ISSN 1996-3351

Asian Journal of Biological Sciences

http://knowledgiascientific.com

Knowledgia SCIENTIFIC
A Place to Publish Outstanding Research

Asian Journal of Biological Sciences 4 (8): 601-608, 2011
ISSN 1996-3351 / DOI: 10.3923/ajbs.2011.601.608

Comparative Studies on Human and Rat Basic Helix-loop-helix Proteins

[1]Nuzhat N. Kabir, [1]Mohammad I. Hossain and [1,2]Julhash U. Kazi
[1]Laboratory of Computational Biochemistry, KN Biomedical Research Institute, Bagura Road, Barisal, Bangladesh
[2]Quality Control Section, Opsonin Pharma Ltd., Bagura Road, 8200 Barisal, Bangladesh

Corresponding Author: Julhash U. Kazi, Quality Control Section, Opsonin Pharma Ltd., Bagura Road, 8200, Barisal, Bangladesh Tel: +880-431-64074 Fax: +880-431-64075

ABSTRACT

The basic Helix-Loop-Helix (bHLH) proteins are a family of transcription factors that are found in almost all eukaryotic genomes. This family of proteins regulates a variety of biological processes including embryonic development. Here we describe the catalog of human, mouse and rat bHLH proteins. Extensive *in silico* studies identified 121 human, 121 mouse and 120 rat bHLH proteins. Human has orthologs for 119 rat bHLH proteins. Two bHLH proteins are found only in human, while human lacks one rodent specific bHLH protein, MYCS. Fifty bHLH protein families are present in all three lineages. Forty nine bHLH proteins contain 16 additional domains. Orthologous bHLH pairs vary in sequence conservation along their length, creating a map of functionally important regions for every bHLH protein pair. Many species-specific sequence inserts exist. This study provides a comprehensive overview of bHLH proteins.

Key words: Basic helix-loop-helix, transcription factor, protein family, embryonic development

INTRODUCTION

The basic Helix-Loop-Helix (bHLH) proteins constitute an ancient superfamily of transcription regulators which are found in organisms from yeast to human. These transcription factors function in critical development processes, including sex determination and the development of muscles as well as the nervous system (Jones, 2004). The bHLH protein contains a conserved domain of approximately 60 amino acids that consists of α DNA-binding basic region followed by two a-helices separated by a variable loop region. The amino-terminal basic region binds to the DNA at a consensus hexanucleotide sequence known as E-box, while the HLH region interacts with other bHLH proteins to form homodimers or heterodimers. Recent, studies using evolutionary relationships classified bHLH proteins into six major groups (A-F) (Atchley and Fitch, 1997; Wang *et al*., 2007). This classification was performed using only the bHLH motif and led to the postulation of six distinct groups based on amino-acid patterns and E-box-binding specificity.

Different E-box consensus sequences are recognized by different group of bHLH proteins. For example, CACCTG or CAGCTG sequence is recognized by group A members while CACGTG or CATGTTG is recognized by group B members (Jones, 2004). The PAS domain containing C group members recognize ACGTG or GCGTG region. The HLH region facilitates interactions with other protein subunits to form homo-dimeric or hetero-dimeric complexes. Group D proteins lack the basic domain and thus are incapable to bind DNA, while form protein-protein dimmers that function as

antagonists of group A proteins. E group proteins recognize N-box sequences (CACGCG or CACGAG). In addition to bHLH domain, E group proteins contain a characteristic domain, referred to as orange domain. F group proteins also lack DNA-binding basic region but contain an additional COE domain which is involved in dimerization and DNA binding.

Recent studies identified hundreds of bHLH genes in organisms whose genome sequences were available. These studies include 8 yeast, 16 *Amphimedon queenslandica*, 33 *Hydra magnipapillata*, 39 chicken, 39 *Brachydanio rerio*, 39 *Caenorhabditis elegans*, 46 *Ciona intestinalis*, 47 *Xenopus laevis*, 50 *Tribolium castaneum*, 50 *Strongylocentrotus purpuratus*, 52 *Bombyx mori*, 57 *Daphnia pulex*, 59 *Drosophila melanogaster*, 63 *Lottia gigantean*, 64 *Capitella* sp. I, 68 *Nematostella vectensis*, 78 *Branchiostoma floridae*, 87 *Tetraodon nigroviridis*, 114 rat, 118 human, 124 mouse, 147 Arabidopsis and 167 rice bHLH proteins (Li *et al*., 2006a, b; Satou *et al*., 2003; Simionato *et al*., 2007; Skinner *et al*., 2010; Stevens *et al*., 2008; Toledo-Ortiz *et al*., 2003; Wang *et al*., 2007; Zheng *et al*., 2009). Based on phylogenetic analyses to the available bHLH proteins, 45 families were identified for all the bHLH genes (Simionato *et al*., 2007). Here we describe the complete catalog of human, mouse and rat bHLH proteins. We compare human bHLH proteins with that of rat.

MATERIALS AND METHODS

Known bHLH proteins and bHLH domains from diverse genomes were used to search human, mouse and rat genetic sequence databases. All publicly available databases including cDNA, EST, gene models and genome assembly were searched using BLAST package. PSI-BLAST was used with an e-value threshold of 0.0001 and h-value of 0.1 for five iterations. The non-redundant set of bHLH proteins was created using MySql (www.mysql.com) relational databases followed by manual inspections. Sequences were mapped to chromosomal bands by using NCBI Map Viewer or UCSC genome browser.

Bidirectional BLAST searches were used to define orthologs across the species primarily. Then each group of bHLH proteins was aligned by CLUSTAL-X (Thompson *et al*., 1997). Phylogenetic studies were performed by using phylogenetic analysis option incorporated in Clustal-X program. Hypertree, a java-based program was used to visualize phylogenetic trees (Bingham and Sudarsanam, 2000). Full length proteins were used to establish the orthology relationships. Percentiles of the identity were determined by bidirectional BLAST searches, followed by manual inspection. Identity per average sequence size was used. Complete domain structures of bHLH proteins were determined by SMART (http://smart.embl-heidelberg.de) and CDD (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) web servers. DnaSP (http://www.ub.edu/dnasp) was used to calculate Ka/Ks ratios.

RESULTS

Comprehensive identification of human, mouse and rat bHLH proteins: Putative bHLH proteins were obtained by searching predicted human, mouse and rat genomic assembly, cDNA, EST and gene models for similarity to known bHLH proteins and bHLH domains. This produced a list of 121 human, 121 of mouse and 120 of rat bHLH proteins. bHLH proteins were classified according to the established hierarchical clustering into six groups and fifty families (Atchley and Fitch, 1997; Simionato *et al*., 2007). This classification was done on the basis of their evolutionary and structural relationships. We were also able to classify human, mouse and rat

bHLH proteins into established six groups (A-F). New proteins which cannot be included into those six groups classified as a new group G. All bHLH proteins were mapped to the chromosomal loci using NCBI Map Viewer or UCSC genome browser. The complete list of human, mouse and rat bHLH proteins can be found at http://sites.google.com/site/bhlhtf/. Each entry includes multiple synonyms, description, Gene ID, chromosomal location and orthology relationships. Gene ID and chromosomal location have been linked to the NCBI databases.

Structure of bHLH proteins: The bHLH proteins function as a diverse set of regulatory factors due to the heterogeneity of DNA sequences recognized and dimers formed. The DNA binding activity is determined by the basic region, while HLH region is recognized as dimerization domain. Other domains within these proteins act as adaptor or link to other proteins. We identified 16 additional domains present in 49 bHLH proteins using SMART and CDD databases (Fig. 1). At least 20 families contain an additional functional domain (Table 1). In general, members of the same bHLH family have the same domain structure, but some domain shuffling is observed, where individual members of families have lost a domain. For example, the PAC domain is found in all four Hif family bHLH proteins as well as in members of the Sim and Clock families, while one of two Ahr family proteins lacks this domain. Within three Orphan family members only MGA processes a TBOX. Two Hif family members HIF3A and NPAS1 lack one HIF-1a CATD domain. All E group members contain an Orange domain while all F group members contain an IPT domain. At least seventy human bHLH proteins contain no additional domains. Some are small proteins containing little more than a bHLH domain while others contain conserved sequences that have not, yet been classified as domains and whose functions are unknown.

Comparisons between human and rat bHLH proteins: Almost all human and rat bHLH proteins are present as orthologous pairs. There are 119 such orthologous pairs. Two bHLH proteins are found only in human. Human lacks one rat bHLH protein (Fig. 2). Orthologs of two human bHLH proteins were absent from rat. Although, rat genome has not been sequenced completely, the absence of the two bHLH proteins is probably not due to incomplete genomic sequence, because they are also absent from EST and cDNA databases and from the mouse genome. One of these

Table 1: Domain structures observed with the changes within families

Groups	Families	Domains
A	MyoD	bHLH, Basic
B	Arnt, bmal	bHLH, PAS, PAC
B	Orphan	bHLH, TBOX
B	Myc	bHLH, MycN, MycLZ
B	Hey/hairy	bHLH, Orange
C	Ahr, clock, hif, trh, trhL	bHLH, PAS, PAC
C	Hif	bHLH, PAS, PAC, HIF-1a CATD
C	Sim	bHLH, PAS, PAC, SIM_C
C	Src	bHLH, PAS, SRC1, NRCA
E	Hairy/E (spl), hey	bHLH, Orange
F	Coe	bHLH, IPT
G	ERC1	bHLH, RBD FIP
G	SMC1A	bHLH, ABC, SMC hinge
G	SYT17	bHLH, C2

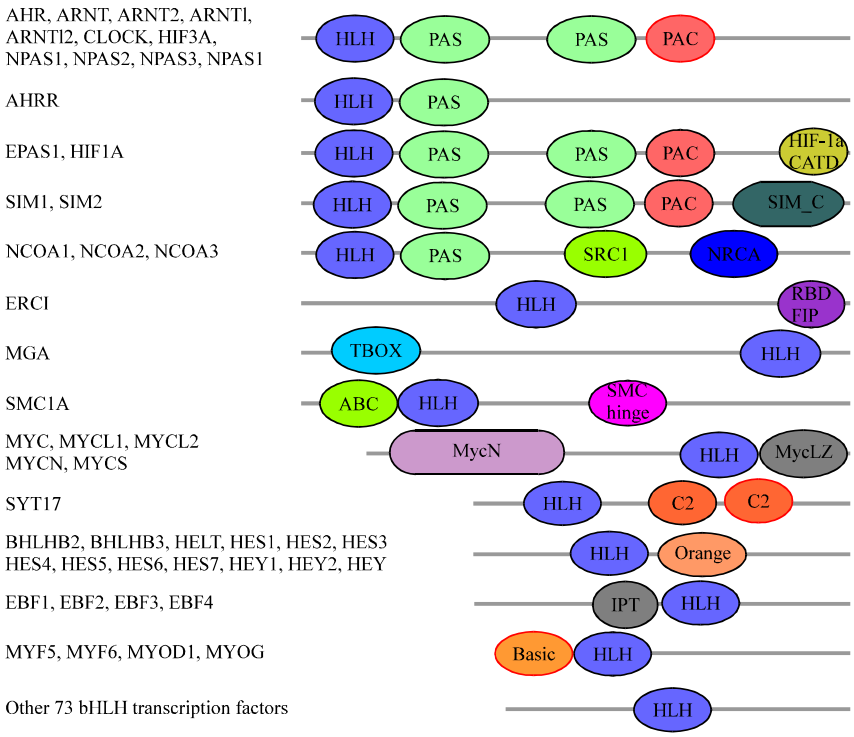
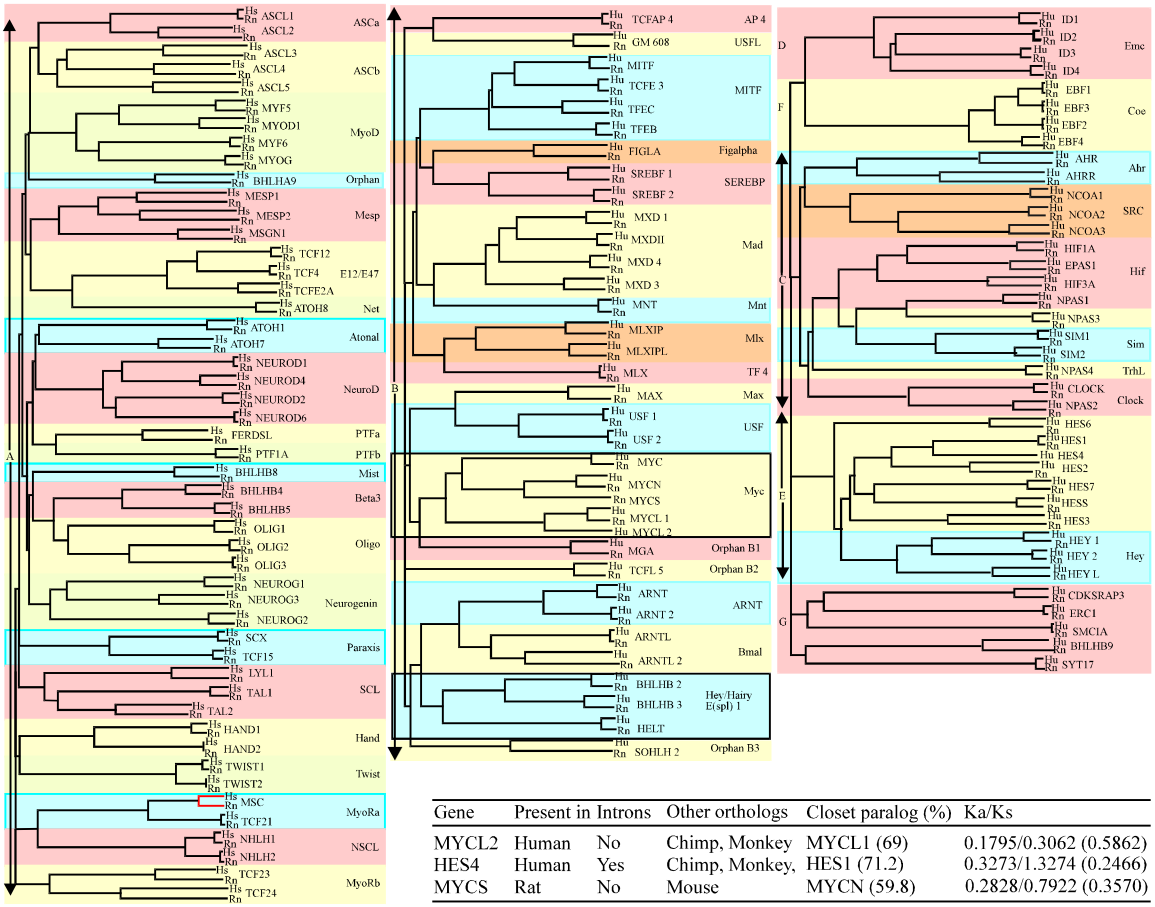


Fig. 1: Domain architectures of bHLH proteins. Human bHLH proteins were used to search for the functional domains in bHLH proteins

proteins, MYCL2 is intronless and has 69% similarity with other bHLH protein MYCL1 and so is probably retrotransposed copy of MYCL1 gene. Other human specific gene, HES4 might be created from the genomic duplication of HES1 gene having 71% similarity. Both of genes are present in chimpanzee and monkey. Rat MYCS is intronless present, only in mouse genome and has 60% similarity with its closest paralog MYCN protein. All of these species specific genes are expressed, and sequence analysis indicates that they continue to be under functional pressures, with a low ratio of nonsynonymous to synonymous substitutions (0.25-0.59) relative to their parental genes.

Functionally important sequences: Alignment of orthologous bHLH protein sequence pairs shows a wide variation in local sequence conservation. The basic structural constraints of the bHLH domain are common across all bHLH proteins, yet there are marked differences in the degree of conservation in different families. Orthologous bHLH domains are on average 97% identical, but some are as low as 66% and 65 pairs are identical across the full domain (Table 2, Fig. 3a). Orthologous full length bHLH proteins are on average 86% identical, but some are as low as 44%, and 28 pairs are more than 95% identical across the full length protein (Fig. 3a). This variability is clearly family-dependent (Fig. 3a). For example, of the four NeuroD family pairs, two are 98% identical and the other two differ by around 10% an average difference of only 94%. While bHLH domain pairs of this family are identical. Collectively, this indicates that massive changes in amino acid destroy some function and have been eliminated by evolution. At the other extreme, Mesp family pairs are 58-77% identical, indicating that the functions of this family of bHLH proteins do not greatly constrain the domain sequence (Fig. 3a). Although, most differences between orthologs



Gene	Present in	Introns	Other orthologs	Closet paralog (%)	Ka/Ks
MYCL2	Human	No	Chimp, Monkey	MYCL1 (69)	0.1795/0.3062 (0.5862)
HES4	Human	Yes	Chimp, Monkey,	HES1 (71.2)	0.3273/1.3274 (0.2466)
MYCS	Rat	No	Mouse	MYCN (59.8)	0.2828/0.7922 (0.3570)

Fig. 2: Phylogenetic analysis of human and rat bHLH proteins. Human and rat bHLH proteins were aligned using CLUSTA-X by dividing into three groups. Un-rooted trees were constructed using phylognetic option incorporated with CLUSTAL-X and visualized by hypertree. Hu, Human and Rn, Rat. Lower panel of the right side describes the nonsynonymous (Ka) to synonymous (Ks) substitution ratios of the three pairs of genes which were calculated by DnaSP

Table 2: Functionally important sequences

Parameters	Observation
Orthologous bHLH domains	97% identical
Lowest identity in orthologous bHLH domains	66%
Identical orthologous bHLH domains	65 pairs
Orthologous full length bHLH proteins	86% identical
Lowest identity in orthologous full length bHLH proteins	44%
More than 95% identical orthologous full length bHLH proteins	28 pairs
Indels of six or more amino acids	30
Nobel insertions in human	18
Nobel insertions in rat	6
Nobel insertions in both	6

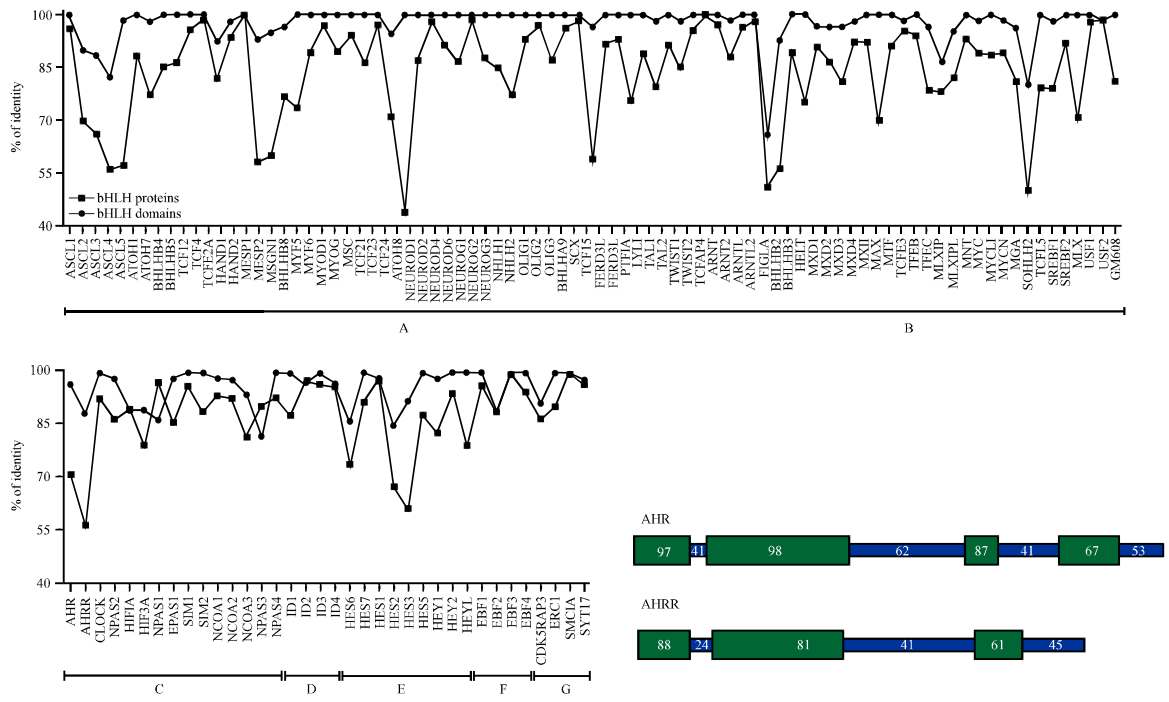


Fig. 3(a-b): Conservation of orthologous bHLH proteins (a) Percentile of identity of orthologous bHLH proteins (square) and bHLH domains (dot) are shown. Conservation within orthologous bHLH domains is family-dependent and (b) Schematic of Ahr protein sequences, with the highly conserved between human and rat boxed in green. Percentage ortholog identity is given for each block and interblock region of lesser conservation

are due to amino acid substitutions, many proteins contain substantial inserts or deletions (indels) between orthologs, which may account for much of their functional differences between species. Thirty of the 119 ortholog pairs (25%) contain indels of six or more amino acids in which 18 have novel insertions in human, 6 in rat and 6 in both. Highly conserved regions map to known domains or reveal previously unknown conserved regions of likely functional importance. For example, the two Ahr proteins have little sequence similarity (26%) outside their bHLH domain (80%). Pairwise alignment of human and rat AHR and AHRR identifies several previously undescribed highly conserved domains, separated by poorly conserved sequences, occurring in similar regions within both ortholog pairs (Fig. 3b).

DISCUSSION

We identified the putative complete set of bHLH proteins encoded by human, mouse and rat genomes. This study describes a catalog of 121 human, 121 mouse and 120 rat bHLH proteins. A previous report gives an alternative count of 118 human bHLH proteins (Simionato *et al*., 2007). This shares 111 genes with our catalog and also includes one unknown gene and six duplicated genes. Another report described 124 bHLH proteins within the mouse genome (Li *et al*., 2006a), in which 117 genes overlap with our prediction. Moreover, we were unable to follow 4 genes and 3 genes might be counted twice. A detailed comparison can be found at http://sites. google. com/site/ bhlhtf/comp.

The bHLH proteins show diversity in functional domain architectures. Presence of 16 functional domain architectures further indicates their diversity in function. Domain architecture of bHLH proteins is family specific. Some domain shuffling is observed, where individual members of families have lost a domain. Members of nine families contain at least one PAS domain. PAS domains have protein-binding and dimerisation functions which are found in a large number of organisms from bacteria to humans (Ponting and Aravind, 1997). PAC domains are found carboxy-terminal to several PAS sequences which are likely to contribute to the PAS structural domain. Although, almost all PAS domain containing bHLH proteins process a PAC domain which is followed by the PAS domains, only Src family members and AHRR have lost that domain. AHRR protein contains a long portion of still unclassified region that might have some functional importance. Thirteen bHLH proteins contain an extra Orange domain which is involved in protein-protein interactions (Leimeister *et al*., 2000). MYC and MAX family proteins contain an extra Leucine Zipper domain along with bHLH domain.

At least 119 rat bHLH proteins have human orthologs while few bHLH genes exhibit lineage specificity. Most of new bHLH proteins within each lineage are derived from retrotransposition rather than genomic duplication. Two human genes are lost from rat genomes. MYCL2 probably a retrotransposed of MYCL1 gene which is present in the common ancestor of human but lost from the mouse lineage. HES4 has been seen only in human, chimp and monkey, but its degree of divergence from HES1 indicates that the duplication that created these genes happened early in vertebrate evolution, and that one copy was later lost. MYCS, an intronless Myc gene family member can only be found in mouse and rat. A quick search with NCBI protein database could not identify MYCS orthologs within the other vertebrate genomes suggesting the rodent specific function of this gene.

Many proteins contain conserved unknown regions. bHLH proteins are highly conserved with in human and rat. On an average 86% amino acids are conserved within this two species and 97% amino acids are identical in bHLH domain indicating strong functional pressure throughout the functional domains. In conclusion, this study provides a catalog and overviews of bHLH proteins including structural insights.

ACKNOWLEDGMENT

This study was supported by a KN Biomedical Research Grant. We are grateful to Nadid Ahnaf Kazi.

REFERENCES

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

Bingham, J. and S. Sudarsanam, 2000. Visualizing large hierarchical clusters in hyperbolic space. Bioinformatics, 16: 660-661.

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

Leimeister, C., K. Dale, A. Fischer, B. Klamt and M.H. de Angelis *et al*., 2000. Oscillating expression of c-Hey2 in the presomitic mesoderm suggests that the segmentation clock may use combinatorial signaling through multiple interacting bHLH factors. Dev. Biol., 227: 91-103.

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, 2006b. Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and arabidopsis. Plant Physiol., 141: 1167-1184.

Ponting, C.P. and L. Aravind, 1997. PAS: A multifunctional domain family comes to light. Curr. Biol., 7: R674-R677.

Satou, Y., K.S. Imai, M. Levine, Y. Kohara, D. Rokhsar and N. Satoh, 2003. A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. I. Genes for bHLH transcription factors. Dev. Genes Evol., 213: 213-221.

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

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.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougi 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.

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.