchor genes and

becomes a good unduplicated reference for chicken/mouse chromosomes. Other evidence may come from the identification of *bHLH* genes in genomes of hagfish or shark, both of which are expected to have a relatively high number of *bHLH* genes.

If this earlier duplication event did happen, it would mean that the present zebrafish genome has undergone at least two rounds of WGD after the lancelet emerged (Panopoulou and Poustka 2005). Is this possible? Because the lancelet has 78 *bHLH* genes, two rounds of WGD would yield at least 200 *bHLH* genes, even after considering intensive gene loss after the WGD. The present zebrafish genome only encodes for 139 *bHLH* genes. Therefore, it seems unlikely that two rounds of WGD could have occurred in the zebrafish genome. However, the zebrafish genome has multiple coding regions of eight *bHLH* genes (Fig. 1). This is very rare among the vertebrates examined so far. The chicken and mouse genomes are found to have only one coding region for each *bHLH* gene, while those of rat and human have two coding regions for only one *bHLH* gene, respectively (data not shown). As transcriptional regulators, it is not reasonable for a *bHLH* gene to have multiple coding regions. Therefore, the multiple coding regions in the zebrafish genome could merely be the redundant copies that have not yet been lost, probably due to their relatively 'recent' origination. (A preliminary list of *bHLH* genes encoded in the chicken *Gallusgallus* genome was obtained as a reference for this study. The amino acid sequences of 104 chicken bHLH motifs together with their protein or EST accession numbers are provided in Supplementary File 2.)

Conclusion

In this study, 139 *bHLH* genes were found in zebrafish. Among them, 12 were newly identified to be encoded in the genome. All zebrafish bHLH members have been defined by their names and families according to various phylogenetic analyses with human bHLH homologues. Phylogenetic analysis has been an effective measure for homologue identification (Atchley and Fitch 1997; Ledent et al. 2002). It is much more reliable than that based on comparison of sequence similarity. Therefore, the names and family information in this report can be used to correct inadequate annotations made for previously identified zebrafish bHLH homologues, most of which were based on sequence similarity comparison. For instance, the zebrafish rotein numbered AAI00123.1 was denoted Tcf3 in GenBank, butt our phylogenetic analyses clearly indicate that it is the homologue of human E2A (Table 1).

Human and zebrafish have their own species-specific *bHLH* genes. We found that 13 human *bHLH* genes have no zebrafish homologues, and 24 zebrafish genes have no human homologues. Eight zebrafish *bHLH* genes were found to have multiple coding regions in the genome. Among them, *Bmal1* and *MITF* are good anchor genes for identification of a fish-specific WGD event in comparison with the mouse and chicken genomes. The identification of zebrafish bHLH family members and investigation of their significance in gene evolutionary events provide useful information for studies on vertebrate development and for related studies in amphibians, reptiles, birds, and other fish species.

Acknowledgments We are grateful to Professor Bin Chen and two anonymous reviewers for constructive comments on the manuscript. This work was supported by grants from the Jiangsu Sci-Tech Support Project—Agriculture (No. BE2008379) and the China National "863" Project (No. 2008AA10Z145).

References

Adolf B, Bellipanni G, Huber V, Bally-Cuif L (2004) atoh1.2 and beta3.1 are two new bHLH-encoding genes expressed in selective precursor cells of the zebrafish anterior hindbrain. Gene Expr Patterns 5:35–41

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714

Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait JH (2004) Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Res 14:1–10

Atchley WR, Fitch WM (1997) A natural classification of the basic helix-loop-helix class of transcription factors. Proc Natl Acad Sci USA 94:5172–5176

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

Chong SW, Nguyen TT, Chu LT, Jiang YJ, Korzh V (2005) Zebrafish id2 developmental expression pattern contains evolutionary conserved and species-specific characteristics. Dev Dyn 234:1055–1063

Germanguz I, Lev D, Waisman T, Kim CH, Gitelman I (2007) Four twist genes in zebrafish, four expression patterns. Dev Dyn 236:2615–2626

Hinits Y, Osborn DP, Carvajal JJ, Rigby PW, Hughes SM (2007) Mrf4 (myf6) is dynamically expressed in differentiated zebrafish skeletal muscle. Gene Expr Patterns 7:738–745

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–957

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624

Ledent V, Vervoort M (2001) The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis. Genome Res 11:754–770

Ledent V, Paquet O, Vervoort M (2002) Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biol 3:RESEARCH0030

Li X, Duan X, Jiang H, Sun Y, Tang Y, Yuan Z, Guo J, Liang W, Chen L, Yin J, Ma H, Wang J, Zhang D (2006) Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. Plant Physiol 141:1167–1184

Massari ME, Murre C (2000) Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. Mol Cell Biol 20:429–440

Naruse K, Tanaka M, Mita K, Shima A, Postlethwait J, Mitani H (2004) A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. Genome Res 14:820–828

Panopoulou G, Poustka AJ (2005) Timing and mechanism of ancient vertebrate genome duplications—the adventure of a hypothesis. Trends Genet 21:559–567

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Simionato E, Ledent V, Richards G, Thomas-Chollier M, Kerner P, Coornaert D, Degnan BM, Vervoort M (2007) Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. BMC Evol Biol 7:33

Skrabanek L, Wolfe KH (1998) Eukaryote genome duplication—Where's the evidence? Curr Opin Genet Dev 8:694–700

Swofford DL (1998) PAUP*: Phylogenetic Analysis Using Parsimony, version 4. Sinauer Associates, Sunderland, MA

Toledo-Ortiz G, Huq E, Quail PH (2003) The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell 15:1749–1770

Wang Y, Chen KP, Yao Q, Wang WB, Zhu Z (2007) The basic helix-loop-helix transcription factor family in *Bombyx mori*. Dev Genes Evol 217:715–723

Wang Y, Chen KP, Yao Q, Wang WB, Zhu Z (2008) The basic helix-loop-helix transcription factor family in the honeybee, Apis mellifera. J Insect Sci 8:insectscience.org/8.40

Woods IG, Kelly PD, Chu F, Ngo-Hazelett P, Yan YL, Huang H, Postlethwait JH, Talbot WS (2000) A comparative map of the zebrafish genome. Genome Res 10:1903–1914