

Evolution, functional divergence and conserved exon–intron structure of bHLH/PAS gene family

Jun Yan · Zhaowu Ma · Xiaopeng Xu · An-Yuan Guo

Received: 11 July 2013 / Revised: 9 September 2013 / Accepted: 15 October 2013 / Published online: 8 November 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract bHLH/PAS genes encode a family of basic helix-loop-helix (bHLH) transcription factors with bHLH, PAS and PAS_3 domain. bHLH/PAS genes are involved in many essential physiological and developmental processes, such as hypoxic response neural development, the circadian clock, and learning ability. Despite their important functions, the origin and evolution of this bHLH/PAS gene family has yet to be elucidated. In this study, we aim to explore the origin, evolution, gene structure conservation of this gene family and provide a model to analyze the evolution of other gene families. Our results show that genes of the bHLH/PAS family only exist in metazoans. They may have originated from the common ancestor of metazoans and expanded into vertebrates. We identified bHLH/PAS genes in more than ten species representing the main lineages and constructed the phylogenetic trees (Bayesian, ML and NJ) to classify them into three groups. The exon–intron structure analysis revealed that a relatively conserved “1001-0210” eight-exon structure exists in most groups

and lineages. In addition, we found the exon fusion pattern in several groups in this conserved eight-exon structure. Further analysis indicated that bHLH/PAS protein paralogs evolved from several gene duplication events followed by functional divergence and purifying selection. We presented a phylogenetic model to describe the evolutionary history of the exon structures of bHLH/PAS genes. Taken together, our study revealed the evolutionary model, functional divergence and gene structure conservation of bHLH/PAS genes. These findings provide clues for the functional and evolutionary mechanism of bHLH/PAS genes.

Keywords bHLH/PAS · Evolution · Conserved exon–intron structure · Functional divergence · Purifying selection

Introduction

The Per-ARNT-Sim (PAS) transcription factors (TFs) which belong to the basic helix-loop-helix (bHLH) TFs are critical regulators of the gene expression network involved in many essential physiological and developmental processes (Kewley et al. 2004). The bHLH/PAS genes were thought to include the aryl hydrocarbon receptor (AHR), hypoxia inducible factors (HIF- α), single minded proteins (SIM), aryl hydrocarbon receptor nuclear translocator (ARNT), aryl hydrocarbon receptor nuclear translocator-like (ARNTL, also called BMAL1 and BMAL2), neuronal PAS domain protein (NPAS) (Gilles-Gonzalez and Gonzalez 2004; Kewley et al. 2004). The bHLH/PAS proteins contain a bHLH domain and PAS domain dimmers labeled as PAS and PAS_3 in the Pfam database (Gilles-Gonzalez and Gonzalez 2004; Kewley et al. 2004). Although the widespread distribution and importance in physiological

Communicated by S. Xu.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-013-0786-0) contains supplementary material, which is available to authorized users.

J. Yan · Z. Ma · X. Xu · A.-Y. Guo (✉)
Hubei Bioinformatics and Molecular Imaging Key Laboratory,
Department of Biomedical Engineering, College of Life
Science and Technology, Huazhong University of Science
and Technology, Wuhan 430074, Hubei,
People's Republic of China
e-mail: guoay@mail.hust.edu.cn

J. Yan
Department of Applied Physics, College of Information
Science and Engineering, Shandong Agricultural University,
Taian 271018, Shandong, People's Republic of China

and developmental processes of bHLH/PAS genes, the process of domain fusion and expansion of this gene family are unknown. We aimed to explore the evolution, origin and gene structure conservation of this gene family, as this may also provide new insights into other gene family evolutions.

bHLH/PAS proteins have multiple functions in many essential physiological and developmental processes, which include dioxin toxicology, hypoxic response, neural development, circadian clock, schizophrenia and learning ability. AHR, as the receptor of Dioxin is well studied in toxicology using zebrafish and mice (Hahn et al. 2006) and also plays essential roles in the immune system (Esser et al. 2009; Stevens et al. 2009). Responses to hypoxia are mediated by HIF-1-3 α , and the oxygen-dependent hydroxylases are inactivated resulting in the accumulation of HIF- α and the formation of a dimer with ARNT under hypoxia (Kewley et al. 2004). ARNT is a general partnering factor of bHLH/PAS proteins such as the AHR, HIF- α , SIM and itself (Kewley et al. 2004). The mechanism of the circadian clock is that the CLOCK and BMAL1 (ARNTL) form a heterodimer to bind to the E-box response elements in the promoter region and stimulate the transcription of other essential clock genes (Griffin et al. 1999; Kume et al. 1999; Preitner et al. 2002). NPAS1/3/4 are associated with neuronal function, social recognition, learning ability (Erbel-Sieler et al. 2004; Pickard et al. 2005; Ploski et al. 2011) and NPAS2 likely functions as part of a molecular clock operative in the mammalian forebrain (Reick et al. 2001).

Gene duplications provide a redundant gene copy for functional innovation (Conant and Wolfe 2008). In 2009, Hahn et al. reviewed the models for the three fates of duplicated genes which are neofunctionalization, subfunctionalization and conservation of function (Hahn 2009). Conant and Wolfe stated that a duplicate gene pair undergoes three life stages to form new functions which are creation, fixation-preservation and subsequent optimization (Conant and Wolfe 2008). Here, we are interested in the specific role of natural selection in the evolutionary history of the bHLH/PAS family. Previous evolutionary studies of bHLH/PAS genes focused on some subfamilies, such as the HIF- α (Zhang et al. 2010), PHD-HIF oxygen-sensing system (Rytkonen et al. 2011), SIM in arthropods (Linne et al. 2012), the animal circadian clock in sea anemone (Reitzel et al. 2010). Yet, there is a gap with regards to a comprehensive study about the molecular evolution of the bHLH/PAS gene family.

In this study, we performed an extensive phylogenetic analysis for the bHLH/PAS gene family. Based on the analysis of exon-intron structure and phylogenetic trees, we proposed a model of bHLH/PAS origination and evolution, including reservation or change of a conserved “1001-0210” eight-exon structure. Analyses of selective pressure and functional divergence were combined to explain the

relationship of the site-specific evolution and functional divergence. Our studies will help in understanding the bHLH/PAS evolution mechanism and provide a starting point for further experimental investigations.

Materials and methods

Identification of bHLH/PAS genes

The proteome sequences of the represented species were obtained from predicted bHLH/PAS genes as per our previous study (Zhou et al. 2012). The proteome of *Dictyostelium discoideum* was downloaded from dictybase (<http://www.dictybase.org>). The proteome of *Strongylocentrotus purpuratus* was downloaded from NCBI (<http://www.ncbi.nlm.nih.gov>). The proteomes of *Monosiga brevicollis*, *Amphimedon queenslandica*, *Nematostella vectensis* and *Branchiostoma floridae* were obtained from JGI (<http://genome.jgi-psf.org/>). The other investigated proteomes were downloaded from Ensembl (<http://www.ensembl.org>).

HMMER in our local server and Pfam 26.0 in batch (<http://pfam.sanger.ac.uk/>) with an *E*-value below 1.0 were performed against all proteomes. The Pfam profile PF00010 (bHLH domain), PF08447 (PAS_3 domain) and PF00989 (PAS domain) were used in HMMER. The CD-HIT (Huang et al. 2010) 90 % sequence identity with a default parameter was applied to remove the polymorphisms, splice variants, pseudogenes and duplicates in JGI and NCBI data. To remove the splice variants from Ensembl, the longest one was kept among the proteins with the same gene id by our perl script.

Multiple alignment and phylogenetic analysis

For the phylogenetic analysis, the truncated sequences including bHLH, PAS and PAS_3 domains and their intervals were used to perform the multiple alignments (Fig. S1) by ClustalX (v2.0) (Larkin et al. 2007). The phylogenetic trees were constructed by three different approaches: Bayesian analyses, neighbor-joining (NJ) method with the JTT model, Maximum Likelihood (ML). Bayesian analysis was performed using MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003) with the mixed amino acid substitution model; a MCMC chain with 12,000,000 generations was used. Markov chains were sampled every 100 generations, and the first 25 % of the trees were discarded as burn-in. Convergence was assessed by checking the average standard deviation of split frequencies (below 0.01). MEGA5 (Tamura et al. 2007) software was used to construct the NJ trees based on the multiple alignments. To assess the confidence degree of nodes in the NJ phylogenetic trees, bootstrap with 1,000 repetitions was performed. The ML

trees were constructed using the PhyML3.0 (Guindon and Gascuel 2003) with 100 bootstrap replicas. The appropriate models of the ML methods including model parameters were calculated by ProtTest2.4 (Abascal et al. 2005) using Akaike Information Criterion (AIC), and the result was JTT + I + G + F.

Identification of exon–intron structure and domain search

Exon–intron structures were based on NCBI (*S. purpuratus*), JGI (*N. vectensis*) and UCSC (*S. purpuratus*, <http://genome.ucsc.edu/>) and Ensembl (others). The information of bHLH, PAS and PAS_3 domains was from Pfam 26.0 in batch (<http://pfam.sanger.ac.uk/>), and was checked and edited manually. The diagram of exon–intron structures was generated using the perl and R script based on the extracted information, followed by manual edition.

Analysis of functional divergence

Type I functional divergence analysis was performed by DIVERGE (version 2.0) software (Gu and Vander Velden 2002). Type I functional divergence which resulted in altered functional constraints between duplicated genes, lead to one of the genes being conserved and the other gene showing great variation. The coefficient of functional divergence (θ) is an indicator of the level of type I functional divergence among two homologous gene clusters. The sites (k) with contribution to the functional divergence were predicted according to their posterior probabilities (Q_k). The sites with $Q_k > 0.67$ were meaningful for the functional divergence. We chose 75 sequences (including bHLH, PAS and PAS_3 domains and their intervals) of deuterostome (including sea urchin, ciona, zebrafish, frog, anole, chicken and human) to calculate the coefficients of functional divergence.

Analysis of selective pressure

The truncated amino acid sequences of bHLH, PAS and PAS_3 domains and their intervals were aligned by ClustalX (v2.0) (Larkin et al. 2007). The corresponding cDNA data was downloaded from Ensembl (<http://www.ensembl.org>). Perl script and local BLAT software were used to obtain the corresponding truncated cDNA. The codon alignment of cDNA was generated by web server PAL2NAL (Suyama et al. 2006). Then CODEML in PAML4.5 (Yang 2007) was used to test the selective pressure with the model M0 (one ratio), M1a (near neutral), M2a (positive selection), M3 (discrete), M7 (beta), M8 (beta and ω). The non-synonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) was calculated to indicate the selective pressure. $\omega > 1$ indicates positive selection, $\omega < 1$ indicates negative selection, and $\omega \approx 1$

indicates neutral evolution. The ML method was used to explore the selective pressure, and the likelihood ratio test (LRT), $2\Delta\ln L = 2(\ln L_1 - \ln L_0)$ obeying the χ^2 distribution with the degrees of freedom (df).

Results

Genome wide identification revealed that bHLH/PAS genes existed in metazoa

To study the origination and evolution of bHLH/PAS genes, we searched bHLH/PAS genes in all species investigated by the definition of containing all three domains: bHLH, PAS, and PAS_3 (see “Materials and methods”, Table 1, Table S1). The results showed that this family existed in metazoa since we identified four sequences with the bHLH, PAS and PAS_3 domains in *A. queenslandica*. Moreover, two sequences with bHLH and PAS domains (jgilMonbr112346l and jgilMonbr1126507l), and the similar sequences with bHLH and PAS_3 domains (jgilMonbr1137483l and jgilMonbr117924l) were found in *M. brevicollis*, suggesting that these three domains may fuse step by step to form the bHLH/PAS genes. Since we are interested in the bHLH/PAS genes with three domains, although we also found some sequences with only two domains in different species, we excluded them in the following analyses. No sequence with these three or two domains was found in *Saccharomyces cerevisiae*, *D. discoideum*, plants, fungi and microbes by BLAST search at NCBI. In invertebrate genomes, there are usually <10 bHLH/PAS genes, and vertebrates always contain more than 10 bHLH/PAS genes (Table 1). In addition, we summarized the detailed gene information of bHLH/PAS in human and zebrafish (Table S2–3).

Phylogenetic analysis of the bHLH/PAS genes

To infer the evolution and classification of bHLH/PAS genes, a phylogenetic Bayesian tree was constructed based on the three domains (bHLH, PAS and PAS_3) and their intervals (Fig. 1). The subtrees of the Bayesian tree are shown in supplementary Fig. S2. The topologies of ML and NJ trees are similar to the Bayesian tree (Fig. S3). Based on the above results, we classified the bHLH/PAS genes into three groups: the first group is AHR, the second group is SIM1/2, HIF-1-3 α and NPAS1/3, the third group is NPAS2/CLOCK, ARNT/ARNT2 and ARNTL/ARNTL2.

The first group is AHR, which existed from *Caenorhabditis elegans* to *Homo sapiens*. There are two clusters which are the deuterostome cluster and the protostome cluster in the AHR clade. We found that there is only one copy of the AHR gene in most vertebrates except zebrafish (three

Table 1 List of bHLH/PAS genes in the genomes investigated

Lineage	Species	Common name	bHLH + PAS + PAS_3	bHLH + PAS	bHLH + PAS_3
Protista	<i>Monosiga brevicollis</i>	Choanoflagellates	0	2	2
Spongia	<i>Amphimedon queenslandica</i>	Sponge	4	0	0
Cnidaria	<i>Nematostella vectensis</i>	Sea anemone	6	3	0
Nematode	<i>Caenorhabditis elegans</i>	Soil nematode	2	2	0
Insecta	<i>Drosophila melanogaster</i>	Fruit fly	7	0	2
Echinodermata	<i>Strongylocentrotus purpuratus</i>	Sea urchin	8	1	0
Urochordata	<i>Ciona intestinalis</i>	Ciona	3	0	1
Cephalochordata	<i>Branchiostoma floridae</i>	Amphioxus	7	1	2
Cyclostomata	<i>Petromyzon marinus</i>	Lamprey	5	1	1
Osteichthyes	<i>Danio rerio</i>	Zebrafish	18	8	1
Amphibia	<i>Xenopus tropicalis</i>	Frog	11	3	1
Reptilia	<i>Anolis carolinensis</i>	Anole	11	4	2
Aves	<i>Gallus gallus</i>	Chicken	14	3	0
Monotremata	<i>Ornithorhynchus anatinus</i>	Platypus	6	3	0
Marsupialia	<i>Monodelphis domestica</i>	Opossum	14	2	1
Rodentia	<i>Mus musculus</i>	Mouse	14	4	1
Primates	<i>Homo sapiens</i>	Human	14	4	1

copies), chicken (three copies) and opossum (three copies). The Lamprey AHR (ENSPMAP0000006554) was also found. This AHR group is clustered with three sequences (ENSDARP00000105317 and ENSDARP00000083340 in zebrafish which were identified as “neuronal PAS domain-containing protein 4-like” by NCBI, C15C8.2b in *C. elegans*) with high common Bayesian posterior probability (91) and high bootstrap value in NJ and ML trees (Fig. S3).

The second group SIM1/2, HIF-1-3 α and NPAS1/3, appeared to be the common ancestor of metazoans because we identified two sponge sequences of *A. queenslandica* (Aqu1.222845 and Aqu1.222846) in this group. The NPAS1/3 clade contained a sequence from *N. vectensis* (jgilNemve1196877), implying that this clade emerged early in the common ancestor of eumetazoan. NPAS1 and NPAS3 diverged in vertebrates, suggesting that they duplicated during the origin of vertebrates since there was only one copy in ciona (ENSCINP00000001875). Though two copies of NPAS1/3 (ENSPMAP00000010632 and ENSPMAP00000008398) were identified in lamprey, they seem to be from the tandem duplication. This is similar in the SIM1/2 clade that diverged in vertebrates, while one sequence in amphioxus (jgilBraf112650331) and *S. purpuratus* (gil115920139) is located in this clade. The fly SIM (FBpp0082178) indicated that the SIM gene emerged at least before the deuterostome and the protostome split. Two copies of SIM1 (ENSDARP00000033085 and ENSDARP0000003984) were found in zebrafish, which may be the result of 3R WGD (whole genome

duplication) in Teleost (Taylor et al. 2001; Vandepoele et al. 2004; Meyer and Van de Peer 2005). One lamprey HIF- α (ENSPMAP0000000148) and one amphioxus HIF- α (jgilBraf112084081) were found. HIF- α was also found in sea anemone (jgilNemve11619591), implying that HIF- α occurred early before bilateral animals emerged. The presence of three members of HIF- α in most vertebrates (HIF-1-3 α) suggests that HIF- α duplicated at the origin of vertebrates, and maybe as a result of 2R WGD.

The third group contains two big clades; the first clade is NPAS2/CLOCK and the second clade is ARNT/ARNT2 and ARNTL/ARNTL2. The NPAS2/CLOCK clade was found in *A. queenslandica* (Aqu1.220834). There was only one copy of genes in this clade in *N. vectensis* (jgilNemve1160110), *Drosophila melanogaster* (FBpp0099478) and *B. floridae* (jgilBraf11636361), while the genes in this clade diverged into NPAS2 and CLOCK in vertebrates. We noticed that a clade containing five sequences (ENSDARP00000016938 in zebrafish, ENSX-ETP00000005089 in frog, ENSGALP00000014735 in chicken, ENSOANP00000018438 in platypus and ENSMODP00000007175 in opossum) is located in the NPAS2 clade of the Bayesian, ML and NJ trees (Fig. 1, Fig. S3). It is surprising that we didn't find genes of this clade in other mammals, which means they may have been lost during evolution. There are two CLOCK genes (ENSDARP00000014377 and ENSDARP00000120780) in zebrafish, as compared to one CLOCK gene in other vertebrates. The clade ARNT/ARNT2 and ARNTL/ARNTL2 also

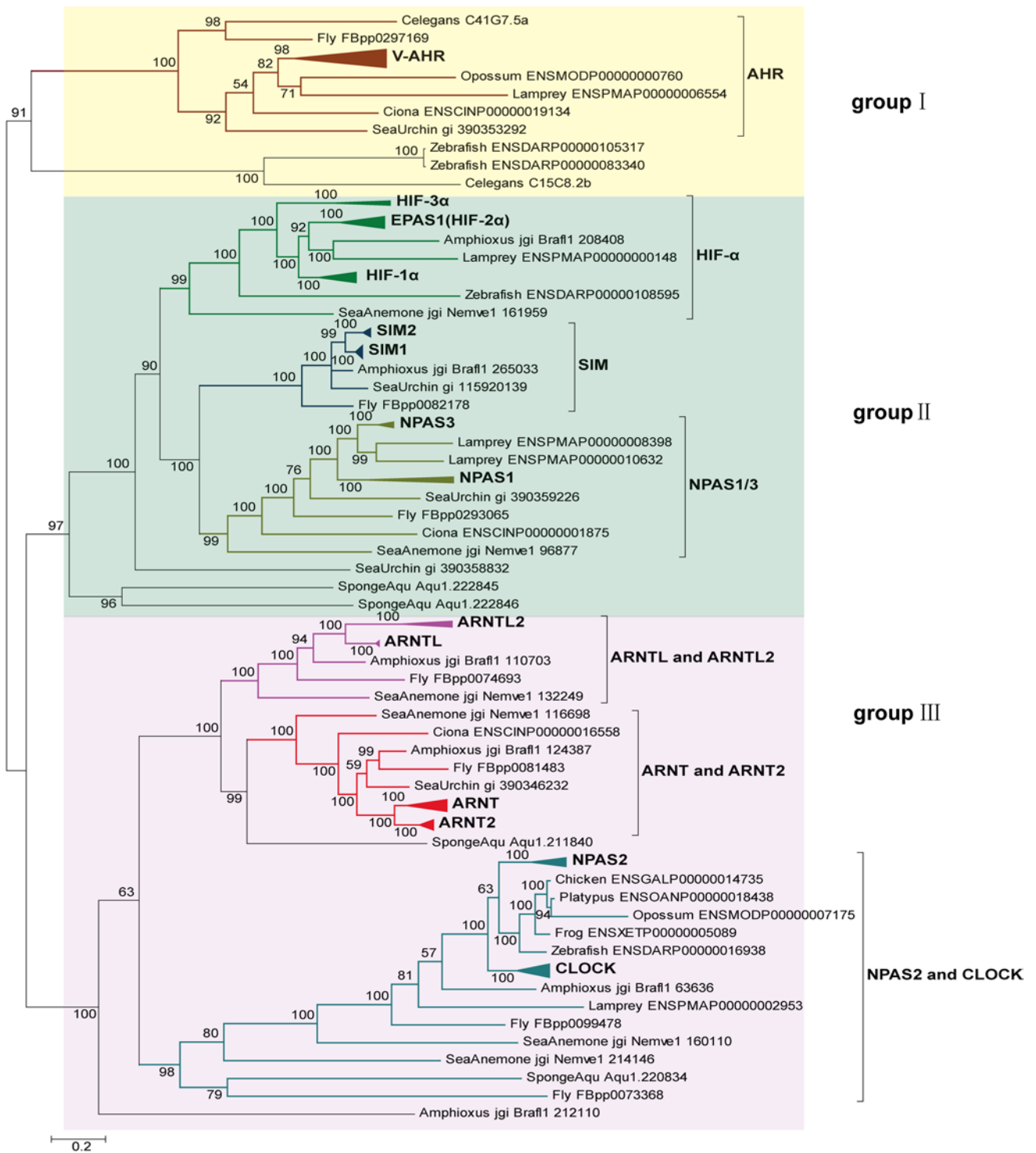


Fig. 1 Phylogenetic tree and gene structure of bHLH/PAS genes. The Bayesian tree was built for the sequences of three domains (bHLH, PAS and PAS_3) and their intervals using the MrBayes 3.1.2

with the mixed amino acid substitution model. The vertebrate bHLH/PAS branches were compressed and their expanded subtrees were in Fig. S2. Three groups (I–III) were labeled in our diagram

contained two big subclades and their common Bayesian posterior probability was very high (100). The subclade of ARNT/ARNT2 appeared as early as the period of the sponge

(Aqu1.211840), and contained only one copy in invertebrates which include *N. vectensis* (jgilNemve1|116698|), *D. melanogaster* (FBpp0081483), *S. purpuratus* (gil390346232|),

Ciona intestinalis (ENSCINP00000016558) and *B. floridae* (jgilBraft1124387). ARNT diverged into ARNT and ARNT2 in vertebrates. This is similar in the ARNTL subclade, which also contained two copies in vertebrates. ARNTL emerged early in *N. vectensis* (jgilNemve1132249), and contained only one copy in invertebrates.

A conserved exon phase pattern “1001-0210” existed in bHLH/PAS genes

To study the gene structures of bHLH/PAS genes, we painted the exon–intron structures in some species based on the information of Ensembl and JGI (Fig. 2, Fig. S4, Fig. S5). Our results showed that most deuterostomes (e.g. gil3903532921 in *S. purpuratus*, ENSCINP00000019134 in *C. intestinalis*) contained a conserved exon–intron structure within the exon phases “1001-0212” in the AHR clade. The AHR genes in flies (FBpp0297169) and nematodes (C41G7.5a) lost this exon–intron structure, which may be a result of the gene rearrangement in protostome (Putnam et al. 2007). There are the similar exon–intron structures in other clades. For example, the exon–intron structure of the amphioxus gene (jgilNemve196877) is “1001-0210” in the NPAS1/3 clade, while the exon–intron structure of

deuterostome genes in this clade evolved into “1001-0212”. The little change in the last exon of “1001-0210” is caused by a prolongation in the C terminal. The exon–intron structure of amphioxus SIM gene (jgilBraft112650331) is still “1001-0210”, however, it became “1001-0212” in vertebrate SIM1/2 (see zebrafish and human SIM1/2). The exon phases of HIF- α in sea anemone and amphioxus were “1001-0210”, while it became “101-0210” in vertebrates as a result of an exon fusion in the second and third exon because the amino acids sequences of this exon is similar with the original two exons (Fig. 2a, Fig. S4a).

In NPAS2 and CLOCK clade, the exon–intron structure changed into “1001-10212” as compared to “1001-0212” in vertebrates NPAS1/3 and SIM1/2. This is not caused by an exon insertion but the rearrangement of the fifth and the sixth exons in “1001-0210” (Fig. S5). We noticed that the exon–intron structure of the amphioxus gene (jgilBraft1163636) in this clade is “1001-1212”, as a result of the same rearrangement (Fig. S5). The exon–intron structure of the sea anemone gene (jgilNemve1214146) in this clade has become “1001-121112”, which is caused by the duplicated two PAS domains in this gene. The other sea anemone gene (jgilNemve11160110) in this clade has even changed into one exon (Fig. S4b).

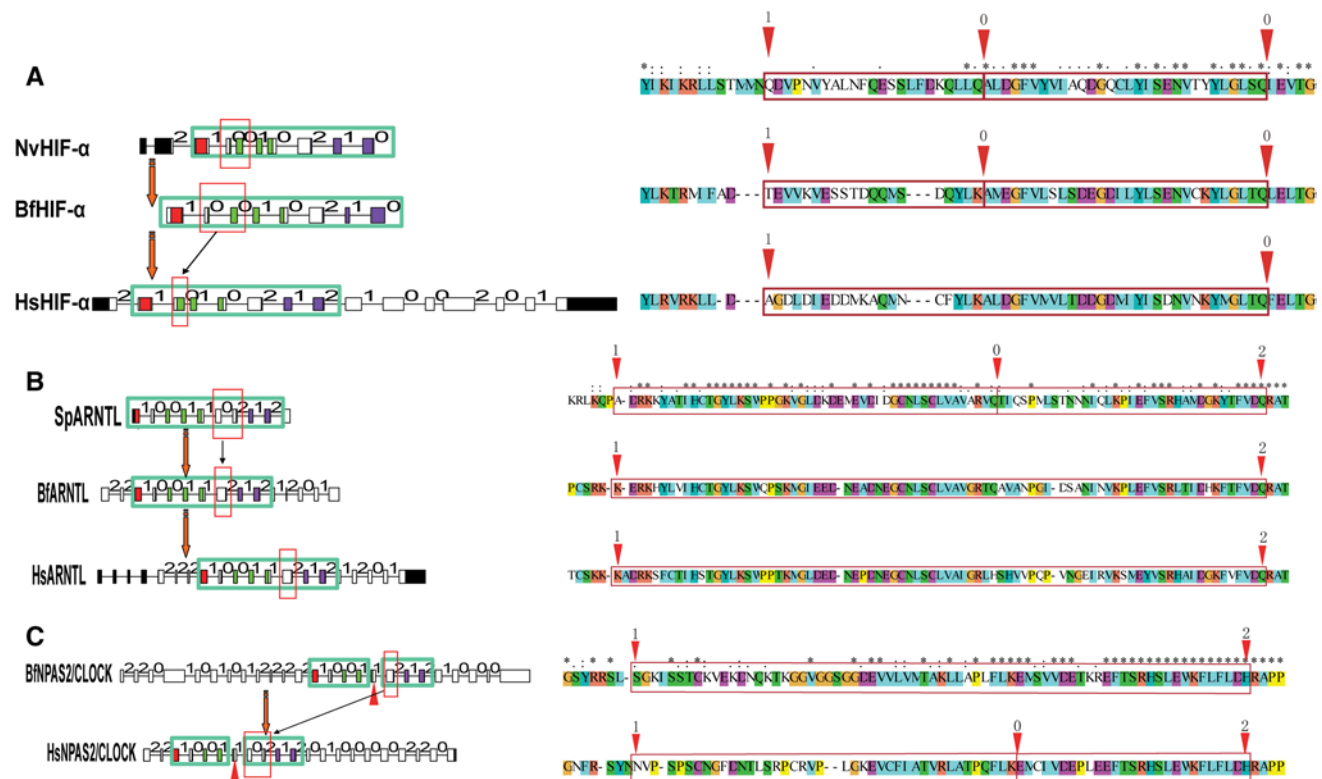


Fig. 2 The detailed exon split and fusion in HIF-1-3 α , ARNTL1/2, NPAS2/CLOCK clades. The one or two exons circled by red boxes and arrows indicated the exon split and fusion, while the corre-

sponding alignments of the sequences are in the right. **a** Exon fusion in HIF-1-3 α . **b** Exon fusion in ARNTL1/2. **c** Exon split in NPAS2/CLOCK (color figure online)

The change in ARNT/ARNT2 and ARNTL/ARNTL2 clade was huge, which is mostly in ARNT/ARNT2 subtree (Fig. S4b). In the ARNTL/ARNTL2 subtree, the exon–intron structure is “1001-1212” in amphioxus (jgilBrafl1110703l), human and zebrafish, suggesting that it is conserved in evolution. The fly ARNTL gene (FBpp0074693) has become four exons with exon phase “2212”. Similar to the exon fusion of HIF- α , we also found that the exon fusion may have occurred in the origins of chordates. The “1001-10212” in sea urchin ARNTL (gil390370839l) became “1001-1212” in amphioxus ARNTL (jgilBrafl1110703l) (Fig. 2b). An exon split was also found in this interval between PAS and PAS_3 domains in NPAS2/CLOCK (Fig. 2c). In the ARNT/ARNT2 subtree, the change of the exon–intron structure occurred during evolution. The exon–intron structure of the sea anemone gene (jgilNemve1116698l) is “1001-10210”. It became “1001-11212” in sea urchin (gil390346232l), and then “1001-111212” in amphioxus (jgilBrafl11124387l). It dramatically changed into “0122-10002” in vertebrates (zebrafish and human ARNT/ARNT2).

Based on the analysis of exon–intron structures, we found a conserved exon phase “1001-0210” which exists in some primitive species including the sea anemone (jgilNemve1196877l) and amphioxus (jgilBrafl11208408l), which may be the ancestral exon phase in the bHLH/PAS family. It was retained or showed little change in most clades except in the ARNT/ARNT2 subtree.

Analysis of functional divergence

Gene family proliferation which resulted from genome wide or local chromosome duplication events provides the raw materials for functional innovations. After gene duplication, one gene copy maintains the original function, while the other copy is free to change toward functional divergence for the accumulating amino acid changes. Type I functional divergence represents the amino acid configurations that one gene duplication copy was conserved but the other copy is highly variable (Gu 1999, 2001) to determine the shift selective constraints in the bHLH/PAS gene family, the coefficients of functional divergence (θ) were calculated based on the pairwise sequence comparisons (Table 2). The results show that all θ values were significantly >0 with $p < 0.05$ (LRT, $df = 1$, 3.841 at 5 % for χ^2), suggesting that a site-specific rate shift after gene duplication is a common phenomenon in the evolution of the bHLH/PAS gene family. Further analysis focused on some groups (Fig. 3). ARNT/ARNT2 and ARNTL/ARNTL2 are grouped in one clade while they have more sites with posterior probabilities (Q_k) that are more than 0.67, implying that they underwent the functional divergence after gene duplication (Fig. 3a). NPAS1/3 and SIM1/2 are also in the neighboring groups but their sites with $Q_k > 0.67$ were less than the ARNT/ARNT2 and ARNTL/ARNTL2 group (Fig. 3b). The θ between HIF-1-3 α and NPAS1/NPAS3 is 0.4752, lower than others, and the sites with $Q_k > 0.67$ are

Table 2 Estimates of the coefficient of type I functional divergence (θ)

Cluster 1	Cluster 2	$\theta \pm SE (ML)$	LRT	P value
AHR	NPAS2 and CLOCK	0.6296 \pm 0.0796	62.5246	<0.01
AHR	ARNTL and ARNTL2	0.7808 \pm 0.1071	53.0693	<0.01
AHR	ARNT and ARNT2	0.7784 \pm 0.0711	119.9351	<0.01
AHR	HIF- α	0.6912 \pm 0.0658	110.3452	<0.01
AHR	NPAS1 and NPAS3	0.5744 \pm 0.0635	81.7291	<0.01
AHR	SIM	0.5952 \pm 0.1126	27.9308	<0.01
NPAS2 and CLOCK	ARNTL and ARNTL2	0.7488 \pm 0.1244	36.2426	<0.01
NPAS2 and CLOCK	ARNT and ARNT2	0.8408 \pm 0.0938	80.2998	<0.01
NPAS2 and CLOCK	HIF- α	0.6344 \pm 0.0878	52.1895	<0.01
NPAS2 and CLOCK	NPAS1 and NPAS3	0.8056 \pm 0.0872	85.2795	<0.01
NPAS2 and CLOCK	SIM	0.8648 \pm 0.1201	51.8311	<0.01
ARNTL and ARNTL2	ARNT and ARNT2	0.6088 \pm 0.1136	28.7163	<0.01
ARNTL and ARNTL2	HIF- α	0.7928 \pm 0.1168	46.03871	<0.01
ARNTL and ARNTL2	NPAS1 and NPAS3	0.7056 \pm 0.1100	41.1802	<0.01
ARNTL and ARNTL2	SIM	0.8834 \pm 0.1454	36.9122	<0.01
ARNT and ARNT2	HIF- α	0.7824 \pm 0.0769	103.5715	<0.01
ARNT and ARNT2	NPAS1 and NPAS3	0.3928 \pm 0.0793	24.5655	<0.01
ARNT and ARNT2	SIM	0.7456 \pm 0.1209	38.0087	<0.01
HIF- α	NPAS1 and NPAS3	0.4752 \pm 0.0669	50.4990	<0.01
HIF- α	SIM	0.4432 \pm 0.1235	12.8800	<0.01
NPAS1 and NPAS3	SIM	0.2352 \pm 0.1175	4.0083	<0.05

θ coefficient of type I functional divergence, $SE (ML)$ standard error of the estimated theta, LRT 2 log-likelihood-ratio against the null hypothesis of $\theta = 0$

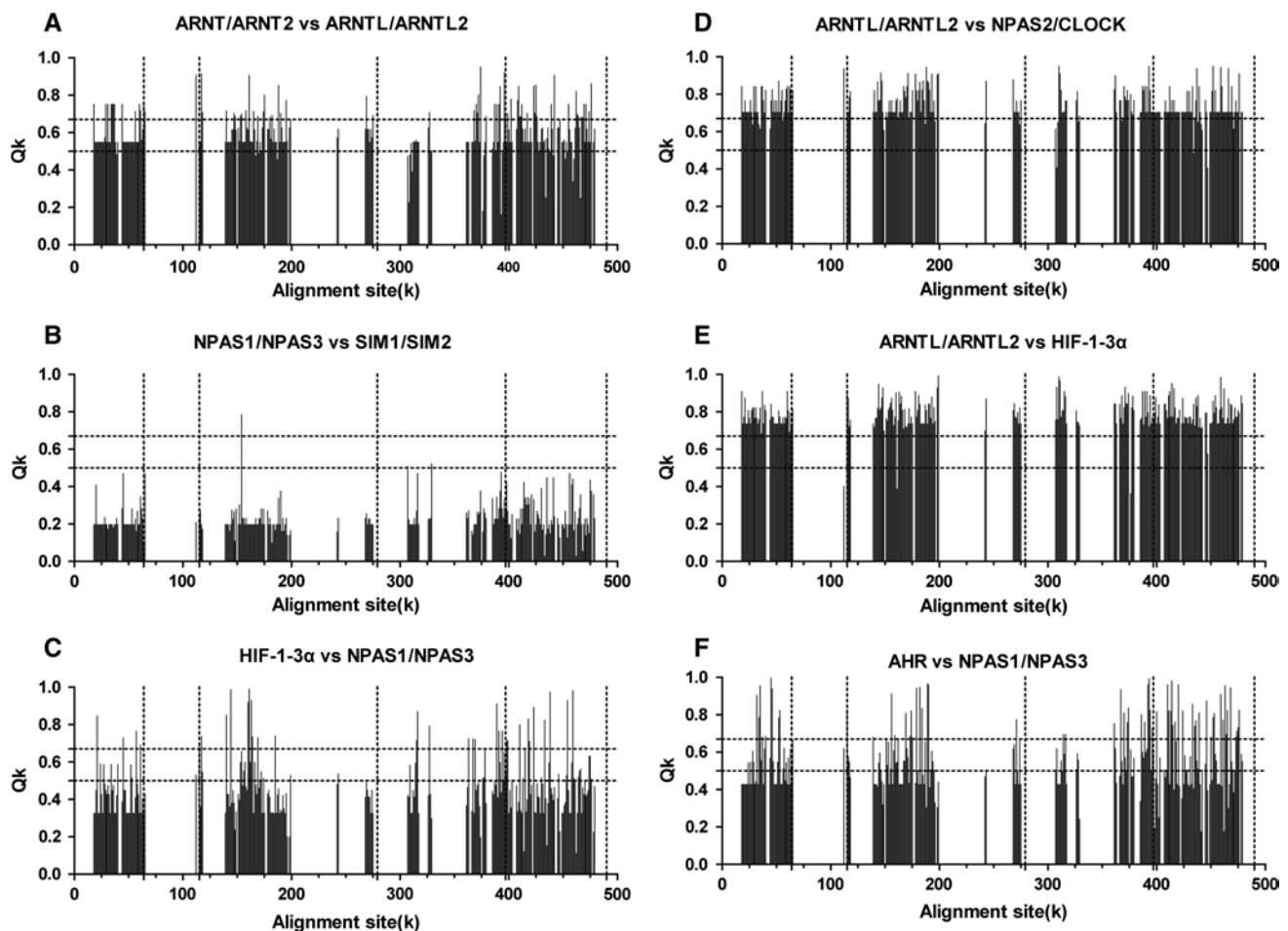


Fig. 3 Site-specific profile for evolution rate change. The *dotted lines* which are perpendicular to X-axis mean the boundaries of bHLH (1-64), PAS (115-279) and PAS_3 (397-490) domains. **a** ARNT/ARNT2 vs ARNTL/ARNTL2. **b** NPAS1/NPAS3 vs

SIM1/SIM2. **c** HIF-1-3 α vs NPAS1/NPAS3. **d** ARNTL/ARNTL2 vs NPAS2/CLOCK. **e** ARNTL/ARNTL2 vs HIF-1-3 α . **f** AHR vs NPAS1/NPAS3

mostly located in PAS_3 domain (Fig. 3c). We also checked the ARNTL/ARNTL2 and NPAS2/CLOCK which has many sites with $Q_k > 0.67$ (Fig. 3d). ARNTL1/2 and HIF-1-3 α are not in neighboring clades and have many functional divergence sites (Fig. 3e). NPAS1/3 and AHR are also not in neighboring clades but have less functional divergence sites (Fig. 3f). We noticed that the three θ s among SIM1/2, HIF-1-3 α and NPAS1/3 are all lower than 0.5, suggesting that they didn't diverge dramatically. ARNT1/2 and ARNTL1/2 are grouped with high bootstraps but their θ is 0.6088, hinting that they underwent functional divergence after duplication.

Analysis of selective pressure

Considering the bHLH, PAS and PAS_3 domains are important for the bHLH/PAS genes, we tested the presence of positive selection at individual amino acid codons based

on the three domains and their intervals by CODEML program (Yang 2007). According to the LRT of site-specific models, model M3 was significantly higher than model M0 ($2\Delta\ln L = 1,805.934$, $p < 0.01$, $df = 4$), indicating that heterogeneous selection is among amino acid sites (Table 3). Three kinds of sites under model M3 had ω values of 0.00929, 0.03115 and 0.09922, suggesting that most amino acid sites underwent strong purifying selection. Model M2a and M1a show no difference ($2\Delta\ln L = 0$, $df = 2$), indicating that no positive selection site is found in these three domains during vertebrate evolution. The fact that model M8 was not significantly higher than model M7 ($2\Delta\ln L = 0$, $df = 2$) also supports the above assertion. These three domains underwent a process of strong purifying selection and the positive selection site may be in the sequences outside these three domains. The other gene families were also reported to have undergone a strong purifying selection (Milenkovic et al. 2010; Teng et al. 2010; Yan

Table 3 Detection of positive selection of bHLH/PAS genes in vertebrates and ciona

Models	np	Estimates of parameters	lnL	LRT pairs	df	2ΔlnL	P
M0: one ratio	1	$\omega = 0.03612$	-48,613.054545	M0/M3	4	1,805.934	<0.01
M1a: neutral	2	$p_0 = 0.99144, (p_1 = 0.00856), \omega_0 = 0.03632, (\omega_1 = 1.00)$	-48,608.1 34821				
M2a: selection	4	$p_0 = 0.99144, p_1 = 0.00511, (p_2 = 0.00345), \omega_0 = 0.03632, (\omega_1 = 1.00), \omega_2 = 1.000$	-48,608.1 34921	M1a/M2	2	0	1
M3: discrete	5	$p_0 = 0.25215, p_1 = 0.45563, (p_2 = 0.29222), \omega_0 = 0.00929, \omega_1 = 0.03115, \omega_2 = 0.09922$	-47,710.0 87328				
M7: beta	2	$p = 1.22018, q = 24.73349$	-47,683.7 86187	M7/M8	2	0	1
M8: beta and ω	4	$p_0 = 0.99999, (p_1 = 0.00001), p = 1.22016, q = 24.73292, \omega = 4.80702$	-47,683.7 90122				

np number of free parameters, lnL log likelihood, LRT likelihood ratio test, df degrees of freedom, 2ΔlnL twice the log-likelihood difference of the models compared

and Cai 2010) especially in HIF-1-3α which also belongs to the bHLH transcription factor (Zhang et al. 2010).

Discussion

In this study, we described the evolution and functional divergence of the bHLH/PAS gene family on a genome-wide scale. We found that bHLH/PAS proteins appeared during the emergence of metazoan (in sponge). Phylogenetic trees indicated that this family can be divided into three groups and its gene members expanded in vertebrates. Exon–intron structure analysis revealed a conserved gene structure with exon phase “1001-0210” in most members and some exon fusion and split events were inferred. Analysis of functional divergence suggested that a site-specific rate shift after gene duplication is a common phenomenon in their evolution. Our study provides new insights for a better understanding into the evolution and origination of the bHLH/PAS family, which may provide clues for further functional analysis.

Origin and evolution of bHLH/PAS gene

The bHLH/PAS proteins are critical regulators of gene expression networks underlying many essential physiological and developmental processes (Kewley et al. 2004). We performed extensive bioinformatics analysis and proposed a model to describe the origin and evolution of the bHLH/PAS gene. In brief this model involved two points: one is that the number of this gene family expanded in vertebrates as compared to invertebrates, the other is that a conserved exon–intron structure with “1001-0210” was retained in most clades during the evolution.

The bHLH proteins are available in plants, fungi and animals (Sailsbery and Dean 2012), whereas PAS domains

can be found in bacterial, eukaryotic and archaeal proteins (Moglich et al. 2009). In our study, we focused on the sequences with both bHLH and PAS domains. The four sequences with two domains in *M. brevicollis* (jgilMonbr1|12346|, jgilMonbr1|26507|, jgilMonbr1|37483| and jgilMonbr1|7924|) implied that the domain fusion or domain rearrangement of bHLH and PAS/PAS_3 occurred before the emergence of metazoa. Considering that the PAS and PAS_3 domains are dimmers (Gilles-Gonzalez and Gonzalez 2004; Kewley et al. 2004), we inferred that the sponge bHLH/PAS genes evolved from a choanoflagellate bHLH/PAS-like primitive sequence by duplication of the PAS domain. Four sponge *A. queenslandica* bHLH/PAS genes are ancestor-like sequences which might be the primitive metazoan bHLH/PAS genes. They were distributed in the root of the two clades group II and group III (Fig. 1), implying that the early metazoan may contain two copies of these families. Based on the conserved exon pattern (Fig. 4a), we inferred that the group I may be a duplication of group II. Most invertebrates contained 4–7 bHLH/PAS genes, suggested that the gene number of this gene family is relatively conserved during the evolution of invertebrates. However, the number of this gene family expanded into 10–14 in vertebrates, which may be a result of 2R WGD (Hokamp et al. 2003; Panopoulou and Poustka 2005).

We identified 18 bHLH/PAS genes in zebrafish (Table S3). The large number of genes may be due to the Teleost specific 3R WGD because we found some genes with two copies in zebrafish, such as SIM1, CLOCK and HIF-2α (Taylor et al. 2001; Vandepoele et al. 2004; Meyer and Van de Peer 2005). The three nematode and seven fly bHLH/PAS genes located in the corresponding vertebrate clusters hint that these genes were conserved in protein level among these three domain (bHLH + PAS + PAS_3) sequences during evolution. However, their exon–intron structures

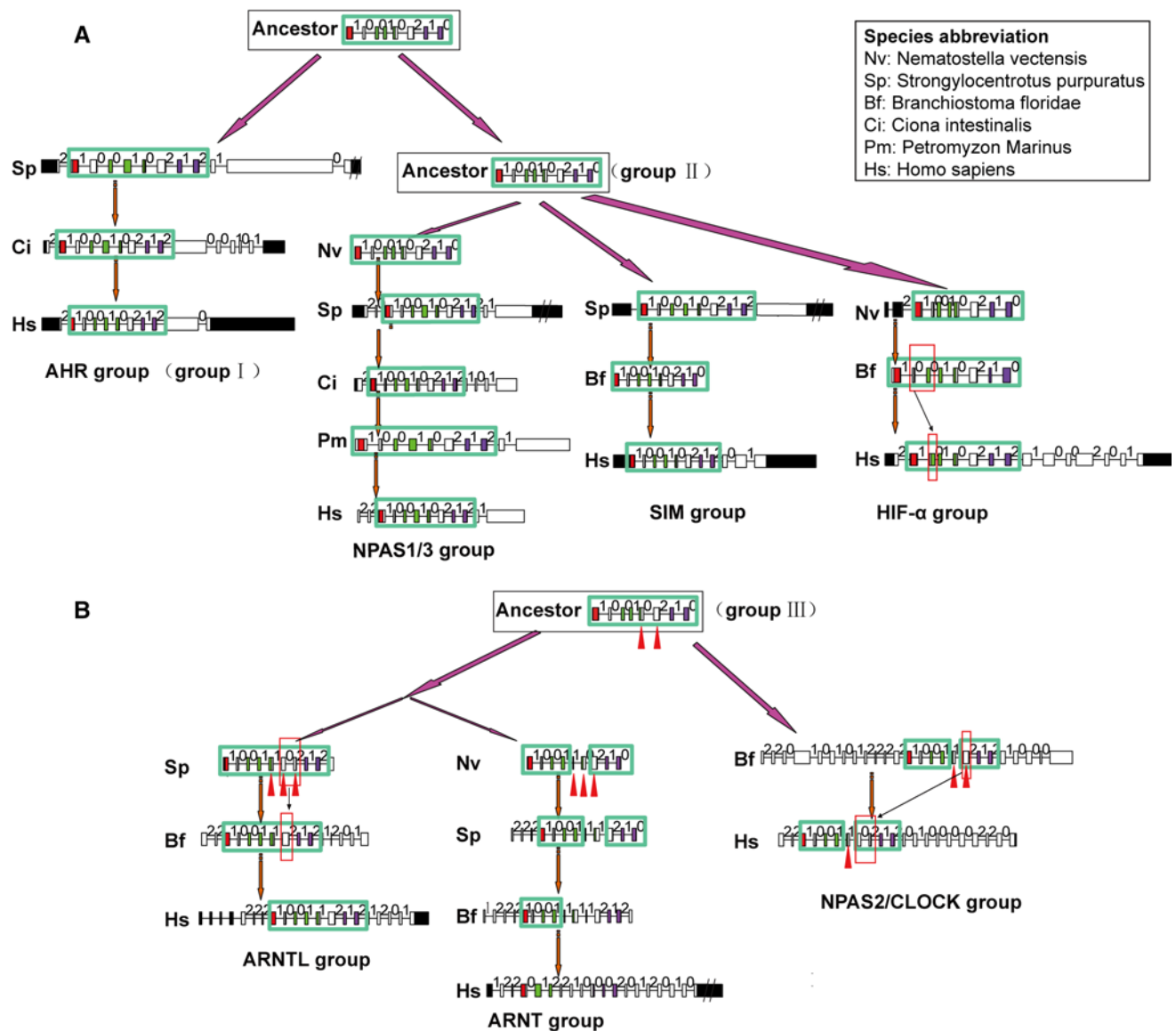


Fig. 4 The evolutionary model of the exon–intron structure of bHLH/PAS genes. The sequences in pictures were from representative primitive species and human. Abbreviations of species names were in the pictures. We found that a conserved exon–intron structure with “1001-0210” existed in many species. The two or three exons indicated by red boxes and arrows show the exon split, fusion or rearrangement in

HIF-1-3 α , ARNTL1/2, NPAS2/CLOCK clades, which were detailed in Fig. 2 and Fig. S5. Three groups (I–III) corresponding to Fig. 1 were labeled in the diagram. **a** AHR, SIM1, SIM2, HIF-1 α , HIF-2 α , HIF-3 α , NPAS1 and NPAS3 retained this conserved exon phases except HIF-1-3 α . **b** NPAS2, CLOCK, ARNT, ARNT2, ARNTL and ARNTL2 changed a little in this conserved exon phases (color figure online)

were changed as compared to the deuterostomes and sea anemone, maybe as a result of rearrangement in protostome (Putnam et al. 2007).

The yellow blocks as shown in Fig. S6 indicate the retentions and changes of the conserved exon–intron structure in three groups. The results show that the ancestor contained the exon–intron structure of eight exons with exon phase “1001-0210”, which can be found in sea anemone, sea urchin and amphioxus. The length outside these three domains (bHLH + PAS + PAS_3) varies among these 11 members of human bHLH/PAS genes, suggesting

that the conservative only existed in sequences of these three domains. The clade of HIF- α , NPAS1/3 and SIM1/2 retained the “1001-0212” and did not change in evolution except HIF- α which showed little change (Fig. 4a). We noticed that other clades clustering together with a high Bayesian posterior probability of 100 in Bayesian tree (Fig. 1) have experienced a little change in the conserved exon phase “1001-0212”. The clade of NPAS2 and CLOCK also retained most of the exon–intron structure with an exon rearrangement and became “1001-10212” (Fig. 4b). The clade of ARNT/ARNT2 changed largely in

vertebrates although the NvARNT still retained the ancestral exon phases and the eight exons (Fig. 4b).

We also noticed three exon changes in HIF-1-3 α , ARNTL1/2 and NPAS2/CLOCK. The two exons in HIF- α with phase “00” in “1001” fused during the emergence of vertebrate (Fig. 2a). An exon fusion occurred in sea urchin and amphioxus (Fig. 2b). An exon split may occur in amphioxus and vertebrate (Fig. 2c). Two exons with phase “02” in “0210” rearranged in ARNTL1/2, ARNT1/2 and NPAS2/CLOCK (Fig. S5). Interestingly, these changes all occurred in intervals between the PAS and PAS_3 domains but not in the three functional domains (bHLH, PAS and PAS_3). In summary, the ancestral exon–intron structure “1001-0210” was retained in most members of this gene family, implying its important functions in evolution.

Functional divergence and purifying selection of the bHLH/PAS genes

Our results show that the functional divergence occurred in early gene duplication events before vertebrates emerged. Strong purifying selection plays a role in the origin of new subfamilies in bHLH/PAS evolution. HIF- α proteins in vertebrates were studied in molecular selection and functional divergence, and they also underwent a strong purifying selection (Zhang et al. 2010). A purifying selection was also detected in other genes such as anoctamin, CYP3, NMDA receptor subunit 2 and UCP genes (Hughes et al. 2009; Milenkovic et al. 2010; Teng et al. 2010; Yan and Cai 2010). Our result is consistent with theirs that a strong purifying selection plays an essential role during the evolution of some gene families, especially the bHLH/PAS subfamily HIF- α proteins in vertebrates (Zhang et al. 2010). Another study also referred the HIF- α and they focused on the evolution of PHD–HIF Oxygen-Sensing System, their result that three HIF isoforms in extant vertebrate genomes are products of 2R is consistent with our findings (Rytönen et al. 2011).

Acknowledgments We would like to thank Dr Siluo Huang (Huazhong University of Science and Technology) for advice on this study. This work was supported by the following funds to A.Y.G: National Natural Science Foundation of China (NSFC) (31171271, 31270885), Young Teachers’ Fund for Doctor Stations, Ministry of Education of China (20110142120042), the Scientific Research Fund for the Returned Overseas Chinese Scholars, Ministry of Education of China, and fund from the State Key Laboratory of Freshwater Ecology and Biotechnology (2012FB02).

References

- Abascal F, Zardoya R et al (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* (Oxford, England) 21(9):2104–2105
- Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* 9(12):938–950
- Erbel-Sieler C, Dudley C et al (2004) Behavioral and regulatory abnormalities in mice deficient in the NPAS1 and NPAS3 transcription factors. *Proc Natl Acad Sci USA* 101(37):13648–13653
- Esser C, Rannug A et al (2009) The aryl hydrocarbon receptor in immunity. *Trends Immunol* 30(9):447–454
- Gilles-Gonzalez MA, Gonzalez G (2004) Signal transduction by heme-containing PAS-domain proteins. *J Appl Physiol* 96(2):774–783
- Griffin EA Jr, Staknis D et al (1999) Light-independent role of CRY1 and CRY2 in the mammalian circadian clock. *Science* (New York, NY) 286(5440):768–771
- Gu X (1999) Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol* 16(12):1664–1674
- Gu X (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol* 18(4):453–464
- Gu X, Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* (Oxford, England) 18(3):500–501
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696–704
- Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered* 100(5):605–617
- Hahn ME, Karchner SI et al (2006) Unexpected diversity of aryl hydrocarbon receptors in non-mammalian vertebrates: insights from comparative genomics. *J Exp Zool A Comp Exp Biol* 305(9):693–706
- Hokamp K, McLysaght A et al (2003) The 2R hypothesis and the human genome sequence. *J Struct Funct Genomics* 3(1–4):95–110
- Huang Y, Niu B et al (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* (Oxford, England) 26(5):680–682
- Hughes DA, Jastroch M et al (2009) Molecular evolution of UCP1 and the evolutionary history of mammalian non-shivering thermogenesis. *BMC Evol Biol* 9:4
- Kewley RJ, Whitelaw ML et al (2004) The mammalian basic helix-loop-helix/PAS family of transcriptional regulators. *Int J Biochem Cell Biol* 36(2):189–204
- Kume K, Zylka MJ et al (1999) mCRY1 and mCRY2 are essential components of the negative limb of the circadian clock feedback loop. *Cell* 98(2):193–205
- Larkin MA, Blackshields G et al (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* (Oxford, England) 23(21):2947–2948
- Linne V, Eriksson BJ et al (2012) Single-minded and the evolution of the ventral midline in arthropods. *Dev Biol* 364(1):66–76
- Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays* 27(9):937–945
- Milenkovic VM, Brockmann M et al (2010) Evolution and functional divergence of the anoctamin family of membrane proteins. *BMC Evol Biol* 10:319
- Moglich A, Ayers RA et al (2009) Structure and signaling mechanism of Per-ARNT-Sim domains. *Structure* 17(10):1282–1294
- Panopoulou G, Poustka AJ (2005) Timing and mechanism of ancient vertebrate genome duplications—the adventure of a hypothesis. *Trends Genet* 21(10):559–567
- Pickard BS, Malloy MP et al (2005) Disruption of a brain transcription factor, NPAS3, is associated with schizophrenia and learning disability. *Am J Med Genet B Neuropsychiatr Genet* 136B(1):26–32
- Ploski JE, Monsey MS et al (2011) The neuronal PAS domain protein 4 (Npas4) is required for new and reactivated fear memories. *PLoS ONE* 6(8):e23760
- Preitner N, Damiola F et al (2002) The orphan nuclear receptor REV-ERB α controls circadian transcription within the positive limb of the mammalian circadian oscillator. *Cell* 110(2):251–260

- Putnam NH, Srivastava M et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* (New York, NY) 317(5834):86–94
- Reick M, Garcia JA et al (2001) NPAS2: an analog of clock operative in the mammalian forebrain. *Science* (New York, NY) 293(5529):506–509
- Reitzel AM, Behrendt L et al (2010) Light entrained rhythmic gene expression in the sea anemone *Nematostella vectensis*: the evolution of the animal circadian clock. *PLoS ONE* 5(9):e12805
- Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* (Oxford, England) 19(12):1572–1574
- Rytkonen KT, Williams TA et al (2011) Molecular evolution of the metazoan PHD-HIF oxygen-sensing system. *Mol Biol Evol* 28(6):1913–1926
- Sailsbery JK, Dean RA (2012) Accurate discrimination of bHLH domains in plants, animals, and fungi using biologically meaningful sites. *BMC Evol Biol* 12:154
- Stevens EA, Mezrich JD et al (2009) The aryl hydrocarbon receptor: a perspective on potential roles in the immune system. *Immunology* 127(3):299–311
- Suyama M, Torrents D et al (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34(Web Server issue):W609–W612
- Tamura KDJ, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599
- Taylor JS, Van de Peer Y et al (2001) Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc London* 356(1414):1661–1679
- Teng H, Cai W et al (2010) Evolutionary mode and functional divergence of vertebrate NMDA receptor subunit 2 genes. *PLoS ONE* 5(10):e13342
- Vandepoele K, De Vos W et al (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci USA* 101(6):1638–1643
- Yan J, Cai Z (2010) Molecular evolution and functional divergence of the cytochrome P450 3 (CYP3) Family in Actinopterygii (ray-finned fish). *PLoS ONE* 5(12):e14276
- Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591
- Zhang X, Wang M et al (2010) Molecular selection and functional divergence of HIF-alpha proteins in vertebrates. *Genetica* 138(11–12):1241–1250
- Zhou M, Yan J et al (2012) Comparative and evolutionary analysis of the HES/HEY gene family reveal exon/intron loss and teleost specific duplication events. *PLoS ONE* 7(7):e40649