Differentiation (2008) 76:1006–1022 DOI: 10.1111/j.1432-0436.2008.00285.x Journal compilation © 2008, International Society of Differentiation

ORIGINAL ARTICLE

Jeffrey D. Stevens · Eric H. Roalson · Michael K. Skinner

Phylogenetic and expression analysis of the basic helix-loop-helix transcription factor gene family: genomic approach to cellular differentiation

Received December 7, 2007; accepted in revised form March 6, 2008

Abstract A phylogenetic analysis of seven different species (human, mouse, rat, worm, fly, yeast, and plant) utilizing all (541) basic helix-loop-helix (bHLH) genes identified, including expressed sequence tags (EST), was performed. A super-tree involving six clades and a structural categorization involving the entire coding sequence was established. A nomenclature was developed based on clade distribution to discuss the functional and ancestral relationships of all the genes. The position/location of specific genes on the phylogenetic tree in relation to known bHLH factors allows for predictions of the potential functions of uncharacterized bHLH factors, including EST's. A genomic analysis using microarrays for four different mouse cell types (i.e. Sertoli, Schwann, thymic, and muscle) was performed and considered all known bHLH family members on the microarray for comparison. Cell-specific groups of bHLH genes helped clarify those bHLH genes potentially involved in cell specific differentiation. This phylogenetic and genomic analysis of the bHLH gene family has revealed unique aspects of the evolution and functional relationships of the different genes in the bHLH gene family.

Key words bHLH · testis · Sertoli cell · Schwann cell · muscle cell · thymic cell · microarray · phylogenetic human · rat · mouse · *Drosophila* · inset · *C. elegans* · Arabidopsis · plant · yeast

Jeffrey D. Stevens · Michael K. Skinner (✉)
Center for Reproductive Biology
School of Molecular Biosciences
Pullman, WA 99164-4231, USA
Tel: +1 509 335 1024
Fax: +1 509 335 2176
E-mail: skinner@wsu.edu

Eric H. Roalson
School of Biological Sciences
Washington State University
Pullman, WA 99164-4231, USA

Grant Sponsorship: This work was supported by an NIH grant to Michael K. Skinner, 5R01 HD04381-04.

Introduction

Identification of the basic helix-loop-helix (bHLH) motif first occurred in 1989 (Murre et al., 1989) when E12 and E47 were discovered in the murine genome. Since this time, numerous bHLH proteins have been identified in animals, plants, and fungi. In 1997, the first large-scale phylogenetic analysis was performed (Atchley and Fitch, 1997) leading to a "natural" classification of different families of bHLH transcription factors. This classification was performed using only the bHLH motif because the flanking regions for proteins from independent clades are very divergent. This classification led to the postulation of four distinct groups based on amino-acid patterns and E-box-binding specificity (Atchley and Fitch, 1997). This classification segregated bHLH proteins under Class A, B, C, or D in an attempt to functionally segregate bHLH proteins. Unfortunately the majority of the bHLH genes do not have known functions or have multiple functions such that only a small sub-group of bHLH proteins can utilize this original classification. Class A includes several tissue-specific bHLH proteins (Hassan and Bellen, 2000) as well as several ubiquitously expressed bHLH proteins such as the E2A gene products E12 and E47, HEB, and E2-2 (Murre et al., 1989; Atchley and Fitch, 1997). Class B proteins represent a

large group of functionally unrelated proteins that are involved in various cellular and developmental processes (Henriksson and Luscher, 1996; Facchini and Penn, 1998; Goding, 2000). Proteins in this group include Myod and myogenin, involved in muscle cell differentiation (Ishibashi et al., 2005; Tang et al., 2006), Ngn and Mash1 involved in neurogenesis (Nakada et al., 2004; Kageyama et al., 2005) and Hand involved in heart development (Thattaliyath et al., 2002). Several of the proteins in this group contain another functionally important motif known as the leucine zipper (Atchley and Fitch, 1997). The leucine zipper (Zip) motif is a protein interaction domain also present in the CREB family of transcription factors and can heterodimerize or homodimerize to bind DNA (Vinson et al., 2006). A subclass of the class B bHLH proteins that function as repressors (i.e. hairy and enhancer-of-split proteins) were first identified in *Drosophila*. Many vertebrate homologs have been subsequently identified including the Hes group of genes (Davis and Turner, 2001). These proteins contain another common structure known as the Orange domain located just C-terminal to the bHLH domain (Taelman et al., 2004). Members of the bHLH/Orange subclass of the class B proteins act as repressors that inhibit target gene expression by acting as direct or indirect DNA-binding-dependent transcriptional repressors or by sequestering positive bHLH factors or their common heterodimer partners (Chin et al., 2000; Giagtzoglou et al., 2003). The function of the Orange domain is not well understood, but may play a role in conferring specificity of binding to certain family members (Dawson et al., 1995) or have a role in transcriptional repression (Castella et al., 2000). An additional structural characteristic of the Hairy and E(spl) bHLH/Orange proteins is the presence of a C-terminal WRPW motif that binds the co-repressor Groucho and its mammalian homologs, the TLE proteins (Paroush et al., 1994; Fisher and Caudy, 1998; Chen and Courey, 2000). Class B bHLH proteins are postulated to not homodimerize, rather they are believed to heterodimerize with Class A bHLH proteins. Class C bHLH proteins also contain one or more PAS domains (Crews, 1998). This domain allows for dimerization between PAS proteins, non-PAS proteins and the binding of small molecules (e.g. dioxin) (Crews, 1998). Examples of class C bHLH proteins include HIF1 involved in regulation of hypoxia, and Sim proteins involved in food intake behavior (Yang et al., 2004). These proteins tend to be ubiquitous and are believed to bind a DNA sequence different from the common E-box (Crews, 1998; Crews and Fan, 1999; Taylor and Zhulin, 1999). Class D includes HLH proteins that lack a basic domain and are thus unable to bind DNA. These proteins are called inhibitors of differentiation (Id). In mammals, there are four known Id proteins that appear to have differential expression based on cell type (Chaudhary et al., 2001). Id1, Id2, and Id3 are thought to be ubiquitously expressed, while Id4 is primarily expressed in the testis (Chaudhary et al., 2001), brain and kidney (van Cruchten et al., 1998). A fifth group of bHLH proteins has been suggested (Crozatier et al., 1996), but phylogenetic analysis of this group is difficult as the HLH domain is highly divergent from the conserved bHLH motif. This group is known as the COE family and is characterized by the presence of an additional COE domain involved in dimerization and DNA binding. Owing to the increased size and diversity of the bHLH gene family, this original classification (i.e. Class A–D) has become inadequate and misleading. A classification that can incorporate the entire gene family and show relatedness is required.

Several large and small scale phylogenetic analysis of the bHLH transcription factor family have been performed for mammals (Atchley and Fitch, 1997; Ledent et al., 2002) and plants (Buck and Atchley, 2003; Heim et al., 2003; Toledo-Ortiz et al., 2003). Most of these analyses have utilized only the bHLH domain while the remaining portion of the protein is considered to be too divergent. A recent analysis in plants has used the entire coding sequence (Li et al., 2006). Owing to a number of additional domains being associated with the bHLH proteins, utilizing the entire coding region in a large-scale phylogenetic analysis for classification is needed to allow better identification of the proper class structure of the bHLH genes. In addition, having a large-scale analysis of the known mammalian bHLH transcription factors will allow for the entire gene family to be used in concert with expression analysis in the investigation of cellular differentiation.

Terminal cellular differentiation occurs when a cell exits the cell cycle, becomes post mitotic, and develops specialized cellular functions associated with the differentiated gene expression profile. These terminally differentiated cells can often not be replaced if lost. Examples of terminally differentiated cells include myocytes (Tam et al., 1995; Wei and Paterson, 2001), neurons (Yoshikawa, 2000), and Sertoli cells (Skinner, 1991). While the role of the bHLH family of transcription factors has been partially identified in this terminal differentiated state for myocytes and neurons (Nakada et al., 2004; Ishibashi et al., 2005; Kageyama et al., 2005; Tang et al., 2006), factors responsible for Sertoli cells to undergo terminal differentiation remain to be elucidated.

Phylogenetic analysis of an entire gene family between species allows for the appropriate structural and functional distribution of genes, identifies gene duplications and species conservation, as well as reveals evolutionary considerations. The phylogenetic analysis of the bHLH gene family identifies new structural and functional relationships between bHLH genes and helps organize the gene family. A comparison of the phylogenetic bHLH gene family information with DNA microarray analyses from divergent cell types identifies the genes specific or unique to the different cells. This phylogenetic and genomic approach enhances the