

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

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

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.

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 ex-

pressed, 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

analysis of a transcription factor family associated with cellular differentiation and allows the investigation of the entire bHLH gene family.

Materials and methods

Identification of bHLH transcription factors

A search of the Ensembl database using known bHLH transcription factors was performed for human, rat, mouse, *Saccharomyces cerevisiae*, *Drosophila*, and *Caenorhabditis elegans*. Following identification of bHLH transcription factors in Ensembl, verification of these genes was performed using a NCBI protein search. A basic phylogenetic tree was generated in Vector NTI to discern clade distribution of the known proteins. One bHLH protein from each clade was then chosen; its bHLH domain selected and blasted in NCBI BLASTp to identify more genes not listed in the Ensembl database as well as possible EST's. The Arabidopsis bHLH genes were taken from a previous analysis (Toledo-Ortiz et al., 2003). After generating the total list of bHLH genes for these seven species (i.e. 541 genes), a phylogenetic tree was generated in Vector NTI to determine if there were duplicate genes listed. Any EST with high homology to a known gene was then aligned (NCBI) with the known gene to determine similarity. If identical, the EST was deleted from the final analysis.

Generation of phylogenetic tree

An initial phylogenetic tree was generated using an amino acid alignment created using ClustalX 1.83 (Thompson et al., 1997) and analyzed with PHYML v.2.4.4 (Guindon and Gascuel, 2003), employing the JTT amino-acid substitution model (Jones et al., 1992). Because of the large numbers of genes (i.e. 541), this starting phylogenetic hypothesis was then divided into six major clades based on the preliminary tree structure for more rigorous analysis. After separating the matrix into separate matrices of 27, 28, 99, 107, 131, and 136 sequences, these amino-acid sequences were realigned with ClustalX to minimize the potential effects of having divergent sequences in the alignment. These final alignments were then analyzed using Bayesian inference analysis and was performed on the various matrices using MrBayes v.3.1 (Huelsenbeck and Ronquist, 2001). Ten million generations were run with four chains (Markov Chain Monte Carlo) and a tree was saved every 100 generations. In order to test for the occurrence of stationarity, convergence and mixing within ten million generations, multiple analyses were started from different random locations in tree space. The posterior probability distributions from these separate replicates were compared for convergence with the same posterior probabilities across branches. Majority rule consensus trees of those sampled in Bayesian inference analyses yielded probabilities that the clades are monophyletic (Lewis, 2001). The trees from the MrBayes analysis were loaded into PAUP*4.0 (Sinauer Associates Inc., Sunderland, MA), discarding the trees generated within the first 2,000,000 generations (those sampled during the "burnin" of the chain (Huelsenbeck and Ronquist, 2001), to only include trees after stationarity was established. Consensus trees were then created to display branches with posterior probabilities greater than 50%. A single tree was not created in this case because a point estimate might mislead interpretation of inferred relationships. The preference is to focus on those branches with strong support.

Affymetrix probe search for bHLH genes

The complete mouse bHLH gene list was taken from the phylogenetic analysis and used in a search of the Affymetrix (Affymetrix,

Santa Clara, CA) database for MGU74v2 array chips in an effort to locate all bHLH genes and their associated Affymetrix ID's on the chips. If the gene name did not yield an Affymetrix ID the actual sequence of the gene was blasted against the MGU74v2 probe set to ensure the gene was or was not on the array chips. Because there are many EST's in the phylogenetic tree, these gene sequences were also used in an Affymetrix blast search for their associated Affymetrix ID. The MGU74v2 array chips contain three separate chips, the A chip contains 12,480 total genes of which the majority are fully annotated in NCBI. The B and C chips contain 12,477 and 11,934 genes respectively, many of these genes are EST's or not fully annotated genes. Of the 107 identified mouse bHLH genes in the phylogenetic analysis, 81 were identified as existing on the MGU74v2 array chips. Owing to limitations in obtaining microarray analysis for other cell types that were also performed on all three MGU74v2 array chips, we limited our bHLH gene list to only those genes on the A chip. Therefore, only a subset of the potential bHLH genes expressed will be assessed.

Microarray analysis

High quality mouse RNA samples of at least 5 µg and with a minimum OD260/280 ratio of 1.8 were analyzed as previously described (McLean et al., 2002; Nef et al., 2005; Small et al., 2005). Briefly, RNA was transcribed into cDNA, which was transcribed into biotin labeled RNA. Biotin labeled RNA was then hybridized to either mouse MGU74v2a arrays containing approximately 37,000 total transcripts (Affymetrix) and visualized by labeling with phycoerythrin-coupled avidin. Hybridized chips were visualized on an Affymetrix Scanner 3000 (Affymetrix). Once raw data was obtained, data was processed using GCOS version 1.1 software (Affymetrix) and analyzed by Genespring version 7.2 (Silicon Genetics, Redwood City, CA) software.

Initial analysis of microarray data was performed as previously described (Small et al., 2005). Microarray hybridization data was examined for physical anomalies on the chip and background noise above a value of 3. Default GCOS statistical values were used for analysis. All probe sets were scaled to a mean of 125. An absolute analysis was performed with GCOS to assess the relative abundance of the transcripts on the chips based on signal and detection calls (present, absent, or marginal). This information was imported into GeneSpring (Silicon Genetics) and normalized using the recommended default normalization methods. This includes setting signal values below 0.01–0.01, total chip normalization to the 50th percentile, and normalization of each chip to the median which allows visualization of data based on relative abundance for a given sample rather than by comparison to a specific control value (Small et al., 2005). Transcripts with statistically significant presence calls ($p < 0.05$) and raw signal values above 75 were selected for comparison to other microarray data. All the cell types used for the microarrays were freshly isolated cell preparations using fluorescent-activated cell sorting (FACS) and/or enzymatic digestion and gravity sedimentations. Gene expression data from mouse Schwann cells (Buchstaller et al., 2004) (GEO:GSE972), mouse thymic medullary epithelial cells (Anderson et al., 2002) (GEO: GSE85) and the mouse muscle cell (GEO:6487) (http://pepr.cnmcresearch.org/browse.do?action=list_prj_exp&projectId=151) were obtained from the Gene Expression Omnibus (GEO) available through NCBI. This data was obtained from MGU74v2a arrays. A minimum of two different microarray chips and experiments were used to obtain the mean data utilized. The raw data was pulled through GCOS and GeneSpring in the same manner as the mouse chips above and was used for comparison of the different cell types. Validation of gene expression from the microarray analysis used previous literature reports confirming (e.g. quantitative PCR) the expression of the selected genes. Over 16 different experiments from the various cell types and literature all confirmed the gene expression from the microarray data.

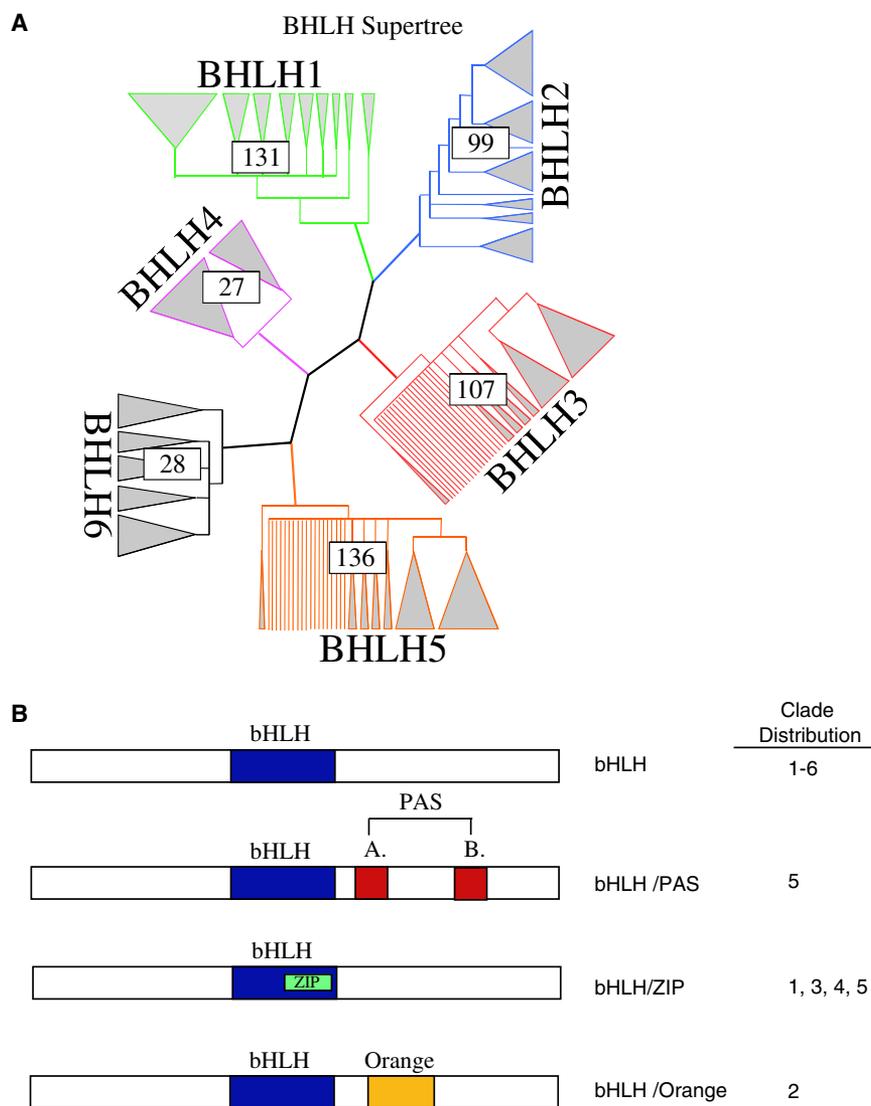


Fig. 1 (A) bHLH supertree showing individual clade distribution, how many genes are in each clade, and representative grouping of individual genes within each clade. (B) bHLH protein categories with additional domains (i.e. Zip, PAS, and Orange) and the clade distribution of the categories.

Results

bHLH phylogenetic analyses

The preliminary maximum likelihood analysis of all the 541 bHLH genes found in the seven species (human, mouse, rat, worm, fly, yeast, plant) resulted in a single topology. Previous phylogenetic analyses of this gene family primarily used only the bHLH region of the protein, limiting the potential informative sequence. Previous studies also limited the number of genes in their total analysis, making *a priori* assumptions about gene copy relatedness before analysis. This has the potential to introduce misinterpretation of bHLH copy relatedness. Therefore, these *a priori* assumptions were minimized in the current study by including the entire coding region where available and all nonidentical gene copies from the seven species. Owing to the size of the matrix involved, it was necessary to break the tree into

smaller sub-trees for more rigorous analysis. These divisions separated major clades of reasonable size from the initial maximum likelihood tree. It should be noted that these divisions were made subjectively and focused on creating related matrices of reasonable size for further analysis. Detailed analysis of these matrices demonstrated the divisions made were accurate. Six clades resulted from this division (Fig. 1A), with clade 1 containing only mammalian genes, clades 2–5 contained genes from a mixture of species, and clade 6 including the plant *Arabidopsis* bHLH genes only. All of the first five clades (Fig. 1A) contain bHLH genes that are typically thought of as Classes A and B factors showing that no unique division based on earlier phylogenetic analysis for these groups is present.

Analysis of each clade was performed to determine similarities among extra domains and allow comparative protein functional analysis (Fig. 1B). Minimally four different categories of bHLH proteins have been

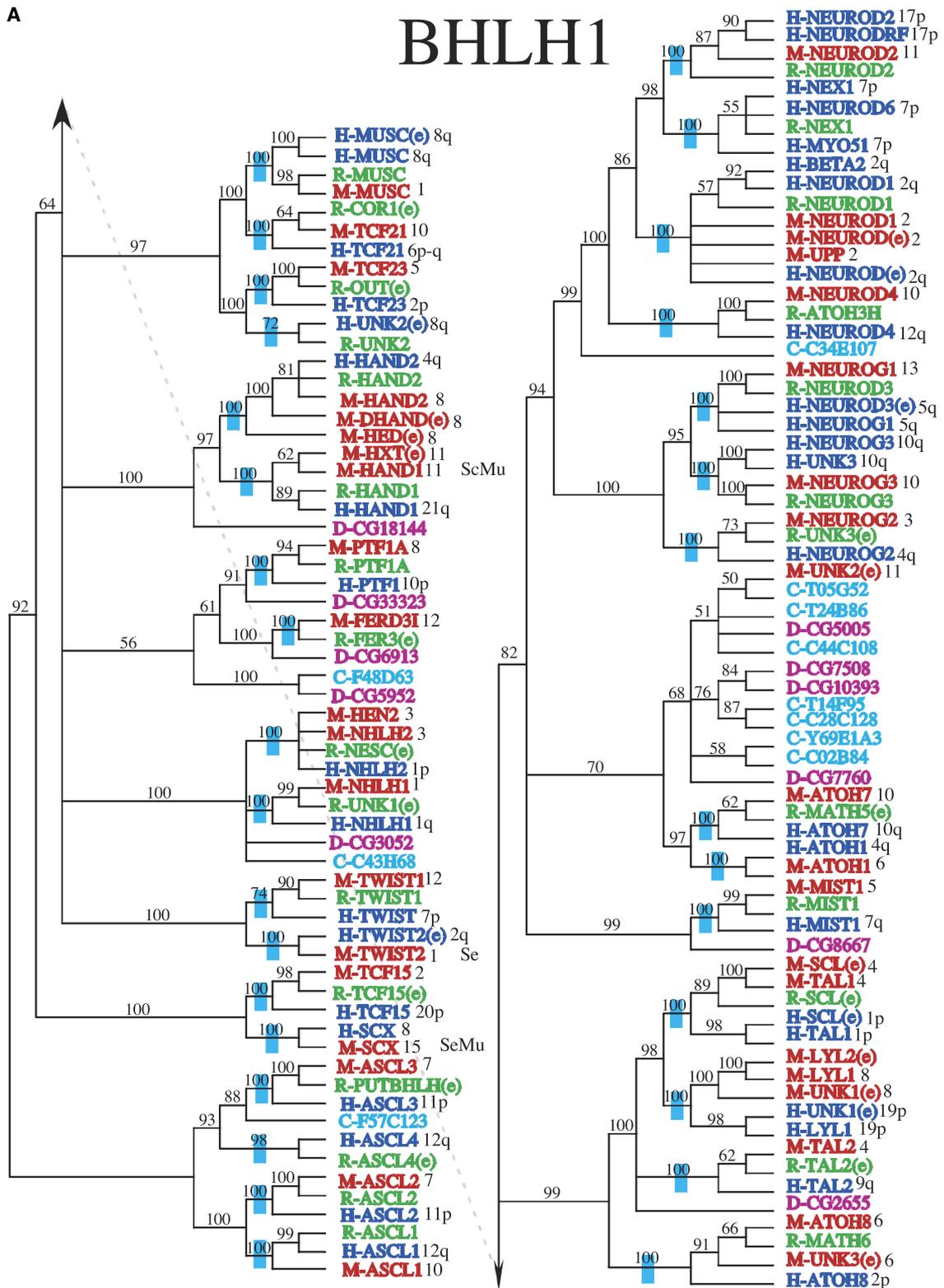


Fig. 2 Phylogenetic relationship of individual genes in (A) clade 1, (B) clade 2, (C) clade 3, (D) clade 4 and 6, and (E) clade 5 with relatedness as determined by Bayesian posterior probability indicated by the blue hatch bar and number to the left of the genes.

Chromosomal location for mouse and human genes is given to the right of the gene names. The species is indicated by blue (human), green (rat), red (mouse), magenta (*Drosophila*), pink (*C. elegans*), and orange (*Arabidopsis*), with the species letter in front of the name.

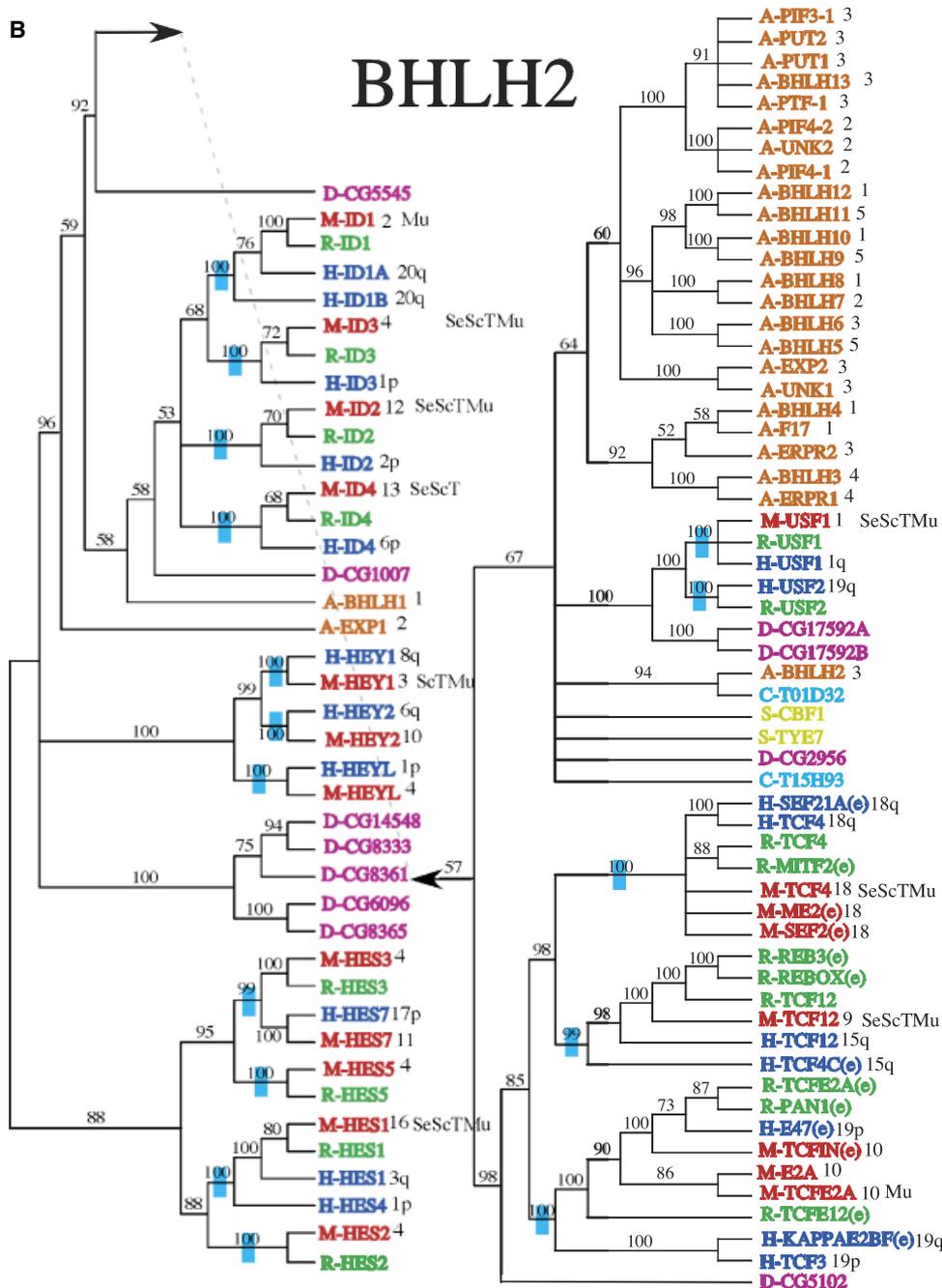


Fig. 2 (Continued)

identified. All contain a bHLH domain and the additional domains include orange, PAS and Zip domains (Fig. 1B). Clade 1 is made up primarily of mammalian genes that were previously considered group B proteins (Fig. 2A). Many of the genes in this clade are also ESTs that have not been fully characterized. Clade 2 is made up of the Id genes that are involved in sequestering other bHLH proteins and inhibiting them from binding DNA (Fig. 2B), as well as the inhibitory Hey and Hes (bHLHb37-b39) genes. Similar to the Id proteins, these

genes are involved in the negative regulation of differentiation by sequestering other bHLH proteins. The Hey and Hes genes contain a second domain, the Orange domain, the function of which is currently unknown. All the bHLH proteins containing an Orange domain localized to clade 2 (Fig. 1B). Clade 2 includes bHLH proteins previously classified in bHLH groups A, B, and D (Fig. 2B) suggesting no phylogenetic segregation of this previous bHLH family categorization. Clade 3 contains a large proportion of Arabidopsis

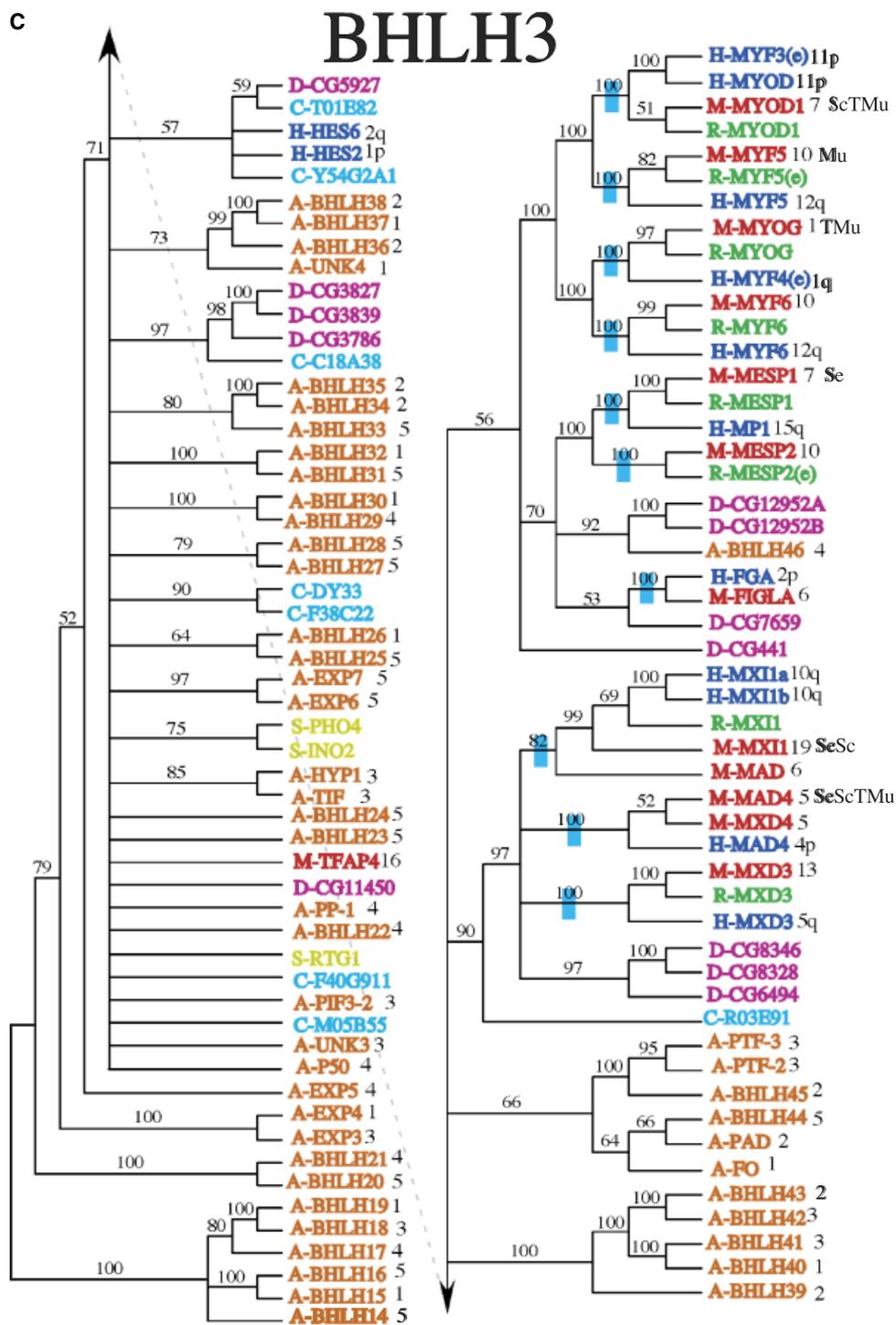


Fig. 2 (Continued)

genes, as well as mammalian genes that are involved in myogenesis and max (bHLHd4-d8) interaction proteins (Fig. 2C). The mammalian genes were all previously classified as group B proteins. Clade 4 contains mostly genes that contain both a bHLH region and a leucine zipper (Zip) region (Fig. 2D). The Zip region is involved in protein interactions and DNA binding, and was pre-

viously classified as group B proteins. Clade 4 is small and distantly related to clades 3 and 5 in the initial analysis, which is why this clade was not included within those larger clades. However, there are several genes in clade 1, 3, and 5 that also contain a leucine zipper region (Figs. 1B,2E), suggesting that the functional structure of this region needs to be explored

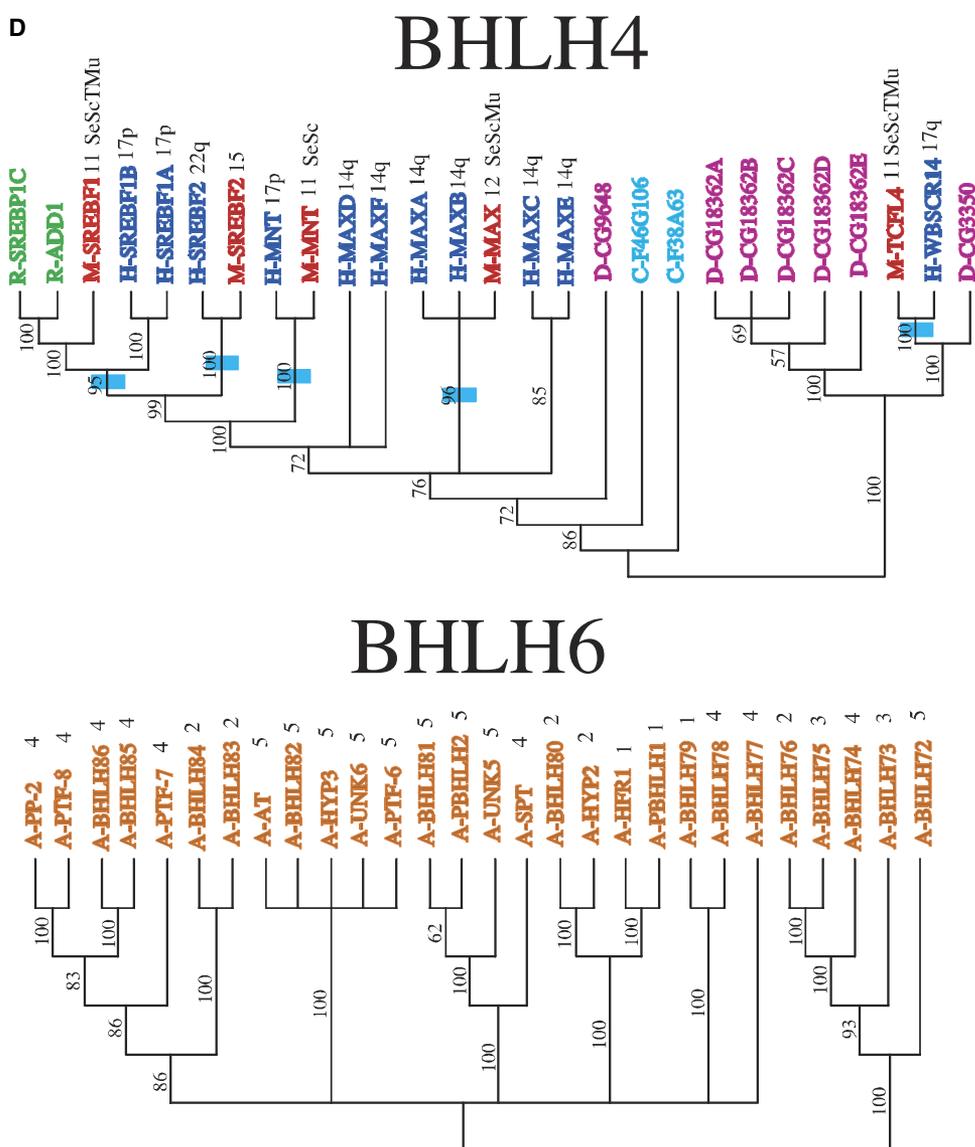


Fig. 2 (Continued)

further. In addition to genes containing a Zip region, clade 5 also contains genes that have one or more PAS domains (Figs. 1B,2E). These genes were previously included in the group C proteins. Clade 6 contains Arabidopsis plant genes only (Fig. 2D). The distribution of the different categories of bHLH proteins that contain other domains (Fig. 1B) demonstrates several of the bHLH categories are isolated to an individual clade. The phylogenetic bHLH family tree generated demonstrates an appropriate segregation of bHLH proteins with similar protein domains and functionally localized genes (Fig. 2). Those genes, previously shown to be related functionally, co-localized into appropriate clades and clusters. Observations suggest that the previous Classes, A, B, C, D categories, do not segregate. Based on the current phylogenetic analysis unknown or non-

characterized genes localized with functionally known bHLH genes can be predicted to potentially have a functional relationship that can now be investigated.

Nomenclature

Further exploration of the supertree and Bayesian consensus trees indicates a preponderance of genes with multiple gene names that have little phylogenetic information and the presence of a large number of EST's without proper names. This has led to the need for a consistent nomenclature to allow a discussion of this gene family. During the blast searches for bHLH genes it became apparent that there are many different names for the same transcription factor. An example of

Table S1. The 541 genes were named and close homologs (e.g. splice variants) were also given letter designates when required [e.g. E2A products E12 (bHLHb21a) and E47 (bHLHb21b) splice variants]. The genomic information of specific genes can include sequencing errors to suggest multiple homologs when the different accession numbers are actually for the same gene. Therefore, some apparently closely related genes may be the same gene [e.g. Hxt and Hand1 (bHLHa27)], but further experimental information is needed for confirmation. This is a limitation of the current study, however; rather than arbitrarily infer relatedness, the current genomic information was used, and all comparative analyses that suggested sequence differences were used to assign separate genes when appropriate.

An example of the use of the nomenclature and the super-tree distribution for the mouse genes is shown in Fig. 3. The mouse has 107 bHLH genes that appear distinct with a clade distribution shown in Fig. 3. This mouse bHLH super-tree demonstrates the clade relationships and presents the related nomenclature. Each species bHLH gene family is listed in Supporting Table S2. Those genes conserved between species can also be observed. Many of the genes have multiple names and due to relatedness on the phylogenetic tree were given the same number. For example, the super-tree in Fig. 3 has several gene clusters with different names, but have been given the same number (e.g. HAND1 and HXT(e), being bHLHa27) and are similar genes, while others (e.g. LYL1, LYL2(e), UNR1e, being bHLHa18) could be similar genes or splice variants. Future analysis of related genes is required to fine-tune this phylogenetic analysis. In the event they are splice variants a letter designation can be provided, versus the same gene when only one name/number should be used. The proposed nomenclature is suggested to allow a large number of EST to be assigned names and to clarify relatedness of clustered genes. Whether the nomenclature is generally used will require acceptance by the bHLH research community. The current nomenclature is not suggested to replace the current bHLH names, but instead to provide an approach to understand and discuss the functional, structural and species relationships of this large gene family.

Epithelial cell distribution of bHLH transcription factors

To further elucidate the phylogenetic analysis and investigate bHLH gene expression differences that define differentiated cells, microarray analysis for four different mouse cell types were analyzed (Table 1). Freshly isolated mouse cell types were collected and used in the microarray analysis. Only a subset of the total bHLH genes was present on the microarray chip used (i.e.

MGU74v2a), such that the genes listed do not reflect the total cohort of bHLH genes potentially expressed. All bHLH genes expressed above a signal of 75 with a statistically significant present call ($p < 0.05$) and were present on the microarray chip are listed for each of the cell types. To verify the utility of such an approach the muscle cell expression data was compared with known information about muscle cell differentiation. Myogenic genes such as Myod1 (bHLHc1) and Myog (bHLHc3) are expressed at high levels in the differentiated muscle cells (Table 1). These genes are known to play a role in the differentiation of muscle cells (Nakada et al., 2004; Ishibashi et al., 2005) which validates the microarray data presented for these genes. The bHLH genes expressed in the muscle, thymic, Schwann, and Sertoli cells are shown in Table 1. Twist2 (bHLHa39) and Mesp1 (bHLHc5) are only expressed in Sertoli cells when compared with the other three cell types. Genes such as Mxi1 (bHLHc11) and Scx (bHLHa41) are differentially expressed among the different cell types. This differential expression pattern may be an indication that these four genes could be important in the differentiation of Sertoli cells. Previously scleraxis (bHLHa41), Id1-4 (bHLHb24-b27), and TCF12 (bHLHb20) were shown to be expressed by Sertoli cells (Chaudhary et al., 1999, 2001; Muir et al., 2005), which validates the microarray data for these genes. Genes that may be expressed at high levels during an earlier stage of differentiation that are then turned off during the maintenance of cellular differentiation are missed in this type of analysis. Harvesting Sertoli cells during different developmental stages for analysis would be one way to obtain this information. Comparing the clade distribution of bHLH genes expressed in the Sertoli cell versus other cell types does not indicate that any one clade is over- or under-represented in the Sertoli cell when compared with the other cell types, Figure S1.

The bHLH gene unique to Schwann cells was Nmyc1 (bHLHe37), with others of interest including Hand1 (bHLHa27), Mnt (bHLHd3), Id2 (bHLHb26), and Mad4 (bHLHc12), Table 1. Previously Schwann cells have been shown to express TCF12 (bHLHb20), Id2 (bHLHb26), Id3 (bHLHb25), and Id4 (bHLHb27) (Stewart et al., 1997; Thatikunta et al., 1999), confirming the microarray data for these genes. The bHLH genes unique to thymic cells are Lmyc1 (bHLHe38) and Msc (bHLHa22), Table 1. Previously thymic cells have been shown to express Id3 (bHLHb25), Id2 (bHLHb26), and TCF12 (bHLHb20) (Blom et al., 1999; Morrow et al., 1999; Bergqvist et al., 2000; Temchura et al., 2005), which validates the microarray data for these genes. As discussed above, the bHLH genes unique to muscle cells include Myog (bHLHc3), Myf5 (bHLHc2), Arnt (bHLHe2), Myod1 (bHLHc1), and Scx (bHLHa41), Table 1. Previously muscle cells have been shown to express Myod1 (bHLHc1),

Table 1 Cell type differential bHLH gene expression

		Sertoli	Schwann	Thymic	Muscle
bHLHc11	Mxi1	611.4	117.5		
bHLHc12	Mad4	422.1	820.6	172.2	174.2
bHLHa41	Scx	361.9			834.5
bHLHb20	Tcf12	324.8	590.1	379	696.6
bHLHd1	Srebf1	298.4	475.3	162.8	666.6
bHLHb26	Id2	267.6	2,362.6	457	113.8
bHLHb12	Usf2	210.7	345.7		241.2
bHLHd13	Tcf4	162.1	143.4	181.4	144
bHLHd6	Max	152.1	221.2		94.9
bHLHe8	Clock	130.9	125.9		90.1
bHLHb39	Hes1	121.8	341.1	93.7	244.3
bHLHb11	Usf1	116.3	94.5	82.6	112.8
bHLHd3	mnt	112.5	116.9		
bHLHc5	Mesp1	112.3			
bHLHe39	Myc	103.9	153.5		
bHLHde40	BHLHb2	101.1	114.1	82.6	474.8
bHLHe5	Arntl	99.5	98.5	89.2	85.6
bHLHb25	Id3	88.5	84.4	102.7	101.5
bHLHb27	Id4	86.1	159.2	124.1	
bHLHb19	Tcf4	82.9	139.1	133	239.3
bHLHa39	Twist2	79.4			
bHLHc1	Myod1		76.6	78.1	646.1
bHLHa27	Hand1		100.5		124
BHLHb31	Hey1		339.4	95.8	103.6
bHLHe37	Nmyc1		108.9		
bHLHa22	Msc			279.7	
bHLHc3	Myog			213.6	2,660.3
bHLHe38	Lmyc1			107.4	
bHLHe2	Arnt				119.4
bHLHb21	Tcfe2a				111.5
bHLHe17	Hif3a				81.7
bHLHe42	Ncoa1				89.6
bHLHb24	Id1				132.3
bHLHc2	Myf5				139.2
bLHe15	Sim2				76.7

Values are mean raw microarray signals and the colors indicate which clade the gene is expressed in (red = clade 3, blue = clade 2, green = clade 1, orange = clade 5, and purple = clade 4). The lack of a number for a corresponding bHLH protein indicates an absence or no expression above 75 for that cell type.

myogenin (bHLHa3), Myf5 (bHLHc2), Hand1 (bHLHa27), and scleraxis (bHLHa41) (Braun et al., 1992; Morikawa and Cserjesi, 2004; Ishibashi et al., 2005; Tang et al., 2006; Pryce et al., 2007), which validates the microarray data for these genes. As observed with the Sertoli cells, no clade distribution was unique to the different cells, Table 1 and Figure S1. Although some subsets of genes did cluster to specific clades (e.g. myogenic) most clades were presented in each cell type, Figure S1. Therefore, clade distribution and functional relationships of expressed bHLH genes was not a major factor in comparing the different cell types. This observation suggests a diversity of bHLH genes with various functions is likely required for cell differentiation. The set of bHLH genes for the different cell types is likely associated with cellular differentiation. In regards to validation of the microarray data, 16 different analyses in the various cell types confirmed the expression observed, validating the array data for these

genes. Those genes listed in Table 1 not previously shown to be expressed now need to be further investigated. Not all the bHLH genes known are present on the microarray chip used, so this is a minimal subset of genes and further analysis of the entire gene family will likely reveal additional insights into the bHLH genes associated with the differentiation of these cell types. The bHLH genes identified are good initial candidates for further analysis of the cellular differentiation of these cell types.

Discussion

Phylogenetic analysis of the bHLH transcription factor family has been performed several times in the past with the first analysis in 1997 (Atchley and Fitch, 1997). In most cases only the bHLH domain was utilized in the analysis and a limited number of genes or species were represented. In the current study, Blast searches were performed of genomes to identify all potential bHLH genes for seven divergent species (i.e. mouse, rat, human, fly, worm, yeast, and plant). This analysis identified 541 total bHLH genes in all the species. Sequence alignments were performed to verify that genes were not identical in sequence before performing the phylogenetic analysis. Phylogenetic analysis was then performed utilizing the entire protein sequence rather than just the bHLH domain. This allowed the inclusion of extra domains associated with many of these proteins (e.g. zip, Orange, PAS), Fig. 1B. These additional domains play known roles in protein interactions (i.e. zip, Orange, and PAS) and may function in transcriptional repression (i.e. Orange). As expected, the majority of the bHLH proteins with similar domains clustered together in the different clades of the phylogenetic analysis, Fig. 1B. This suggests the gene groupings are appropriately related to structure and potentially similar in function.

The bHLH gene family has been previously divided into several “groups” (Atchley and Fitch, 1997; Ledent et al., 2002; Buck and Atchley, 2003; Heim et al., 2003; Toledo-Ortiz et al., 2003), yet there has been a lack of consistency in how many groups are accepted and relevant. While group A, B, and D proteins are generally classified consistently, group E and C proteins may be classified with the group B proteins. Previous classification revolves around E-box-binding specificity and bHLH domain amino-acid patterns (Atchley and Fitch, 1997). However, amino-acid patterns in the current study would indicate that previous classification of this gene family is incorrect. For example, previously identified group B proteins are spread throughout all five clades containing mammalian bHLH genes. Similarly, the ability of the genes to homodimerize or heterodimerize does not correlate to a given region of

the super-tree. Genes that can only heterodimerize are spread throughout all of the clades. E-box-binding specificity was not considered in the current study as this specificity may change with different dimerization partners. In addition, the majority of genes in the family have no known function or name. Therefore, the current study provides a system of classification that better fits the entire coding sequence and presents a phylogenetic basis of classification.

The Bayesian consensus trees suggest several interesting evolutionary patterns. First, it is clear that many of the duplication events predate vertebrate diversification (Figs. 2A–2F). Neither *Drosophila* nor *C. elegans* have as many bHLH gene family copies as mammalian species (Supporting Table S2). However, many bHLH clades are sister to *Drosophila* and/or *C. elegans* sequences suggesting these might be orthologs of the vertebrate copies and may provide candidates for studying functional diversification, Supporting Table S2. In some cases, the *Drosophila* or *C. elegans* sequence is sister to a single clade of vertebrate sequences, such as is found with the relationship among D-CG8667 (bHLHa16) and the vertebrate Mist1 (bHLHa15) sequences (Fig. 2), suggesting that this ortholog originated before the split of *Drosophila* and the vertebrates. In others, these non-vertebrate sequences are sister to a set of clades of vertebrate gene copies, as is found with D-CG5102 (bHLHb22) sister to the combination of the TCF4 (bHLHb19), TCF12 (bHLHb20), and TCF3 (bHLHb21) clades (Fig. 2B), which suggests that there were several duplication events before vertebrate diversification but after divergence from *Drosophila*. In all, it appears that approximately 89 paralogs of the bHLH gene family were present in the ancestor of human, mouse, and rat, as is noted by the blue bars on branches where these gene copies coalesce, Figs. 2A–2F.

While there are fewer bHLH gene copies in *Drosophila* and *C. elegans* with some paralogs that seem closely related to those found in vertebrates, there do seem to be some “insect specific” duplication events, particularly notable in the bHLH1, 2, and 4 clades (Figs. 2A–2C). This suggests that separate and different functions might be present for these paralogs as related copies are not found in vertebrates. Similarly, most of the Arabidopsis paralogs do not show close relationships to any of the animal sequences, and many of the duplicates form clades separate from other sequences (i.e. the bHLH clade 6; Fig. 2C). As this is the only plant included, the timing of these duplications is unclear, however, given the very few sequences with similarity to any of the animal sequences, it appears that few “orthologs” predate the divergence of plants and animals. The yeast sequences similarly do not group with sequences from any of the other lineages, not surprising given the phylogenetic distance between fungi and plants and animals. Comparatively, yeast has far fewer paralogs of the bHLH gene family than the other

species, and most of them seem quite divergent from each other suggesting very old duplications. Sequences from additional fungi will be necessary to explore the origins of these duplication events.

As complete or nearly complete sequences are present for the human, mouse, and Arabidopsis genomes, chromosome location of the bHLH paralogs can be mapped to chromosomal location to explore the issue of possible origins/mechanisms for paralogous copies. These are mapped onto the phylogenies following the sequences names (Figs. 2A–2E). For instance, in the bHLH 4 clade (Fig. 2D), the paralogs MaxA–MaxF (bHLHd4–d8) are all found on chromosome arm 14q, suggesting these were either tandem duplications or duplications to near chromosomal locations. Conversely, the human paralogs Hey1 (bHLHb29), Hey2 (bHLHb32), and HeyL (bHLHb33), all closely related, are found on chromosome arms 8q, 6q, and 1p (Fig. 2B), suggesting a very different pattern of duplication.

Several expressed sequence tags (ESTs) have been included in these analyses due to the lack of a characterized copy identical to the EST copy. The placement of these ESTs follow a few basic patterns. In some cases, the EST fills in a gap where a gene copy is expected but not yet characterized. One example of this is the rat Tal2 (bHLHa19) EST (Fig. 2A). Characterized copies of Tal2 (bHLHa19) are present for mouse and human, and the uncharacterized rat EST is placed in the tree sister to the mouse copy. Because no other closely related rat copies exist, observations suggest this is the likely functional ortholog of the other vertebrate Tal2 (bHLHa19) genes. Alternatively, some ESTs are sister to functionally characterized genes from the same organism. One example is the human Scl (bHLHa17) EST sister to the human Tal1 (bHLHa17) gene (Fig. 2A). These sequences are not identical suggesting either duplication, modified splicing, or possibly mistakes in sequencing. Similarly there are fully characterized genes with several names that either contain errors in sequencing or have actual differences in their amino-acid sequence as suggested by sequence alignment. These putative paralogs need to be further analyzed in order to clarify their identity and relationships. There are a number of clades where there are not copies from human, mouse, and rat. In many cases this might be due to incomplete knowledge of the rat genome versus the human and mouse genomes. Alternatively, this may represent missed copies in mouse or human, or possibly more recent loss of those gene copies. An example of this is the presence of rat and human copies of the Nex1 (bHLHa2) (Fig. 2A), but no mouse copy.

Several functional domains other than the bHLH domain are associated with many of these proteins. Previous focus has been on the bHLH domain, but the other domains should not be ignored as they may play a vital role in protein dimerization and/or DNA binding. The first additional domain discussed is the leucine

zipper coiled-coil dimerization domain (Zip) that is found in several families of transcription factors. Zip proteins can homodimerize or heterodimerize to bind DNA (Vinson et al., 2006). The presence of a bHLH domain along with the Zip domain may influence protein interactions and dimerization partners (Baxevanis and Vinson, 1993). In a recent study (Muir et al., 2008) it was found that bHLH and basic leucine zipper (bZip) proteins can directly interact leading to the complexity of the protein function through heterodimerization. Interestingly, the Zip domains in bHLH proteins is located in the C terminal helix domain (Fig. 1B) (Baxevanis and Vinson, 1993). The bHLH/Zip proteins did not localize to an individual clade suggesting a more general function for the Zip domain.

Similar to the bHLH/Zip proteins, the bHLH/PAS proteins have unique functions including ligand binding and protein-protein interactions. Unlike bHLH/Zip proteins, the PAS region is located a short distance C-terminal of the bHLH domain (Fig. 1B). In contrast to the bHLH/ZIP proteins, all of the bHLH/PAS proteins are located within one clade. The structure of the PAS portion of the bHLH/PAS proteins consist of a five- or six-stranded antiparallel β -sheet that is flanked by α -helices and loops (Crews and Fan, 1999). This structure forms a fold in which ligands can bind, potentially altering the conformation of the bHLH/PAS protein thereby allowing interactions with downstream signaling components (Crews and Fan, 1999). While many of the bHLH/PAS genes are conserved between mammals, *Drosophila* and *C. elegans*, the function of these genes differs. For example, the Aryl hydrocarbon receptor (Ahr) (i.e. dioxin receptor) and ARNT proteins form a bHLH protein DNA-binding complex that controls the physiological response to environmental compounds (e.g. dioxin) (Rowlands and Gustafsson, 1997) in mammals, but not in non-mammalian species (Hahn et al., 1997). The bHLH/PAS gene HIF (bHLH) has a role in cellular oxidative stress responses. The relationship of the different bHLH/PAS genes in clade 5, Fig. 3, suggests functional relationships between the different bHLH/PAS genes should be considered.

Similar to the bHLH/PAS proteins, the bHLH/Orange proteins are located within one clade and the Orange domain is located a short distance C-terminal of the bHLH domain (Davis and Turner, 2001), Fig. 1B. Interestingly, these proteins are located within clade 2 that contains the Id proteins and the bHLH/Orange proteins that function as transcriptional repressors. A direct interaction has been observed between bHLH/PAS and bHLH/Orange proteins (Chin et al., 2000). Further analysis indicates that transcriptional repression requires both the bHLH and Orange domains of these proteins. It has been speculated that the Orange domain function may be associated with subfamily specificity (Dawson et al., 1995). bHLH/Orange proteins bind to either an E-box or an N-box. Binding to an

E-box raises the possibility that bHLH/Orange proteins can compete with bHLH activator proteins, thus functioning as a repressor. The WRPW motif is present in most bHLH/Orange proteins and is able to bind to the transcriptional co-repressor groucho and its mammalian homolog the TLE proteins (Paroush et al., 1994). Groucho/TLE proteins do not bind directly to DNA, rather they are recruited to target genes by a variety of DNA bound repressors (Fisher and Caudy, 1998) and appear to function in part by recruiting histone deacetylases to repress target genes (Chen and Courey, 2000). The bHLH/Orange proteins appear to be functionally related with other bHLH repressors localized to clade 2.

This phylogenetic analysis developed a classification of the bHLH gene family that can be used to elucidate cellular differentiation considering the gene family as a whole. The current study utilized a bioinformatics approach combining the phylogenetic information with microarray analysis to determine the association of various members of the bHLH gene family in cellular differentiation. The potential role of clade distribution in relation to expressed bHLH genes was examined. To verify the utility of this approach muscle cell microarray data was analyzed for bHLH transcription factors known to play a role in muscle cell differentiation. The myogenic proteins (Myog (bHLHc3), MyoD (bHLHc1), and myf5 (bHLHc2) are all expressed in the differentiated muscle cell as was expected. Scx (bHLHa41) that has a known role in chondrocyte development during embryogenesis is also expressed at a high level in the differentiated muscle cell. Making a direct comparison among four different cell types (i.e. Sertoli, Schwann, thymic, muscle) allows for differential expression of bHLH genes to be determined. All the cell types had unique bHLH genes expressed, as well as others expressed in common. These groups of bHLH genes expressed by the various cell types provide candidates for further analysis of the role of bHLH genes in cellular differentiation. For example, four genes that may play a role in pubertal differentiation of Sertoli cells were identified, Mxi1 (bHLHc11), Srebf1 (bHLHd1), Scx (bHLHa41), and Id4 (bHLHb27). Future studies involving knockdown and over-expression of these genes will help elucidate their respective roles in Sertoli cell differentiation. A limitation to this microarray analysis is that the chip used only provides a subset of bHLH genes within the total bHLH family. In addition, only a single adult developmental time-point was used in the analysis. Many bHLH genes expressed during development and involved in the induction of cellular differentiation may be absent from this bHLH gene set. An example is the role that Tcf2a (bHLHb21) has in thymic cell differentiation (Jones and Zhuang, 2007), but was absent from the bHLH gene set, Table 1. Therefore, further developmental analysis using the entire bHLH family is needed in the future. In relation to

the current study, the bHLH gene sets associated with the different cell types demonstrated the bHLH genes expressed represented multiple clades, Figure S1 and Table 1. Cellular differentiation appears to require a mixture of different functionally related bHLH genes present throughout the phylogenetic tree and clades.

The current study uses phylogenetic analysis in concert with microarray data to classify the bHLH gene family and identify genes that may be important in the process of cellular differentiation. Investigation of the entire gene family significantly enhances the ability to study the role of bHLH genes in cellular differentiation and development. Observations provided suggest a consideration that the group or network of bHLH genes present may be more physiologically relevant and important than the role of an individual gene. Owing to the differential functions and ability to heterodimerize, understanding the interplay of the network of bHLH genes is critical to investigating the role of the bHLH family in cellular differentiation. The phylogenetic bHLH super-tree presented and bHLH gene relationships identified will be critical for future investigations.

Acknowledgments We acknowledge the technical assistance of Dr. Ingrid Sadler-Riggleman and Dr. Marina Savenkova. We thank John R. Clark for help with design and preparation of Fig. 3 and thank Ms. Jill Griffin for assistance in preparing the manuscript. None of the authors have any financial interests or disclosures. This work was supported by NIH grants to Michael K. Skinner.

References

- Anderson, M.S., Venanzi, E.S., Klein, L., Chen, Z., Berzins, S.P., Turley, S.J., von Boehmer, H., Bronson, R., Dierich, A., Benoist, C. and Mathis, D. (2002) Projection of an immunological self shadow within the thymus by the aire protein. *Science* 298:1395–1401.
- Atchley, W.R. and Fitch, W.M. (1997) A natural classification of the basic helix-loop-helix class of transcription factors. *Proc Natl Acad Sci USA* 94:5172–5176.
- Baxevasis, A.D. and Vinson, C.R. (1993) Interactions of coiled coils in transcription factors: where is the specificity? *Curr Opin Genet Dev* 3:278–285.
- Bergqvist, I., Eriksson, M., Saarikettu, J., Eriksson, B., Corneliusen, B., Grundstrom, T. and Holmberg, D. (2000) The basic helix-loop-helix transcription factor E2-2 is involved in T lymphocyte development. *Eur J Immunol* 30:2857–2863.
- Blom, B., Heemskerk, M.H., Verschuren, M.C., van Dongen, J.J., Stegmann, A.P., Bakker, A.Q., Couwenberg, F., Res, P.C. and Spits, H. (1999) Disruption of alpha beta but not of gamma delta T cell development by overexpression of the helix-loop-helix protein Id3 in committed T cell progenitors. *EMBO J* 18:2793–2802.
- Braun, T., Bober, E. and Arnold, H.H. (1992) Inhibition of muscle differentiation by the adenovirus E1a protein: repression of the transcriptional activating function of the HLH protein Myf-5. *Genes Dev* 6:888–902.
- Buchstaller, J., Sommer, L., Bodmer, M., Hoffmann, R., Suter, U. and Mantei, N. (2004) Efficient isolation and gene expression profiling of small numbers of neural crest stem cells and developing Schwann cells. *J Neurosci* 24:2357–2365.
- Buck, M.J. and Atchley, W.R. (2003) Phylogenetic analysis of plant basic helix-loop-helix proteins. *J Mol Evol* 56:742–750.
- Castella, P., Sawai, S., Nakao, K., Wagner, J.A. and Caudy, M. (2000) HES-1 repression of differentiation and proliferation in PC12 cells: role for the helix 3-helix 4 domain in transcription repression. *Mol Cell Biol* 20:6170–6183.
- Chaudhary, J., Johnson, J., Kim, G. and Skinner, M.K. (2001) Hormonal regulation and differential actions of the helix-loop-helix transcriptional inhibitors of differentiation (Id1, Id2, Id3, and Id4) in Sertoli cells. *Endocrinology* 142:1727–1736.
- Chaudhary, J., Kim, G. and Skinner, M.K. (1999) Expression of the basic helix-loop-helix protein REBalpha in rat testicular Sertoli cells. *Biol Reprod* 60:1244–1250.
- Chen, G. and Courey, A.J. (2000) Groucho/TLE family proteins and transcriptional repression. *Gene* 249:1–16.
- Chin, M.T., Maemura, K., Fukumoto, S., Jain, M.K., Layne, M.D., Watanabe, M., Hsieh, C.M. and Lee, M.E. (2000) Cardiovascular basic helix loop helix factor 1, a novel transcriptional repressor expressed preferentially in the developing and adult cardiovascular system. *J Biol Chem* 275:6381–6387.
- Crews, S.T. (1998) Control of cell lineage-specific development and transcription by bHLH-PAS proteins. *Genes Dev* 12:607–620.
- Crews, S.T. and Fan, C.M. (1999) Remembrance of things PAS: regulation of development by bHLH-PAS proteins. *Curr Opin Genet Dev* 9:580–587.
- Crozatier, M., Valle, D., Dubois, L., Ibsouda, S. and Vincent, A. (1996) Collier, a novel regulator of *Drosophila* head development, is expressed in a single mitotic domain. *Curr Biol* 6:707–718.
- Davis, R.L. and Turner, D.L. (2001) Vertebrate hairy and enhancer of split related proteins: transcriptional repressors regulating cellular differentiation and embryonic patterning. *Oncogene* 20:8342–8357.
- Dawson, S.R., Turner, D.L., Weintraub, H. and Parkhurst, S.M. (1995) Specificity for the hairy/enhancer of split basic helix-loop-helix (bHLH) proteins maps outside the bHLH domain and suggests two separable modes of transcriptional repression. *Mol Cell Biol* 15:6923–6931.
- Facchini, L.M. and Penn, L.Z. (1998) The molecular role of Myc in growth and transformation: recent discoveries lead to new insights. *FASEB J* 12:633–651.
- Fisher, A. and Caudy, M. (1998) The function of hairy-related bHLH repressor proteins in cell fate decisions. *Bioessays* 20:298–306.
- Giagtzoglou, N., Alifragis, P., Koumbanakis, K.A. and Delidakis, C. (2003) Two modes of recruitment of E(spl) repressors onto target genes. *Development* 130:259–270.
- Goding, C.R. (2000) Mitf from neural crest to melanoma: signal transduction and transcription in the melanocyte lineage. *Genes Dev* 14:1712–1728.
- Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
- Hahn, M.E., Karchner, S.I., Shapiro, M.A. and Perera, S.A. (1997) Molecular evolution of two vertebrate aryl hydrocarbon (dioxin) receptors (AHR1 and AHR2) and the PAS family. *Proc Natl Acad Sci USA* 94:13743–13748.
- Hassan, B.A. and Bellen, H.J. (2000) Doing the MATH: is the mouse a good model for fly development? *Genes Dev* 14:1852–1865.
- Heim, M.A., Jakoby, M., Werber, M., Martin, C., Weisshaar, B. and Bailey, P.C. (2003) The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Mol Biol Evol* 20:735–747.
- Henriksson, M. and Luscher, B. (1996) Proteins of the Myc network: essential regulators of cell growth and differentiation. *Adv Cancer Res* 68:109–182.
- Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

- Ishibashi, J., Perry, R.L., Asakura, A. and Rudnicki, M.A. (2005) MyoD induces myogenic differentiation through cooperation of its NH₂- and COOH-terminal regions. *J Cell Biol* 171:471–482.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282.
- Jones, M.E. and Zhuang, Y. (2007) Acquisition of a functional T cell receptor during T lymphocyte development is enforced by HEB and E2A transcription factors. *Immunity* 27:860–870.
- Kageyama, R., Ohtsuka, T., Hatakeyama, J. and Ohsawa, R. (2005) Roles of bHLH genes in neural stem cell differentiation. *Exp Cell Res* 306:343–348.
- Ledent, V., Paquet, O. and Vervoort, M. (2002) Phylogenetic analysis of the human basic helix-loop-helix proteins. *Genome Biol* 3:RESEARCH0030.
- Lewis, P.O. (2001) Phylogenetic systematics turns over a new leaf. *Trends Ecol Evolut* 16:30–37.
- Li, X., Duan, X., Jiang, H., Sun, Y., Tang, Y., Yuan, Z., Guo, J., Liang, W., Chen, L., Yin, J., Ma, H., Wang, J. and Zhang, D. (2006) Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. *Plant Physiol* 141:1167–1184.
- McLean, D.J., Friel, P.J., Pouchnik, D. and Griswold, M.D. (2002) Oligonucleotide microarray analysis of gene expression in follicle-stimulating hormone-treated rat Sertoli cells. *Mol Endocrinol* 16:2780–2792.
- Morikawa, Y. and Cserjesi, P. (2004) Extra-embryonic vasculature development is regulated by the transcription factor HAND1. *Development* 131:2195–2204.
- Morrow, M.A., Mayer, E.W., Perez, C.A., Adlam, M. and Siu, G. (1999) Overexpression of the helix-loop-helix protein Id2 blocks T cell development at multiple stages. *Mol Immunol* 36:491–503.
- Muir, T., Sadler-Riggleman, I. and Skinner, M.K. (2005) Role of the basic helix-loop-helix transcription factor, scleraxis, in the regulation of Sertoli cell function and differentiation. *Mol Endocrinol* 19:2164–2174.
- Muir, T., Wilson-Rawls, J., Stevens, J.D., Rawls, A., Kang, C. and Skinner, M.K. (2008) Integration of CREB and bHLH transcriptional signaling pathways through direct heterodimerization of proteins. *Mol Reprod Develop*, doi 10.1002/mrd.20802.
- Murre, C., McCaw, P.S. and Baltimore, D. (1989) A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. *Cell* 56:777–783.
- Nakada, Y., Hunsaker, T.L., Henke, R.M. and Johnson, J.E. (2004) Distinct domains within Mash1 and Math1 are required for function in neuronal differentiation versus neuronal cell-type specification. *Development* 131:1319–1330.
- Nef, S., Schaad, O., Stallings, N.R., Cederroth, C.R., Pitetti, J.L., Schaer, G., Malki, S., Dubois-Dauphin, M., Boizet-Bonhoure, B., Descombes, P., Parker, K.L. and Vassalli, J.D. (2005) Gene expression during sex determination reveals a robust female genetic program at the onset of ovarian development. *Dev Biol* 287:361–377.
- Paroush, Z., Finley, R.L. Jr, Kidd, T., Wainwright, S.M., Ingham, P.W., Brent, R. and Ish-Horowicz, D. (1994) Groucho is required for *Drosophila neurogenesis*, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. *Cell* 79:805–815.
- Pryce, B.A., Brent, A.E., Murchison, N.D., Tabin, C.J. and Schweitzer, R. (2007) Generation of transgenic tendon reporters, ScxGFP and ScxAP, using regulatory elements of the scleraxis gene. *Dev Dyn* 236:1677–1682.
- Rowlands, J.C. and Gustafsson, J.A. (1997) Aryl hydrocarbon receptor-mediated signal transduction. *Crit Rev Toxicol* 27:109–134.
- Skinner, M.K. (1991) Cell-cell interactions in the testis. *Endocr Rev* 12:45–77.
- Small, C.L., Shima, J.E., Uzumcu, M., Skinner, M.K. and Griswold, M.D. (2005) Profiling gene expression during the differentiation and development of the murine embryonic gonad. *Biol Reprod* 72:492–501.
- Stewart, H.J., Zoidl, G., Rossner, M., Brennan, A., Zoidl, C., Nave, K.A., Mirsky, R. and Jessen, K.R. (1997) Helix-loop-helix proteins in Schwann cells: a study of regulation and subcellular localization of Ids, REB, and E12/47 during embryonic and postnatal development. *J Neurosci Res* 50:684–701.
- Taelman, V., Van Wayenbergh, R., Solter, M., Pichon, B., Pieler, T., Christophe, D. and Bellefroid, E.J. (2004) Sequences downstream of the bHLH domain of the *Xenopus* hairy-related transcription factor-1 act as an extended dimerization domain that contributes to the selection of the partners. *Dev Biol* 276:47–63.
- Tam, S.K., Gu, W., Mahdavi, V. and Nadal-Ginard, B. (1995) Cardiac myocyte terminal differentiation. Potential for cardiac regeneration. *Ann NY Acad Sci* 752:72–79.
- Tang, H., Veldman, M.B. and Goldman, D. (2006) Characterization of a muscle-specific enhancer in human MuSK promoter reveals the essential role of myogenin in controlling activity-dependent gene regulation. *J Biol Chem* 281:3943–3953.
- Taylor, B.L. and Zhulin, I.B. (1999) PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol Mol Biol Rev* 63:479–506.
- Temchura, V.V., Frericks, M., Nacken, W. and Esser, C. (2005) Role of the aryl hydrocarbon receptor in thymocyte emigration in vivo. *Eur J Immunol* 35:2738–2747.
- Thatikunta, P., Qin, W., Christy, B.A., Tennekoon, G.I. and Rutkowski, J.L. (1999) Reciprocal Id expression and myelin gene regulation in Schwann cells. *Mol Cell Neurosci* 14:519–528.
- Thattaliyath, B.D., Firulli, B.A. and Firulli, A.B. (2002) The basic-helix-loop-helix transcription factor HAND2 directly regulates transcription of the atrial natriuretic peptide gene. *J Mol Cell Cardiol* 34:1335–1344.
- Thompson, J.A., Gold, P.J. and Fefer, A. (1997) Outpatient chemoimmunotherapy for the treatment of metastatic melanoma. *Semin Oncol* 24:S44–S48.
- Toledo-Ortiz, G., Huq, E. and Quail, P.H. (2003) The Arabidopsis basic/helix-loop-helix transcription factor family. *Plant Cell* 15:1749–1770.
- van Cruchten, I., Cinato, E., Fox, M., King, E.R., Newton, J.S., Riechmann, V. and Sablitzky, F. (1998) Structure, chromosomal localisation and expression of the murine dominant negative helix-loop-helix Id4 gene. *Biochim Biophys Acta* 1443:55–64.
- Vinson, C., Acharya, A. and Taparowsky, E.J. (2006) Deciphering B-ZIP transcription factor interactions in vitro and in vivo. *Biochim Biophys Acta* 1759:4–12.
- Wei, Q. and Paterson, B.M. (2001) Regulation of MyoD function in the dividing myoblast. *FEBS Lett* 490:171–178.
- Yang, C., Boucher, F., Tremblay, A. and Michaud, J.L. (2004) Regulatory interaction between arylhydrocarbon receptor and SIM1, two basic helix-loop-helix PAS proteins involved in the control of food intake. *J Biol Chem* 279:9306–9312.
- Yoshikawa, K. (2000) Cell cycle regulators in neural stem cells and postmitotic neurons. *Neurosci Res* 37:1–14.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. bHLH gene family nomenclature.

Table S2. Species distribution of bHLH homolog.

Fig. S1. Mouse bHLH super-tree for the 5 different clades and all 107 genes with the cellular expression of specific genes listed for Sertoli, Schwann, thymic and

muscle cells. The expression is shown in the pie chart diagrams for each gene.

Please note: Blackwell Publishing are not responsible for the content or functionality of

any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Supplemental Table S1
bHLH Gene Family Nomenclature

New name	Current name	GenBank #	Subtree
bHLHa1	H-NEUROD2	NP006151	1
bHLHa1	H-NEURODRF	BAA77569	1
bHLHa1	M-NEUROD2	Q62414	1
bHLHa1	R-NEUROD2	NP062199	1
bHLHa2	H-MYO51	AAG43167	1
bHLHa2	H-NEX1	AAP97259	1
bHLHa2	H-NEUROD6	NP073565	1
bHLHa2	R-NEX1	XP575489	1
bHLHa3	H-BETA2	BAA36519	1
bHLHa3	H-NEUROD1	NP002491	1
bHLHa3	M-NEUROD1	NP035024	1
bHLHa3	R-NEUROD1	NP062091	1
bHLHa3	M-UPP	BAC25725	1
bHLHa3	M-NEUROD(e)	AAC52203	1
bHLHa3	H-NEUROD(e)	AAC51318	1
bHLHa4	H-NEUROD4	NP067014	1
bHLHa4	M-NEUROD4	NP031527	1
bHLHa4	R-ATOH3H	XP222370	1
bHLHa5	C-C34E107	NP498115	1
bHLHa6	H-NEUROD3(e)	AAB37575	1
bHLHa6	H-NEUROG1	NP006152	1
bHLHa6	M-NEUROG1	NP035026	1
bHLHa6	R-NEUROD3	NP062080	1
bHLHa7	H-NEUROG3	NP066279	1
bHLHa7	H-UNK3	AAH74939	1
bHLHa7	M-NEUROG3	NP033849	1
bHLHa7	R-NEUROG3	NP067732	1
bHLHa8	H-NEUROG2	NP076924	1
bHLHa8	M-NEUROG2	NP033848	1
bHLHa8	R-UNK3(e)	XP227716	1
bHLHa9	C-C44C108	NP509952	1
bHLHa9	D-CG5005	NP477302	1
bHLHa9	C-T24B86	NP495938	1
bHLHa9	C-T05G52	NP499150	1
bHLHa9	M-UNK2(e)	CAI25286	1
bHLHa10	D-CG10393	NP477446	1
bHLHa10	C-C28C128	NP501445	1
bHLHa10	C-T14F95	NP508410	1
bHLHa10	D-CG7508	NP731223	1
bHLHa11	C-C02B84	NP509367	1
bHLHa11	C-Y69E1A3	NP500236	1

bHLHa12	D-CG7760	NP477344	1
bHLHa13	H-ATOH7	NP660161	1
bHLHa13	M-ATOH7	NP058560	1
bHLHa13	R-MATH5(e)	XP345103	1
bHLHa14	H-ATOH1	NP005163	1
bHLHa14	M-ATOH1	NP031526	1
bHLHa15	H-MIST1	NP803238	1
bHLHa15	M-MIST1	NP034930	1
bHLHa15	R-MIST1	NP036995	1
bHLHa16	D-CG8667	NP523611	1
bHLHa17	H-SCL(e)	AAA36598	1
bHLHa17	H-TAL1	NP003180	1
bHLHa17	M-SCL(e)	CAB72256	1
bHLHa17	M-TAL1	NP035657	1
bHLHa17	R-SCL(e)	XP233430	1
bHLHa18	H-UNK1(e)	AAA88084	1
bHLHa18	H-LYL1	NP005574	1
bHLHa18	M-LYL2(e)	B43814	1
bHLHa18	M-UNK1(e)	BAC36194	1
bHLHa18	M-LYL1	NP032561	1
bHLHa19	H-TAL2	NP005412	1
bHLHa19	R-TAL2(e)	XP345532	1
bHLHa19	M-TAL2	NP033343	1
bHLHa20	D-CG2655	NP525055	1
bHLHa21	H-ATOH8	NP116216	1
bHLHa21	M-ATOH8	NP722473	1
bHLHa21	R-MATH6	XP575552	1
bHLHa21	M-UNK3(e)	BAC25499	1
bHLHa22	H-MUSC(e)	AAH06313	1
bHLHa22	H-MUSC	NP005089	1
bHLHa22	M-MUSC	NP034957	1
bHLHa22	R-MUSC	XP232587	1
bHLHa23	H-TCF21	NP938206	1
bHLHa23	M-TCF21	NP035675	1
bHLHa23	R-COR1(e)	XP341738	1
bHLHa24	H-TCF23	XP372923	1
bHLHa24	M-TCF23	NP444315	1
bHLHa24	R-OUT(e)	XP345644	1
bHLHa25	H-UNK2(e)	DAA00306	1
bHLHa25	R-UNK2	XP578439	1
bHLHa26	H-HAND2	NP068808	1
bHLHa26	R-HAND2	NP073187	1
bHLHa26	M-HAND2	NP034532	1
bHLHa26	M-DHAND(e)	AAC52338	1
bHLHa26	M-HED(e)	AAA86274	1

bHLHa27	H-HAND1	NP004812	1
bHLHa27	M-HXT(e)	AAA86273	1
bHLHa27	M-HAND1	NP032239	1
bHLHa27	R-HAND1	NP067603	1
bHLHa28	D-CG18144	NP609370	1
bHLHa29	H-PTF1	NP835455	1
bHLHa29	M-PTF1A	NP061279	1
bHLHa29	R-PTF1A	NP446416	1
bHLHa30	D-CG33323	NP996177	1
bHLHa31	M-FERD3I	NP277057	1
bHLHa31	R-FER3(e)	XP345659	1
bHLHa32	D-CG6913	NP524322	1
bHLHa33	C-F48D63	NP508725	1
bHLHa33	D-CG5952	NP524376	1
bHLHa34	H-NHLH2	NP005590	1
bHLHa34	M-HEN2	Q64221	1
bHLHa34	M-NHLH2	NP848892	1
bHLHa34	R-NESC(e)	XP227524	1
bHLHa35	H-NHLH1	NP005589	1
bHLHa35	M-NHLH1	NP035046	1
bHLHa35	R-UNK1(e)	XP222898	1
bHLHa36	D-CG3052	NP476989	1
bHLHa37	C-C43H68	NP508440	1
bHLHa38	H-TWIST	NP000465	1
bHLHa38	M-TWIST1	NP035788	1
bHLHa38	R-TWIST1	NP445982	1
bHLHa39	H-TWIST2(e)	AAH33168	1
bHLHa39	M-TWIST2	NP031881	1
bHLHa40	H-TCF15	NP004600	1
bHLHa40	M-TCF15	NP033354	1
bHLHa40	R-TCF15(e)	XP230722	1
bHLHa41	H-SCX	DAA00239	1
bHLHa41	M-SCX	NP942588	1
bHLHa42	H-ASCL3	NP065697	1
bHLHa42	M-ASCL3	NP064435	1
bHLHa42	R-PUTBHLH(e)	BAB83912	1
bHLHa43	C-F57C123	NP508119	1
bHLHa44	H-ASCL4	NP982260	1
bHLHa44	R-ASCL4(e)	XP235013	1
bHLHa45	R-ASCL2	NP113691	1
bHLHa45	H-ACSL2	NP005161	1
bHLHa45	M-ASCL2	NP032580	1
bHLHa46	H-ASCL1	AAH31299	1
bHLHa46	R-ASCL1	NP071779	1
bHLHa46	M-ACSL1	NP032579	1

bHLHb1	A-BHLH13	NP851021	2
bHLHb1	A-PIF3-1	BAC56978	2
bHLHb1	A-PUT2	AAN15368	2
bHLHb1	A-PUT1	NP191465	2
bHLHb1	A-PTF-1	AAM10954	2
bHLHb2	A-PIF4-2	NP565991	2
bHLHb2	A-PIF4-1	NP850381	2
bHLHb2	A-UNK2	AAK25931	2
bHLHb3	A-BHLH12	NP176820	2
bHLHb3	A-BHLH11	NP198596	2
bHLHb4	A-BHLH10	NP564293	2
bHLHb4	A-BHLH9	NP680385	2
bHLHb5	A-BHLH8	NP174355	2
bHLHb5	A-BHLH7	NP181028	2
bHLHb6	A-BHLH6	NP190602	2
bHLHb6	A-BHLH5	NP201507	2
bHLHb7	A-EXP2	NP566567	2
bHLHb7	A-UNK1	AAG51338	2
bHLHb8	A-BHLH4	NP564782	2
bHLHb8	A-F17	AAF98179	2
bHLHb9	A-ERPR2	NP566639	2
bHLHb10	A-BHLH3	NP192426	2
bHLHb10	A-ERPR1	NP193865	2
bHLHb11	H-USF1	NP009053	2
bHLHb11	M-USF1	NP033506	2
bHLHb11	R-USF1	NP113965	2
bHLHb12	H-USF2	NP003358	2
bHLHb12	R-USF2	NP112401	2
bHLHb13	D-CG17592A	NP572167	2
bHLHb13	D-CG17592B	NP726931	2
bHLHb14	A-BHLH2	NP188848	2
bHLHb14	C-T01D32	NP506391	2
bHLHb15	S-CBF1	NP012594	2
bHLHb16	S-TYE7	NP014989	2
bHLHb17	D-CG2956	NP523816	2
bHLHb18	C-T15H93	NP496070	2
bHLHb19	H-SEF21A(e)	AAA60310	2
bHLHb19	H-TCF4	NP003190	2
bHLHb19	R-MITF2(e)	XP222520	2
bHLHb19	R-TCF4	NP445821	2
bHLHb19	M-SEF2(e)	CAA62868	2
bHLHb19	M-TCF4	Q60722	2
bHLHb19	M-ME2(e)	AAB32662	2
bHLHb20	H-TCF4C(e)	NP996923	2
bHLHb20	H-TCF12	NP003196	2

bHLHb20	M-TCF12	NP035674	2
bHLHb20	R-TCF12	NP037308	2
bHLHb20	R-REB3(e)	AAB25128	2
bHLHb20	R-REBOX(e)	AAB25129	2
bHLHb21	H-E47(e)	AAC41693	2
bHLHb21	H-KAPPAE2BF(e)	AAA56830	2
bHLHb21	H-TCF3	NP003191	2
bHLHb21	M-TCFIN(e)	CAA35541	2
bHLHb21	R-PAN1(e)	CAA44199	2
bHLHb21	R-TCFE2A(e)	AAH78750	2
bHLHb21	M-TCFE2A	P15806	2
bHLHb21	M-E2A	BAA06218	2
bHLHb21	R-TCFE12(e)	I78853	2
bHLHb22	D-CG5102	P11420	2
bHLHb23	D-CG5545	NP523592	2
bHLHb24	H-ID1A	NP002156	2
bHLHb24	H-ID1B	NP851998	2
bHLHb24	R-ID1	NP036929	2
bHLHb24	M-ID1	NP034625	2
bHLHb25	H-ID3	NP002158	2
bHLHb25	M-ID3	NP032347	2
bHLHb25	R-ID3	AAH64658	2
bHLHb26	H-ID2	NP002157	2
bHLHb26	M-ID2	NP034626	2
bHLHb26	R-ID2	NP037192	2
bHLHb27	H-ID4	NP001537	2
bHLHb27	M-ID4	NP112443	2
bHLHb27	R-ID4	NP783172	2
bHLHb28	D-CG1007	NP523876	2
bHLHb29	A-BHLH1	NP177590	2
bHLHb30	A-EXP1	NP566098	2
bHLHb31	H-HEY1	NP036390	2
bHLHb31	M-HEY1	AAH86635	2
bHLHb32	H-HEY2	NP036391	2
bHLHb32	M-HEY2	NP038932	2
bHLHb33	H-HEYL	NP055386	2
bHLHb33	M-HEYL	NP038933	2
bHLHb34	D-CG14548	NP524505	2
bHLHb34	D-CG8333	NP524504	2
bHLHb35	D-CG8361	NP536753	2
bHLHb36	D-CG6096	NP524511	2
bHLHb36	D-CG8365	NP524513	2
bHLHb37	M-HES3	NP032263	2
bHLHb37	R-HES3	NP073178	2
bHLHb37	H-HES7	NP115969	2

bHLHb37	M-HES7	NP149030	2
bHLHb38	M-HES5	NP034549	2
bHLHb38	R-HES5	NP077359	2
bHLHb39	H-HES1	NP005515	2
bHLHb39	M-HES1	NP032261	2
bHLHb39	R-HES1	NP077336	2
bHLHb39	H-HES4	NP066993	2
bHLHb40	M-HES2	NP032262	2
bHLHb40	R-HES2	NP062109	2
bHLHc1	H-MYF3(e)	CAA35640	3
bHLHc1	H-MYOD	NP002469	3
bHLHc1	M-MYOD1	P10085	3
bHLHc1	R-MYOD1	NP788268	3
bHLHc2	H-MYF5	NP005584	3
bHLHc2	M-MYF5	NP032682	3
bHLHc2	R-MYF5(e)	XP235101	3
bHLHc3	H-MYF4(e)	AAP35897	3
bHLHc3	M-MYOG	AAH48683	3
bHLHc3	R-MYOG	NP058811	3
bHLHc4	H-MYF6	NP002460	3
bHLHc4	M-MYF6	NP032683	3
bHLHc4	R-MYF6	NP037304	3
bHLHc5	H-MP1	NP061140	3
bHLHc5	M-MESP1	NP032614	3
bHLHc5	R-MESP1	XP218826	3
bHLHc6	M-MESP2	NP032615	3
bHLHc6	R-MESP2(e)	XP218825	3
bHLHc7	A-BHLH46	NP195114	3
bHLHc7	D-CG12952A	NP524287	3
bHLHc7	D-CG12952B	NP731326	3
bHLHc8	H-FGA	NP001004311	3
bHLHc8	M-FIGLA	NP036143	3
bHLHc9	D-CG7659	NP524124	3
bHLHc10	D-CG441	NP524516	3
bHLHc11	H-MXI1a	NP005953	3
bHLHc11	H-MXI1b	NP569157	3
bHLHc11	R-MXI1	NP037292	3
bHLHc11	M-MXI1	NP034977	3
bHLHc11	M-MAD	P50538	3
bHLHc12	H-MAD4	NP006445	3
bHLHc12	M-MAD4	NP034883	3
bHLHc12	M-MXD4	Q60948	3
bHLHc13	H-MXD3	NP112590	3
bHLHc13	M-MXD3	NP057871	3
bHLHc13	R-MXD3	NP665716	3

bHLHc14	D-CG6494	NP523997	3
bHLHc14	D-CG8328	NP524503	3
bHLHc14	D-CG8346	NP524509	3
bHLHc15	C-R03E91	NP509136	3
bHLHc16	A-BHLH45	NP181657	3
bHLHc16	A-PTF-3	AAM10940	3
bHLHc16	A-PTF-2	AAM10941	3
bHLHc17	A-PAD	AAC34218	3
bHLHc17	A-BHLH44	NP199992	3
bHLHc18	A-FO	AAF88076	3
bHLHc19	A-BHLH43	NP181646	3
bHLHc19	A-BHLH42	NP191236	3
bHLHc19	A-BHLH41	NP189199	3
bHLHc19	A-BHLH40	NP564944	3
bHLHc20	A-BHLH39	NP181549	3
bHLHc21	D-CG5927	NP525094	3
bHLHc22	C-T01E82	NP496204	3
bHLHc23	H-HES6	NP061115	3
bHLHc24	H-HES2	CAB46199	3
bHLHc25	C-Y54G2A1	NP500281	3
bHLHc26	A-BHLH37	NP563749	3
bHLHc26	A-BHLH38	NP181843	3
bHLHc26	A-BHLH36	NP181757	3
bHLHc27	A-UNK4	NP683482	3
bHLHc28	C-C18A38	NP495131	3
bHLHc28	D-CG3786	NP476694	3
bHLHc28	D-CG3839	NP476623	3
bHLHc28	D-CG3827	NP476803	3
bHLHc29	A-BHLH34	NP179861	3
bHLHc29	A-BHLH35	NP179860	3
bHLHc30	A-BHLH33	NP569014	3
bHLHc31	A-BHLH32	NP177064	3
bHLHc31	A-BHLH31	NP198702	3
bHLHc32	A-BHLH30	NP175518	3
bHLHc32	A-BHLH29	NP567431	3
bHLHc33	A-BHLH28	NP568850	3
bHLHc33	A-BHLH27	NP201335	3
bHLHc34	C-DY33	NP492372	3
bHLHc34	C-F38C22	NP502928	3
bHLHc35	A-BHLH26	NP175399	3
bHLHc35	A-BHLH25	NP199178	3
bHLHc36	A-EXP7	NP201239	3
bHLHc36	A-EXP6	NP199812	3
bHLHc37	S-INO2	NP010408	3
bHLHc37	S-PHO4	NP116692	3

bHLHc38	A-TIF	AAK96799	3
bHLHc38	A-HYP1	CAB71865	3
bHLHc39	A-BHLH24	NP196430	3
bHLHc40	A-BHLH23	NP200609	3
bHLHc41	M-TFAP4	NP112459	3
bHLHc42	D-CG11450	NP524820	3
bHLHc43	A-PP-1	CAB79132	3
bHLHc44	A-BHLH22	NP195330	3
bHLHc45	S-RTG1	NP014574	3
bHLHc46	C-F40G911	NP497173	3
bHLHc47	A-PIF3-2	BAC10690	3
bHLHc48	C-M05B55	NP492034	3
bHLHc49	A-UNK3	AAM67566	3
bHLHc50	A-P50	NP195468	3
bHLHc51	A-EXP5	NP194747	3
bHLHc52	A-EXP4	NP563839	3
bHLHc52	A-EXP3	NP566260	3
bHLHc53	A-BHLH20	NP196537	3
bHLHc53	A-BHLH21	NP191923	3
bHLHc54	A-BHLH19	NP172483	3
bHLHc54	A-BHLH18	NP189011	3
bHLHc55	A-BHLH17	NP195179	3
bHLHc56	A-BHLH16	NP568745	3
bHLHc56	A-BHLH15	NP177524	3
bHLHc57	A-BHLH14	NP201067	3
bHLHd1	H-SREBF1A	NP001005291	4
bHLHd1	H-SREBF1B	NP004167	4
bHLHd1	M-SREBF1	NP035610	4
bHLHd1	R-ADD1	A48085	4
bHLHd1	R-SREBP1C	AAG28734	4
bHLHd2	H-SREBF2	NP004590	4
bHLHd2	M-SREBF2	XP127995	4
bHLHd3	H-MNT	Q99583	4
bHLHd3	M-MNT	CAI24246	4
bHLHd4	H-MAXD	NP660089	4
bHLHd5	H-MAXF	NP932061	4
bHLHd6	H-MAXA	NP002373	4
bHLHd6	M-MAX	NP032584	4
bHLHd6	H-MAXB	NP660087	4
bHLHd7	H-MAXC	NP660088	4
bHLHd8	H-MAXE	NP660092	4
bHLHd9	D-CG9648	NP649097	4
bHLHd10	C-F46G106	NP510223	4
bHLHd11	C-F38A63	NP505856	4
bHLHd12	D-CG18362A	NP724325	4

bHLHd12	D-CG18362B	NP610105	4
bHLHd12	D-CG18362C	NP724326	4
bHLHd12	D-CG18362E	NP724327	4
bHLHd12	D-CG18362D	NP724328	4
bHLHd13	M-TCFL4	NP035680	4
bHLHd13	H-WBSCR14	NP937848	4
bHLHd13	D-CG3350	NP651556	4
bHLHe1	H-ARNT2	NP055677	5
bHLHe1	R-ARNT2	NP036913	5
bHLHe1	M-ARNT2	NP031514	5
bHLHe2	H-ARNT	CAI12797	5
bHLHe2	M-ARNT	NP033839	5
bHLHe2	R-ARNT	NP036912	5
bHLHe3	C-C25A111	NP492687	5
bHLHe4	D-CG11987	NP731308	5
bHLHe5	H-ARNTL	NP001169	5
bHLHe5	M-ARNTL	NP031515	5
bHLHe5	R-ARNTL	NP077338	5
bHLHe6	H-ARNTL2	NP064568	5
bHLHe6	M-ARNTL2	NP758513	5
bHLHe7	D-CG8727	NP524168	5
bHLHe8	H-CLOCK	O15516	5
bHLHe8	M-CLOCK	NP031741	5
bHLHe8	R-CLOCK	NP068628	5
bHLHe9	H-NPAS2	Q99743	5
bHLHe9	M-NPAS2	NP032745	5
bHLHe10	D-CG7391	NP996021	5
bHLHe11	H-NPAS1	NP002508	5
bHLHe11	M-NPAS1	NP032744	5
bHLHe12	H-NPAS3	NP775182	5
bHLHe12	M-NPAS3	NP038808	5
bHLHe13	D-CG6883	NP523872	5
bHLHe14	H-SIM1	NP005059	5
bHLHe14	M-SIM1	NP035506	5
bHLHe15	H-SIM2	NP005060	5
bHLHe15	M-SIM2	NP035507	5
bHLHe16	D-CG7771	NP731771	5
bHLHe17	H-HIF3A	NP690007	5
bHLHe17	M-HIF3A	NP058564	5
bHLHe17	R-HIF3A	NP071973	5
bHLHe17	H-EPAS1	NP001421	5
bHLHe18	C-41G75	NP492530	5
bHLHe19	H-BHLHB1(e)	AAF61215	5
bHLHe19	M-OLIG2	NP058663	5
bHLHe19	R-OLIG2(e)	XP221668	5

bHLHe19	H-OLIG2	NP005797	5
bHLHe20	H-OLIG3	CAI20348	5
bHLHe20	M-OLIG3	NP443734	5
bHLHe20	R-OLIG3(e)	XP218772	5
bHLHe21	H-OLIG1VAR(e)	BAD93028	5
bHLHe21	M-OLIG1	NP058664	5
bHLHe21	R-OLIG1	NP068538	5
bHLHe21	H-OLIG1	NP620450	5
bHLHe22	H-BHLHB5	NP689627	5
bHLHe22	M-BHLHB5	NP067535	5
bHLHe22	R-BETA2	XP345191	5
bHLHe23	H-BHLHB4	CAD37449	5
bHLHe23	M-BHLHB4	NP542372	5
bHLHe24	A-BHLH70	NP180680	5
bHLHe24	A-BHLH71	NP180679	5
bHLHe25	A-BHLH69	NP177366	5
bHLHe25	A-BHLH68	NP189056	5
bHLHe26	A-BHLH67	NP176552	5
bHLHe26	A-BHLH66	NP680372	5
bHLHe27	A-MYC	NP191957	5
bHLHe28	A-BHLH65	NP192720	5
bHLHe29	A-BHLH64	NP179283	5
bHLHe29	A-BTF	AAC64222	5
bHLHe30	A-BHLH63	NP195520	5
bHLHe31	A-BHLH60	NP199495	5
bHLHe31	A-BHLH62	NP193376	5
bHLHe31	A-BHLH61	NP567195	5
bHLHe31	A-BHLH59	NP195791	5
bHLHe32	H-MITF	I38024	5
bHLHe32	M-MITF	NP032627	5
bHLHe32	R-MITF	XP232215	5
bHLHe33	H-TCFE3	NP006512	5
bHLHe33	M-TCFE3	NP766060	5
bHLHe34	H-TCFEC	NP036384	5
bHLHe34	M-TCFEC	NP112475	5
bHLHe35	H-TFEB	NP009093	5
bHLHe35	M-TCFEB	NP035679	5
bHLHe36	H-MONDOA	NP055753	5
bHLHe36	M-WBSCR14	NP067430	5
bHLHe37	H-MYCN	P04198	5
bHLHe37	R-MYCS	NP068609	5
bHLHe37	M-NMYC1	NP032735	5
bHLHe38	H-MYCL1	P12524	5
bHLHe38	M-LMYC1	P10166	5
bHLHe39	H-MYC	NP002458	5

bHLHe39	M-MYC	P01108	5
bHLHe39	R-MYC	NP036735	5
bHLHe40	H-BHLHB2	NP003661	5
bHLHe40	M-BHLHB2	NP035628	5
bHLHe40	R-BHLHB2	NP445780	5
bHLHe41	R-BHLHB3	NP579837	5
bHLHe41	M-BHLHB3	NP077789	5
bHLHe41	H-BHLHB3	NP110389	5
bHLHe42	H-NCOA11	NP003734	5
bHLHe42	H-NCOA13	NP671766	5
bHLHe42	H-NCOA12	NP671756	5
bHLHe42	M-NCOA1	NP035011	5
bHLHe42	H-NCOA3	CAI42141	5
bHLHe43	D-CG13316A	NP570071	5
bHLHe43	D-CG13316B	NP726879	5
bHLHe44	D-CG1194	NP523657	5
bHLHe45	A-ERFP	NP194639	5
bHLHe46	C-C17C31	NP495065	5
bHLHe46	C-C17C37	NP495063	5
bHLHe47	C-F31A32	NP510838	5
bHLHe47	C-F31A34	NP510837	5
bHLHe48	A-BHLH58	NP200506	5
bHLHe48	A-T10024	AAD39586	5
bHLHe49	D-CG17100	NP524775	5
bHLHe50	D-CG8704	NP476923	5
bHLHe51	D-CG10446	NP523599	5
bHLHe52	A-EXP9	NP189001	5
bHLHe53	A-BHLH57	NP188770	5
bHLHe54	A-CDP	NP565839	5
bHLHe55	A-PTF-5	AAM10949	5
bHLHe56	D-CG7951	NP524584	5
bHLHe57	D-CG10798	NP525062	5
bHLHe58	A-PIF3-3	NP849626	5
bHLHe59	D-CG1705	NP511126	5
bHLHe60	A-BHLH56	NP172746	5
bHLHe61	A-ERPR4	NP174087	5
bHLHe62	A-ERPR3	NP188700	5
bHLHe63	D-CG7664	NP476605	5
bHLHe64	C-Y7D3B7	NP499472	5
bHLHe65	S-RTG3	NP009447	5
bHLHe66	S-HMS1	NP014675	5
bHLHe67	A-PUT3	CAB86924	5
bHLHe68	A-BHLH55	NP195372	5
bHLHe68	A-PTF-4	AAN18284	5
bHLHe68	A-BHLH54	NP849508	5

bHLHe69	A-EXP8	AAD15506	5
bHLHe70	A-BHLH53	NP173950	5
bHLHe70	A-BHLH52	NP177058	5
bHLHe71	A-BHLH51	NP187390	5
bHLHe71	A-BHLH50	NP199667	5
bHLHe72	A-BHLH49	NP180003	5
bHLHe72	A-BHLH48	NP194827	5
bHLHe72	A-BHLH47	NP567245	5
bHLHf1	A-PP-2	CAA22971	6
bHLHf1	A-PTF-8	CAD58595	6
bHLHf2	A-BHLH86	NP194608	6
bHLHf2	A-BHLH85	NP974634	6
bHLHf3	A-PTF-7	CAE12173	6
bHLHf4	A-BHLH84	NP179608	6
bHLHf4	A-BHLH83	NP849996	6
bHLHf5	A-AT	AAS79546	6
bHLHf5	A-BHLH82	NP200935	6
bHLHf5	A-HYP3	CAG25857	6
bHLHf5	A-UNK6	BAB08482	6
bHLHf5	A-PTF-6	AAP86213	6
bHLHf6	A-BHLH81	NP201512	6
bHLHf6	A-PBHLH2	BAD43719	6
bHLHf6	A-UNK5	AAM61195	6
bHLHf6	A-SPT	NP568010	6
bHLHf7	A-BHLH80	NP182220	6
bHLHf7	A-HYP2	AAO37214	6
bHLHf7	A-HFR1	NP563650	6
bHLHf7	A-PBHLH1	BAC41923	6
bHLHf8	A-BHLH79	NP174776	6
bHLHf8	A-BHLH78	NP192657	6
bHLHf9	A-BHLH77	NP191916	6
bHLHf10	A-BHLH76	NP182204	6
bHLHf10	A-BHLH75	NP567121	6
bHLHf10	A-BHLH73	NP187263	6
bHLHf10	A-BHLH72	NP200133	6
bHLHf10	A-BHLH74	NP192055	6

Supplemental Table S2
Species Distribution of bHLH Homologs

Names	Human	Rat	Mouse	Drosophila	C. Elegans	Yeast	Arabidopsis
bHLHa1	Neurod2, Neurodrf	Neurod2	Neurod2				
bHLHa2	Myo51, Nex1, Neurod6	Nex1					
bHLHa3	Beta2, Neurod1, Neurod(e)	Neurod1	Neurod1, Upp, Neurod(e)				
bHLHa4	Neurod4	Atoh3h	Neurod4				
bHLHa5					C34E107		
bHLHa6	Neurod3(e), Neurog1	Neurod3	Neurog1				
bHLHa7	Neurog3, Unk3	Neurog3	Neurog3				
bHLHa8	Neurog2	Unk3(e)	Neurog2				
bHLHa9			Unk2(e)	CG5005	C44C108, T24B86, T05G52		
bHLHa10				CG10393, CG7508	C28C128, T14F95		
bHLHa11					C02B84, Y69E1A3		
bHLHa12				CG7760			
bHLHa13	Atoh7	Math5(e)	Atoh7				
bHLHa14	Atoh1		Atoh1				
bHLHa15	Mist1	Mist1	Mist1				
bHLHa16				CG8667			
bHLHa17	Scl(e), Tal1	Scl(e)	Tal1, Scl(e)				
bHLHa18	Unk1(e), Lyl1		Lyl2(e), Unk1(e), Lyl1				
bHLHa19	Tal2	Tal2(e)	Tal2				
bHLHa20				CG2655			
bHLHa21	Atoh8	Math6	Atoh8, Unk3(e)				
bHLHa22	Musc(e), Musc	Musc	Musc				
bHLHa23	Tcf21	Tcf21	Cor1(e)				
bHLHa24	Tcf23	Tcf23	Out(e)				
bHLHa25	Unk2(e)	Unk2					
bHLHa26	Hand2	Hand2	Hand2, Dhand(e), Hed(e)				
bHLHa27	Hand1	Hand1	Hand1, Hxt(e)				
bHLHa28				CG18144			

bHLHa29	Ptf1	Ptf1a	Ptf1a				
bHLHa30				CG33323			
bHLHa31		Fer3(e)	Ferd3I				
bHLHa32				CG6913			
bHLHa33				CG5952	F48D63		
bHLHa34	Nhlh2	Nesc(e)	Nhlh2, Hen2				
bHLHa35	Nhlh1	Unk1(e)	Nhlh1				
bHLHa36				CG3052			
bHLHa37					C43H68		
bHLHa38	Twist	Twist1	Twist1				
bHLHa39	Twist2(e)		Twist2				
bHLHa40	Tcf15	Tcf15(e)	Tcf15				
bHLHa41	Sex		Sex				
bHLHa42	Ascl3	Putbhlh(e)	Ascl3				
bHLHa43					F57C123		
bHLHa44	Ascl4	Ascl4(e)					
bHLHa45	Ascl2	Ascl2	Ascl2				
bHLHa46	Ascl1	Ascl1	Ascl1				
bHLHb1							Bhlh13, Pif3-1, Put2, Put1, Ptf-1
bHLHb2							Pif4-2, Pif4- 1, Unk2
bHLHb3							Bhlh12, Bhlh11
bHLHb4							Bhlh10, Bhlh9
bHLHb5							Bhlh8, Bhlh7
bHLHb6							Bhlh6, Bhlh5
bHLHb7							Exp2, Unk1
bHLHb8							Bhlh4, F17
bHLHb9							Erpr2
bHLHb10							Bhlh3, Erpr1
bHLHb11	Usf1	Usf1	Usf1				
bHLHb12	Usf2	Usf2					
bHLHb13				CG17592A, CG17592B			
bHLHb14					T01D32		Bhlh2
bHLHb15						CBF1	
bHLHb16						TYE7	
bHLHb17				CG2956			
bHLHb18					T15H93		
bHLHb19	Sef21a(e), Tcf4	Mitf2(e), Tcf4	Sef2(e), Tcf4, Me2(e)				
bHLHb20	Tcf4c(e),	Tcf12,	Tcf12				

	Tcf12	Reb3(e), Rebox(e)					
bHLHb21	E47(e), Kappa2bf(e), Tcf3	Pan1(e), Tcfe2a(e), Tcfe12(e)	Tcfin(e), Tcfe2a, E2a				
bHLHb22				CG5102			
bHLHb23				CG5545			
bHLHb24	Id1a, Id1b	Id1	Id1				
bHLHb25	Id3	Id3	Id3				
bHLHb26	Id2	Id2	Id2				
bHLHb27	Id4	Id4	Id4				
bHLHb28				CG1007			
bHLHb29							Bhlh1
bHLHb30							Exp1
bHLHb31	Hey1		Hey1				
bHLHb32	Hey2		Hey2				
bHLHb33	Hey1		Hey1				
bHLHb34				CG14548, CG8333			
bHLHb35				CG8361			
bHLHb36				CG6096, CG8365			
bHLHb37	Hes7	Hes3	Hes3, Hes7				
bHLHb38		Hes5	Hes5				
bHLHb39	Hes1, Hes4	Hes1	Hes1				
bHLHb40		Hes2	Hes2				
bHLHc1	Myf3(e), Myod	Myod1	Myod1				
bHLHc2	Myf5	Myf5(e)	Myf5				
bHLHc3	Myf4(e)	Myog	Myog				
bHLHc4	Myf6	Myf6	Myf6				
bHLHc5	Mpl	Mesp1	Mesp1				
bHLHc6		Mesp2(e)	Mesp2				
bHLHc7				CG12952a, CG12952b			Bhlh46
bHLHc8	GA		Fig1a				
bHLHc9				CG7659			
bHLHc10				CG441			
bHLHc11	Mxi1a, Mxi1b	Mxi1	Mxi1, Mad				
bHLHc12	Mad4		Mad4, Mxd4				
bHLHc13	Mxd3	Mxd3	Mxd3				
bHLHc14				CG6494, CG8328, CG8346			
bHLHc15					R03E91		
bHLHc16							Bhlh45, Ptf- 3, Ptf-2

bHLHc17							Pad, Bhlh44
bHLHc18							F0
bHLHc19							Bhlh43 Bhlh42, Bhlh41, Bhlh40
bHLHc20							Bhlh39
bHLHc21				CG5927			
bHLHc22					T01E82		
bHLHc23	Hes6						
bHLHc24	Hes2						
bHLHc25					Y54G2A1		
bHLHc26							Bhlh37, Bhlh38, Bhlh36
bHLHc27							Unk4
bHLHc28				CG3786, CG3839, CG3827	C18A38		
bHLHc29							Bhlh34, Bhlh35
bHLHc30							Bhlh33
bHLHc31							Bhlh32, Bhlh31
bHLHc32							Bhlh30, Bhlh29
bHLHc33							Bhlh28, Bhlh27
bHLHc34					DY33, F38C22		
bHLHc35							Bhlh26, Bhlh25
bHLHc36							Exp7, Exp6
bHLHc37						INO2, PHO4	
bHLHc38							Tif, Hyp1
bHLHc39							Bhlh24
bHLHc40							Bhlh23
bHLHc41			Tfap4				
bHLHc42				CG11450			
bHLHc43							PP-1
bHLHc44							Bhlh22
bHLHc45						RTG1	
bHLHc46					F40G911		
bHLHc47							Pif3-2
bHLHc48					M05B55		
bHLHc49							Unk3
bHLHc50							P50
bHLHc51							Exp5

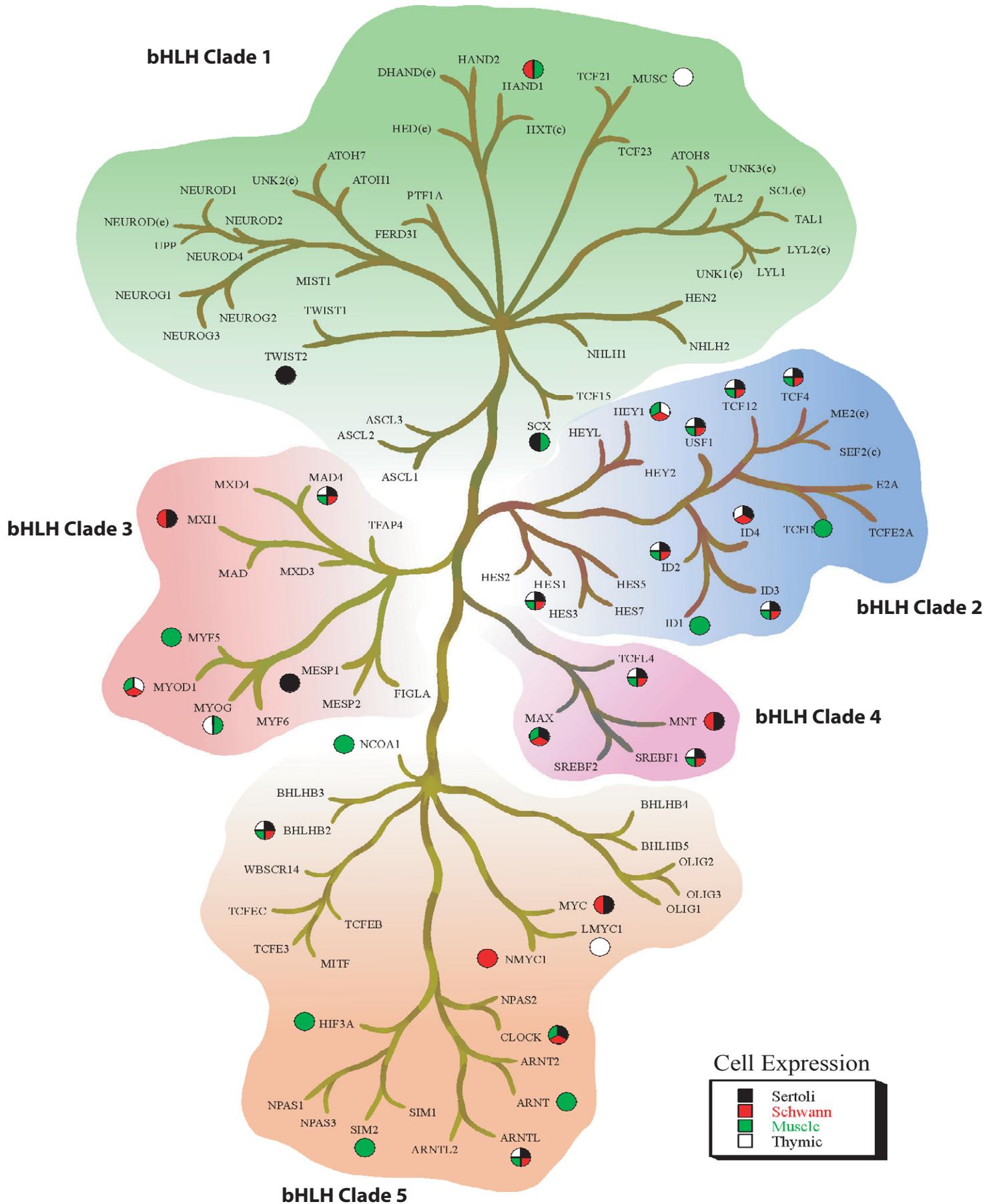
bHLHc52							Exp4, Exp3
bHLHc53							Bhlh20, Bhlh21
bHLHc54							Bhlh19, Bhlh18
bHLHc55							Bhlh17
bHLHc56							Bhlh16, Bhlh15
bHLHc57							Bhlh14
bHLHd1	Srebf1a, Srebf1b	Add1, Srebp1c	Srebf1				
bHLHd2	Srebf2		Srebf2				
bHLHd3	Mnt		Mnt				
bHLHd4	Maxd						
bHLHd5	Maxf						
bHLHd6	Maxa, Max, Maxb						
bHLHd7	Maxc						
bHLHd8	Maxe						
bHLHd9				CG9648			
bHLHd10					F46G106		
bHLHd11					F38A63		
bHLHd12				CG18362a, CG18362b, CG18362c, CG18362e, CG18362d			
bHLHd13	Wbscr14		Tefl4	CG3350			
bHLHe1	Arnt2	Arnt2	Arnt2				
bHLHe2	Arnt	Arnt	Arnt				
bHLHe3					C25A111		
bHLHe4				CG11987			
bHLHe5	Arntl	Arntl	Arntl				
bHLHe6	Arntl2		Arntl2				
bHLHe7				CG8727			
bHLHe8	Clock	Clock	Clock				
bHLHe9	Npas2		Npas2				
bHLHe10				CG7391			
bHLHe11	Npas1		Npas1				
bHLHe12	Npas3		Npas3				
bHLHe13				CG6883			
bHLHe14	Sim1		Sim1				
bHLHe15	Sim2		Sim2				
bHLHe16				CG7771			
bHLHe17	Hif3a, Epas1	Hif3a	Hif3a				
bHLHe18					41G75		
bHLHe19	Bhlh1(e), Olig2	Olig2(e)	Olig2				

bHLHe20	Olig3	Olig3(e)	Olig3				
bHLHe21	Olig1 var(e), Olig1	Olig1	Olig1				
bHLHe22	Bhlhb5	Beta2	Bhlhb5				
bHLHe23	Bhlhb4		Bhlhb4				
bHLHe24							Bhlh70, Bhlh71
bHLHe25							Bhlh69, Bhlh68
bHLHe26							Bhlh67, Bhlh66
bHLHe27							Myc
bHLHe28							Bhlh65
bHLHe29							Bhlh64, Btf
bHLHe30							Bhlh63
bHLHe31							Bhlh60, Bhlh62, Bhlh61, Bhlh59
bHLHe32	Mif	Mitf	Mitf				
bHLHe33	Tcfe3		Tcfe3				
bHLHe34	Tfec		Tfec				
bHLHe35	Tfeb		Tfeb				
bHLHe36	Mondoa		Wbscr14				
bHLHe37	Mycn	Mycs	Nmyc1				
bHLHe38	Mycl1		Lmyc1				
bHLHe39	Myc	Myc	Myc				
bHLHe40	Bhlhb2	Bhlhb2	Bhlhb2				
bHLHe41	Bhlhb3	Bhlhb3	Bhlhb3				
bHLHe42	Ncoa11, Ncoa13, Ncoa12, Ncoa3	Ncoa1					
bHLHe43				CG13316a, CG13316b			
bHLHe44				CG1194			
bHLHe45							Erfp
bHLHe46					C17C31, C17C37		
bHLHe47					F31A32, F31A34		
bHLHe48							Bhlh58, T10024
bHLHe49				CG17100			
bHLHe50				CG8704			
bHLHe51				CG10446			
bHLHe52							Exp9
bHLHe53							Bhlh57
bHLHe54							Cdp

bHLHe55						Ptf-5
bHLHe56				CG7951		
bHLHe57				CG10798		
bHLHe58						Pif3-3
bHLHe59				CG1705		
bHLHe60						Bhlh56
bHLHe61						Erpr4
bHLHe62						Erpr3
bHLHe63				CG7664		
bHLHe64					Y7D3B7	
bHLHe65						Rtg3
bHLHe66						Hms1
bHLHe67						Put3
bHLHe68						Bhlh55, Ptf-4, Bhlh54
bHLHe69						Exp8
bHLHe70						Bhlh53, Bhlh52
bHLHe71						Bhlh51, Bhlh50
bHLHe72						Bhlh49, Bhlh48, Bhlh47
bHLHf1						PP-2, Ptf-8
bHLHf2						Bhlh86, Bhlh85
bHLHf3						Ptf-7
bHLHf4						Bhlh84, Bhlh83
bHLHf5						At,Bhlh82, Hyp3, Unk6, Ptf-6
bHLHf6						Bhlh81, Pbhlh2, Unk5, Spt
bHLHf7						Bhlh80, Hyp2, Hfr1, Pbhlh1
bHLHf8						Bhlh79, Bhlh78
bHLHf9						Bhlh77
bHLHf10						Bhlh76, Bhlh75, Bhlh73, Bhlh72, Bhlh74

Multiple bHLH in a single box/cell are similar and closely related genes.

Supplemental Figure S1 (color)



Supplemental Figure S1: Mouse bHLH super-tree for the 5 different clades and all 107 genes with the cellular expression of specific genes listed for Sertoli, Schwann, thymic and muscle cells. The expression is shown in the pie chart diagrams for each gene.