The Basic Helix–Loop–Helix Transcription Factor Family in Plants: A Genome-Wide Study of Protein Structure and Functional Diversity

Marc A. Heim, Marc Jakoby,* Martin Werber,* Cathie Martin,† Bernd Weisshaar,* and Paul C. Bailey†*

*Max-Planck-Institute for Plant Breeding Research, Cologne, Germany; and †Department of Cell and Developmental Biology, John Innes Centre, Colney, Norwich, United Kingdom

Basic helix–loop–helix (bHLH) transcription factors (TFs) belong to a family of transcriptional regulators present in three eukaryotic kingdoms. Many different functions have been identified for these proteins in animals, including the control of cell proliferation and development of specific cell lineages. Their mechanism for controlling gene transcription often involves homodimerization or heterodimerization. In plants, little is known about the bHLH family, but we have determined that there are 133 *bHLH* genes in *Arabidopsis thaliana* and have confirmed that at least 113 of them are expressed. The *AtbHLH* genes constitute one of the largest families of transcription factors in *A. thaliana* with significantly more members than are found in most animal species and about an equivalent number to those in vertebrates. Comparisons with animal sequences suggest that the majority of plant *bHLH* genes have evolved from the ancestral group B class of *bHLH* genes. By studying the *AtbHLH* genes collectively, twelve subfamilies have been identified. Within each of these main groups, there are conserved amino acid sequence motifs outside the DNA binding domain. Potential gene redundancy among members of smaller subgroups has been analyzed, and the resulting information is presented to provide a simplified visual interpretation of the gene family, identifying related genes that are likely to share similar functions. Based on the current characterization of a limited number of plant bHLH proteins, we predict that this family of TFs has a range of different roles in plant cell and tissue development as well as plant metabolism.

Introduction

In 1989 Murre, McCaw, and Baltimore identified a region present in ten DNA binding proteins from animals that shared a significant number of identical amino acids. This region has become known as the basic helix–loop–helix (bHLH) domain. The proteins containing this domain have broad functions in regulating cell proliferation and cellular differentiation pathways and include human c-Myc, the myoblast determination factor MyoD, as well as Achaete and Scute from *Drosophila melanogaster* that define the initial steps in neural development. Also in 1989, Ludwig et al. reported the molecular identification of Lc, a regulator of anthocyanin biosynthesis in *Zea mays*, and showed that the predicted protein shared the bHLH domain. With the identification of the Ino4p protein from yeast (Berben et al. 1990), it became clear that bHLH proteins constitute a ubiquitous family of regulators in eukaryotes and that the bHLH domain is an ancient component of transcriptional regulation. Moreover, recent genome sequencing and expressed sequence tag (EST) programs indicated the existence of many more *bHLH* genes in various eukaryotic species.

Typically, a bHLH domain comprises a stretch of about 18 hydrophilic and basic amino acids at the N-terminal end of the domain, followed by two regions of hydrophobic residues predicted to form amphipathic α-helices separated by an intervening loop (Murre et al. 1994). Studies with mammalian bHLH proteins have shown that the conserved HLH structure is required for dimerization between two bHLH proteins (Ferré-D'Amaré et al. 1993, 1994; Ellenberger et al. 1994). The specificity for a particular partner protein resides in the α-helices, and for a sequence-specific interaction with DNA, the basic regions of both polypeptides are required. Although the bHLH domain is conserved between animals and plants, the dimeric structure of the DNA binding complex in plants has only been inferred from the conserved amino acid sequence. Some, although not all, plant bHLH proteins which have been studied so far have been shown to modulate gene expression by binding DNA.

Generally, eukaryotic transcription factors (TFs) consist of at least two discrete domains—a DNA binding domain and an activation or repression domain—that operate together to modulate the rate of transcriptional initiation from target gene promoters (Ptashne 1988). Domains other than the bHLH DNA binding domain are known to be important for the regulation of gene expression by bHLH proteins, and these regions may be conserved at the amino acid level in related proteins from different species. This has been shown for the E12 family of bHLH proteins in animals, which are a ubiquitously expressed class of animal TFs involved in the development of many cell lineages. E12 factors act via heterodimerization with other, tissue-specific bHLH proteins (Quong et al. 1993).

To date, few plant *bHLH* genes have been studied in detail, but those that have, have provided insights into the central roles of TFs in biology, and into their biochemical function. Genetic analysis of the anthocyanin biosynthetic pathway in *Z. mays* has identified a group of *bHLH* genes required for production of the purple anthocyanin pigments: *R*(*R-s* and *R-p*), *B*, *Lc*, *Sn* (Neuffer, Coe, and Wessler 1997, pp. 367–369) and *R-ch Hopi* (Petroni et al. 2000). The encoded bHLH proteins interact with members of another group of TFs in *Z. mays*—Cl or P, both R2R3-MYB proteins—and together they control pigmentation in tissues wherever the two proteins are expressed together. Other members of the *R* gene family in *Z. mays* and other

Key words: bHLH, *Arabidopsis thaliana*, transcription control, genomics.

E-mail: weisshaa@mpiz-koeln.mpg.de.

Mol. Biol. Evol. 20(5):735–747. 2003
DOI: 10.1093/molbev/msg088

species are also believed to interact with MYB proteins to regulate anthocyanin production in a tissue-specific manner. The bHLH/MYB partnership has been shown to be important in the differentiation of *A. thaliana* trichomes through the interaction of a bHLH protein encoded by *GLABRA3* (*GL3*) and the R2R3-MYB protein GLABROUS1 (GL1; Payne, Zhang, and Lloyd 2000). Phytochrome interacting factor 3 (PIF3) is a bHLH protein necessary for light signaling mediated by the photoreceptor phytochrome B (PhyB) in *A. thaliana* (Ni, Tepperman, and Quail 1998; Halliday et al. 1999). Recently it has been shown that PhyB in its active form interacts with DNA-bound PIF3, suggesting a mechanism for direct activation of photoresponsive genes by a light signal (Martinez-Garcia, Huq, and Quail 2000).

Within the *A. thaliana* genome (The Arabidopsis Genome Initiative 2000), 32 families of genes have been identified as encoding TFs that contain three or more members (Riechmann and Ratcliffe 2000). Each family is characterized by a unique region of highly conserved amino acid sequence, which usually comprises the DNA binding domain. From comparative analysis of these conserved domains, it is apparent that within each gene family there are subgroups of genes that are structurally more closely related to each other than to other members of the family. Regions of TFs outside the DNA binding domain often contain short amino acid sequence motifs or domains that are also conserved. The prediction is that proteins within a subgroup that share these smaller motifs will share similar (but not necessarily identical) functions. Where information on function is available, different members of a subgroup of genes tend to share very similar or related functions, even though the biological functions of the particular TF family as a whole may be very broad (for example the AP2 family [Riechmann and Meyerowitz 1998], the R2R3-MYB family [Stracke, Werber, and Weisshaar 2001], or the MADS-box family [Becker et al. 2000]). Therefore an assessment of the structural relationships between all *A. thaliana bHLH* genes should provide a guide for prediction of gene function and for elucidation of the range of activities carried out by members of the *bHLH* gene family. At least for related groups of animal genes, the amino acid sequence of the bHLH motif has retained sufficient information to identify evolutionary relationships (Atchley and Fitch 1997). In this article, structural analysis is presented for all *A. thaliana* genes encoding proteins with a bHLH motif. Gene and protein sequence prediction is based on sequence information from newly isolated cDNAs. The results of the structural relationships between the proteins are discussed in the context of the functional diversity of this family of TFs in plants.

Materials and Methods

Identification of bHLH Motifs in the *A. thaliana* Genome Sequence

Using a consensus sequence based on that predicted by Atchley and Fitch (1997) for the bHLH motif, *A. thaliana* genomic DNA sequence was scanned using the TBLASTN algorithm (Altschul et al. 1997). In addition, DNA sequences corresponding to open reading frames (ORFs) predicted by the *A. thaliana* genome annotation (TAIR: www.arabidopsis.org; MIPS: mips.gsf.de/proj/thal/) were scanned for the bHLH motif. To this collection of putative *bHLH* genes, additional candidate *bHLH* genes were added which encode proteins containing the INTERPRO domains IPR001092, IPR003015 (www.ebi.ac.uk/interpro/) and/or the PROSITE motif PS00038 (www.expasy.ch/prosite). Finally, following the identification of groups of genes with distinct but similar sequence, the PSI-Blast algorithm (Altschul et al. 1997) was used to search for more *bHLH* genes belonging to each group that may have escaped the initial database searches. Previously, some genes had been classified as of the *bHLH* type which did not encode a bona fide bHLH motif according to our definition (lacking several amino acids that are highly conserved). These were excluded from the analysis.

The exon-intron structure prediction for each gene was improved by using a combination of gene-modeling programs (GeneMark: opal.biology.gatech.edu/GeneMark; GeneScan: genome.dkfz-heidelberg.de/cgi-bin/GENSCAN/genscan.cgi) and comparison with available ESTs. Where ESTs were not available, the predicted protein and the genes most closely related to it were aligned and the gene model was altered manually if there was evidence for greater similarity in gene structure than predicted. In some cases, small exons in the basic region were missing in the existing predictions (e.g., *AtbHLH16*).

Isolation and Sequencing of cDNAs for the *A. thaliana* bHLH TFs

To verify the ORFs for each *bHLH* gene, and to confirm expression, cDNAs corresponding to the majority of genes were isolated using the Rapid Amplification of cDNA Ends (RACE) protocol (Frohman, Dush, and Martin 1988). If available, cDNAs encoding bHLH proteins identified in EST collections were sequenced. All sequencing was on both strands. DNA sequences were determined by the MPIZ DNA core facility on Applied Biosystems (Weiterstadt, Germany) Abi Prism 377 and 3100 sequencers using BigDye-terminator chemistry. Premixed reagents were from Applied Biosystems. Oligonucleotide primers were purchased from Metabion (Martinsried, Germany) or Invitrogen (Carlsbad, Calif.).

Expression Analysis

Total RNA was isolated from cultured *A. thaliana* cells (At7; Trezzini, Horrichs, and Somssich 1993) that were submitted to eight different treatments: untreated, incubation for 6 h at 4°C in the dark, incubation at 37°C in the dark, incubation under UV light at 28°C, and incubation with 10 μM 1-aminocyclopropan-1-carboxylic acid at 28°C. To simulate a stress response to pathogen attack, cell cultures were also incubated with 1 μM Flagellin22 at 28°C, with 10 μM methyl-jasmonate for 40 min at 28°C or 10 μM salicylic acid for 40 min at 28°C. In addition, RNA was isolated from roots, leaves, stems, and flowers and siliques of 6- to 8-week-old plants

grown in the greenhouse under long day conditions. Using CDSIII-NotI primer (ATTCTAGAGGCCGAGGCGGCC-GCCATG(T_{30})VN), 5 μg of each RNA was reverse transcribed in a 20 μl reaction with Superscript II RT polymerase (Invitrogen). After addition of 100 μl water, 1 μl of the diluted reaction was used for a first round of polymerase chain reaction (PCR) with a gene-specific primer, and CDSIII-NotI. PCR was carried out in a 96-well format, where 8 different genes could be probed against the array of 12 cDNA samples. After a first denaturation step of 2 min at 94°C and 30 cycles of 20 s at 94°C, 20 s at 60°C, and 2 min at 72°C, an aliquot of 1 μl of each reaction was transferred to a new 96-well plate and subjected to a second round of PCR with the same conditions as in round one, except that a second, nested, gene-specific primer was included. Resulting PCR fragments were analyzed on gels, and bands of the expected size were extracted and subsequently inserted into pTOPO (Invitrogen). Two to six clones were sequenced for each gene.

Clustering of *bHLH* Genes into Groups and Subgroups

For each gene, only the amino acid sequence from the bHLH domain was used for creating a manually edited alignment for sequence comparison with the ClustalW (Thompson, Higgins, and Gibson 1994) program. A similarity tree was constructed (Neighbor-Joining algorithm) followed by bootstrap analysis using PHYLIP (Felsenstein 1993). Based on the sequence information, 500 trees were calculated and all branches that appeared in more than 250 trees were used to calculate the final clustering. In addition, we used the MEME analysis tool (Bailey and Elkan 1994) to search for motifs shared by the AtbHLH proteins.

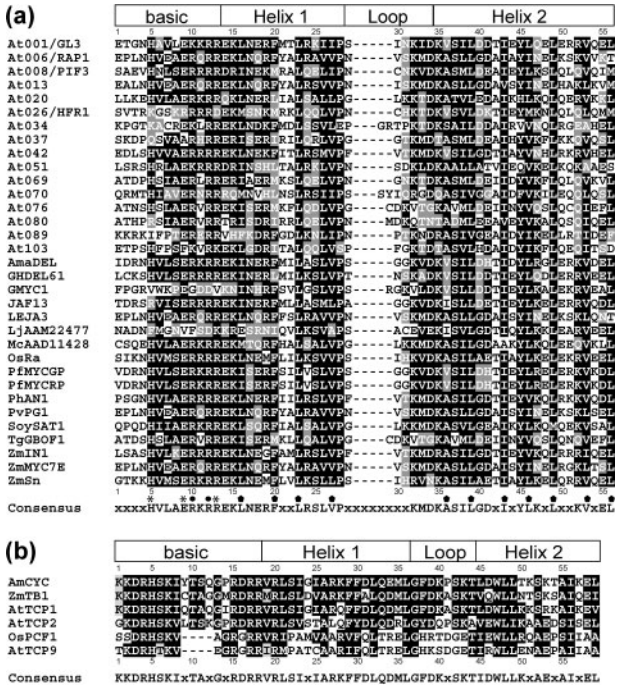


FIG. 1.—(*a*) Alignment of selected bHLH domains from plants. Icons between alignment and consensus indicate residues important for DNA binding or protein–protein interaction (asterisk, amino acid contacts with nucleotide bases; dot, amino acid contacts with DNA backbone; pentagon, nonpolar residues in protein–protein interaction). Shown at the top are the boundaries used in this study to demarcate the basic region and the two α-helices. Many plant bHLH domains possess a configuration of amino acid residues (H-E-R) at positions 5, 9, and 13 (marked with asterisks) that is also conserved in and defines the animal group B proteins. (*b*) Alignment of selected bHLH domains present in plant TCP proteins. Convergence in structure is indicated by boxes above. Convergence to the classical bHLH proteins is restricted to three-dimensional structure but might extend to function.

Results and Discussion

The *A. thaliana* Genome Contains 133 Genes with a bHLH Motif

A consensus sequence derived from an alignment of well-known plant sequences in the region surrounding the HLH motif was used to identify *bHLH* genes in *A. thaliana* (fig. 1*a*). Variations from this consensus were observed in a number of genes, but a high degree of conservation was observed in those positions that are known to have important functions in DNA binding and protein dimerization. The frequencies of each amino acid at every position in the bHLH domain deduced from all 133 *bHLH* genes are shown in figure 2. Within those bHLH proteins with proven ability to bind DNA, the amino acids at positions 5, 9, and 13 are the most critical. Non-plant bHLH proteins with His-Glu-Arg (H-E-R) at positions 5, 9, and 13 have been shown to bind to a variation of the E-box hexanucleotide sequence (E-box: CANNTG, variation: CACGTG), with His5 and Glu9 residues in contact with the outer two nucleotides and Arg13 in contact with the two inner nucleotides of the motif (Brownlie et al. 1997; Atchley, Terhalle, and Dress 1999; Ledent and Vervoort 2001). The DNA backbone is contacted by the basic residues at positions 10 and 12, and these are also conserved in the majority of plant proteins. The highly conserved hydrophobic residues in helix 1 and 2 are believed to be necessary for dimerization. For example, the structure of the human MAX protein dimer bound to DNA requires Leu23 in helix 1 to contact the corresponding Leu23 in the second bHLH polypeptide and confer stability to the DNA–protein dimer complex (Brownlie et al. 1997). In *A. thaliana* a leucine residue is present at position 23 in every bHLH protein, which emphasizes the likely importance of this residue in dimerization. Despite being well studied, there are some plant proteins (the R-like proteins) for which there is no evidence of dimerization. Based on the analysis presented (fig. 2) and the obvious importance of Leu23 in MAX dimerization, however, it is reasonable to suppose that R proteins do form dimers. Dimer formation is also known to be stabilized further by conserved hydrophobic residues in helix 2, which are located to one side of the helix. These hydrophobic residues are also conserved in plant proteins in a similar configuration, as detected by helical-wheel analysis (data not shown).

In total, 133 *AtbHLH* genes were identified, fitting well to the consensus presented in figure 2. This contrasts with the number of 139 claimed by Riechmann and Ratcliffe (2000). Unfortunately, the annotation data used

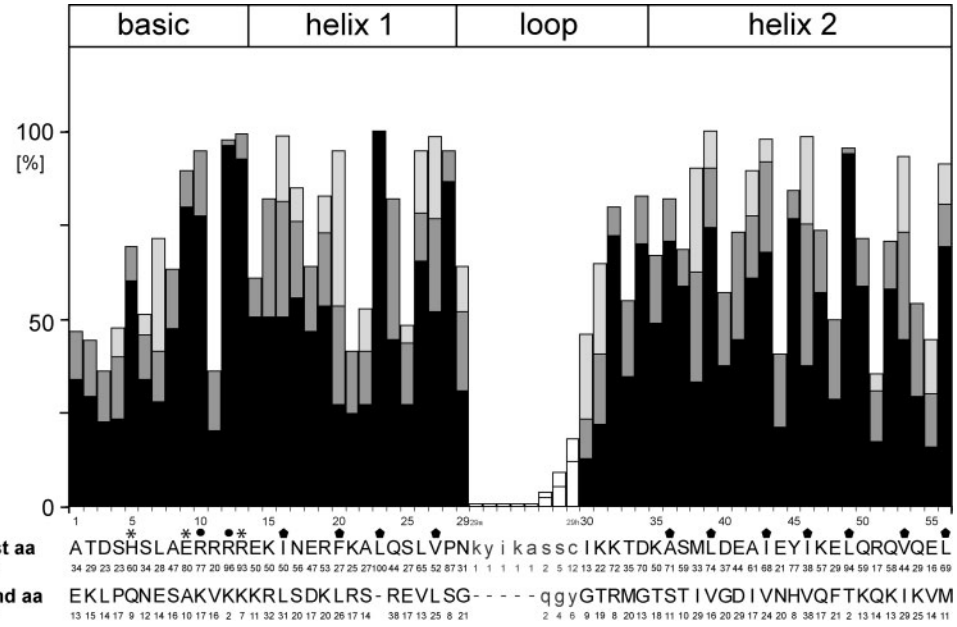


FIG. 2.—A bar chart showing the frequencies for the most common amino acid in each position across the bHLH domain. The characteristics of functionally important amino acids are indicated by icons above the consensus (see legend to fig. 1). The most frequent amino acid is shown in black; the second most frequent amino acid, in gray. Pale gray sections of a column represent those genes possessing an alternative amino acid with similar chemical characteristics. Variable loop residues not present in all bHLHs are indicated with white columns. As in animal bHLH domains, the loop between helix 1 and helix 2 is of variable length, consisting of up to 14 amino acids with 6 amino acids in most cases.

by those authors are not available for comparison. A generic name (AtbHLH001 to AtbHLH133) was given to each *bHLH* gene, including those that had already been named in previous publications. The numbering system we chose provides a unique identifier for each *bHLH* gene as proposed for the *A. thaliana*, MYB, WRKY, and bZIP TFs (Kranz et al. 1998; Romero et al. 1998; Eulgem et al. 2000; Jakoby et al. 2002).

The predicted gene structure of 87 *bHLH* genes was verified by performing 3′ RACE-PCR on 12 different cDNA sources. The results highlighted a number of incorrectly annotated ORFs (37% of all annotated bHLH genes) that had been predicted by the *A. thaliana* genome annotation. Forty-seven entries were corrected using our RACE-PCR results, including *AtbHLH125* (At1g62975), which was newly annotated; this genome region was originally described as a noncoding repeat-rich region. These new cDNA accessions have been deposited at GenBank with the assigned AtbHLHxyz gene name and the new ORF structure has been submitted to MAtDB at the Munich Institute for Protein Sequences (MIPS). The results of the RT-PCR experiment also provided a preliminary impression of the expression profile of these genes (fig. 3). In total, at least 113 *bHLH* genes can be confirmed as being expressed in *A. thaliana*, including genes previously characterized. Most of the genes tested on all 12 cDNA sources have a very broad expression spectrum; very few are specific for particular tissues (*AtbHLH039*, *040*, *050*, and *101*) or treatments (*AtbHLH085*). Because this PCR-based approach was designed for the maximum amplification of cDNAs, it does not provide any insight into the relative expression levels of the genes. However, our results provide the basis for the creation of cDNA (and other) arrays with a high coverage of *bHLH* genes.

The TCP (TB1-CYC-PCFs) family is a small subgroup of proteins sharing a common motif that is predicted to form a basic-helix–loop–helix structure (Cubas et al. 1999). Two rice proteins from this gene family, PCF1 and PCF2, have been shown to bind DNA, and the importance of the bHLH motif in the binding ability of PCF1 has been demonstrated by deletion analysis (Kosugi and Ohashi 1997). However, although these TCP proteins have a DNA binding domain with a predicted secondary structure similar to bHLH proteins, their primary amino acid sequences are unrelated to that of all other members of the bHLH family, including other eukaryotes (fig. 1*b*). Therefore, TCP proteins are not considered further in our study. It is possible that the TCP bHLH-like structure and the bHLH structure of the conventional bHLH proteins provide an example of convergent evolution.

The Structural Relationships Between Plant *bHLH* Genes

Based on the amino acid alignments of 133 bHLH domains, a comparative tree was constructed. The clustering showed that there are 12 major groups (subfamilies) of related sequences which were strongly supported by bootstrap analysis. The conclusions from this analysis are presented in figure 4, with closely related genes within subfamilies grouped together. Comparing all aspects of gene structure, genes within each subfamily contain a similar number of introns with conserved positions, and show similar predicted lengths for the encoded proteins and

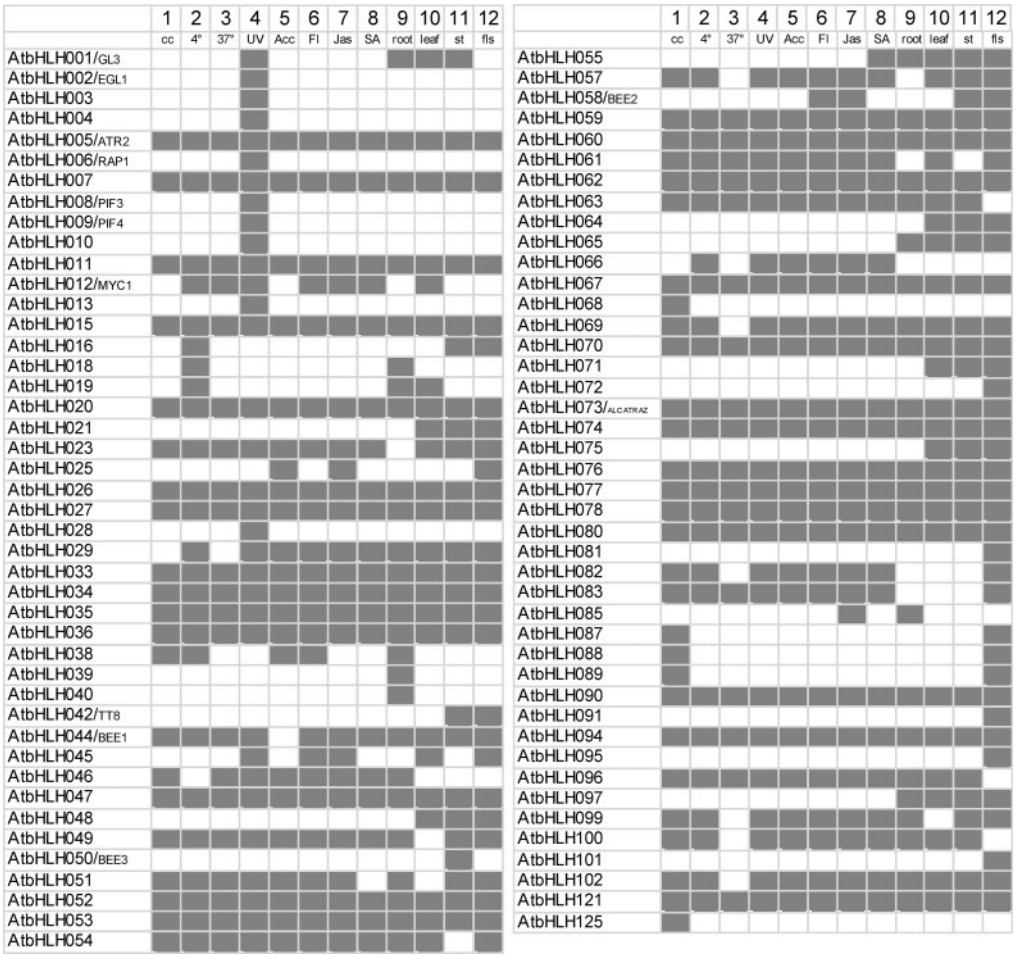


FIG. 3.—Expression patterns for 87 *A. thaliana* bHLH genes detected by RT-PCR. Detectable expression is indicated by a gray box; a white box indicates that no expression was detectable. Columns 1–8, RNA isolated from cell cultures: 1, untreated; 2, incubated at 4°C for 6 h in the dark; 3, incubated at 37°C for 6 h in the dark; 4, incubated under UV light for 6 h; 5, incubated with 10 μM 1-aminocyclopropan-1-carboxylic acid for 6 h; 6, incubated with 1 μM Flagellin22 for 6 h; 7, incubated with 10 μM methyl-jasmonate for 40 min; 8, incubated with 10 μM salicylic acid for 40 min. Columns 9–12, RNA isolated plant tissues: 9, roots; 10, leaves; 11, stems; 12, flowers and siliques. For genes that have been published under various names see figure 4 legend.

similar positions for the bHLH domain within the protein. In many subfamilies, the proteins also share related amino acid sequences outside the DNA binding domain (fig. 4). These motifs give added support to the clustering obtained using the DNA binding domain alone and in several cases help to separate the subfamilies further into subgroups. Some of these motifs probably act as activation domains and may also be important for interaction with other modules of the transcription complex, or they may be targets of signal transduction chains. For example, proteins belonging to sub-groups IIId, IIIe, and IIIf have a conserved stretch of amino acids positioned toward the N-termini of the proteins, a region that is found only in plant bHLH proteins (fig. 4). Amino acids within this region are predicted to form α-helixes, some of which are highly amphipathic (Goff, Cone, and Chandler 1992). A conserved aspartate residue in this region (within motif 11) of the AtbHLH005/ATR2 protein appears to be functionally important for correct expression of several downstream target genes that belong to the tryptophan biosynthesis pathway (Smolen et al. 2002). Between the bHLH domain and the conserved N-terminal domain is a sequence rich in acidic amino acids with significant negative charge (Lc: Ludwig et al. 1989; B: Radicella, Turks, and Chandler 1991; DEL: Goodrich, Carpenter, and Coen 1992), a feature of activating domains identified in known transcriptional activators (Ptashne 1988). Using *trans*-activation assays Goff, Cone, and Chandler (1992) demonstrated that a region of the anthocyanin regulator B, comprising the N-terminal domain and part of the acidic region, was important for activation of a *Bz1* promoter-reporter construct. However, activation only occurred when the C1 MYB protein was also present in the assay, indicating that an interaction of B with C1 was necessary to obtain an active transcription complex. The interacting region within C1 was mapped to the MYB DNA binding domain (Goff, Cone, and Chandler 1992), and more recently, particular amino acids within the MYB domain of C1 have been shown to be important for an interaction with the R bHLH protein (Grotewold et al. 2000).

Group	Generic name[a]	Synonym[b]	AGI genecode	Accession	Exon Number	5-9-13 Motif
Ia	AtbHLH067 *		At3g61950	AF488600	5	H-E-R
	AtbHLH097		At3g24140	AF488624	4	H-E-R
	AtbHLH070 *		At2g46810	AF488602	4	H-E-R
	AtbHLH071		At5g46690	AF488603	2	H-E-R
	AtbHLH057		At4g01460	AF488590	3	H-E-R
	AtbHLH099		At5g65320	AF488625	3	H-E-R
	AtbHLH096		At1g72210	AJ459771	3	H-E-R
	AtbHLH094		At1g22490	AF488622	3	H-E-R
	AtbHLH098		At5g53210	n.a.	4	H-E-R
	AtbHLH045		At3g06120	AF488580	3	H-E-R
Ib	AtbHLH095		At1g49770	AF488623	3	H-E-R
	AtbHLH055		At1g12540	AF488589	3	H-E-R
	AtbHLH125		At1g62975	AF506369	3	H-E-R
	AtbHLH126 **		At4g25410	Z46563	3	H-E-R
	AtbHLH118 *		At4g25400	n.a.	3	H-E-R
	AtbHLH120 *		At5g51790	n.a.	3	H-E-R
	AtbHLH036 **		At5g51780	AF488575	3	H-E-R
	AtbHLH100 *		At2g41240	AF488626	2	H-E-R
	AtbHLH101		At5g04150	AJ519810	3	H-E-R
	AtbHLH038 *		At3g56970	AF488576	2	H-E-R
	AtbHLH039 **		At3g56980	AF488577	2	H-E-R
II	AtbHLH091 *		At2g31210	AJ519809	3	P-E-R
	AtbHLH010 **		At2g31220	AF251695	3	T-E-R
	AtbHLH089 *		At1g06170	AF488619	3	I-E-R
IIIa	AtbHLH022		At4g21330	n.a.	3	N-E-R
	AtbHLH029		At2g28160	AF488570	4	T-E-R
	AtbHLH021		At2g16910	AF488565	5	N-E-R
IIIb	AtbHLH093		At5g65640	AF488621	4	N-E-R
	AtbHLH061		At5g10570	AF488594	4	N-E-R
	AtbHLH033		At1g12860	AF488572	4	N-E-R
	AtbHLH116		At3g26744	AY079016	4	N-E-R
IIIc	AtbHLH090		At1g10610	AF488620	8	N-E-R
	AtbHLH035		At5g57150	AF488574	5	N-E-R
	AtbHLH027		At4g29930	AF488569	5	N-E-R
IIId	AtbHLH003		At4g16430	AF251688	1	H-E-R
	AtbHLH013	MYC7E	At1g01260	AY120752	1	H-E-R
	AtbHLH017		At2g46510	AY094399	1	H-E-R
	AtbHLH014		At4g00870	AJ519812	2	H-E-R
IIIe	AtbHLH006	RAP-1*	At1g32640	X99548	1	H-E-R
	AtbHLH004 *		At4g17880	AF251689	1	H-E-R
	AtbHLH005 **		At5g46760	AF251690	1	H-E-R
	AtbHLH028 *		At5g46830	AF252636	1	H-E-R
IIIf	AtbHLH042	TT8	At4g09820	AJ277509	7	H-E-R
	AtbHLH012	ATMYC1*	At4g00480	AF251697	8	L-D-R
	AtbHLH001 *	GI3	At5g41315	AF246291	7	H-E-R
	AtbHLH002 *	EGL1*	At1g63650	AF027732	7	H-E-R
IVa	AtbHLH020 *		At2g22770	AF488564	3	H-E-R
	AtbHLH019 *		At2g22760	AF488563	3	H-E-R
	AtbHLH018 **		At2g22750	AF488562	3	H-E-R
	AtbHLH025 *		At4g37850	AF488567	3	H-E-R
IVb	AtbHLH047		At3g47640	AF488582	4	K-E-R
	AtbHLH011		At4g36060	AF251696	5	K-E-R
	AtbHLH121		At3g19860	AF488633	4	K-E-R
IVc	AtbHLH105		At5g54680	AF488629	5	K-E-R
	AtbHLH115		At1g51070	AF488632	5	K-E-R
	AtbHLH034 *		At3g23210	AF488573	4	K-E-R
	AtbHLH104 *		At4g14410	AF488628	4	K-E-R
IVd	AtbHLH041		At5g56960	n.a.	5	H-E-R
	AtbHLH092		At5g43650	AY065390	3	H-E-R
Va	AtbHLH102		At1g69010	AF488627	7	H-E-R
	AtbHLH046		At5g08130	AF488581	10	H-E-R
Vb	AtbHLH131		At4g38071	n.a.	2	H-E-R
	AtbHLH030		At1g68810	AY072161	2	H-E-R
	AtbHLH032		At3g25710	AF488571	2	H-E-R
	AtbHLH107 *		At3g56770	n.a.	2	H-E-R
	AtbHLH106 *		At2g41130	AY074639	2	H-E-R
	AtbHLH051 **		At2g40200	AF488586	2	H-E-R

Conserved Domains[c]

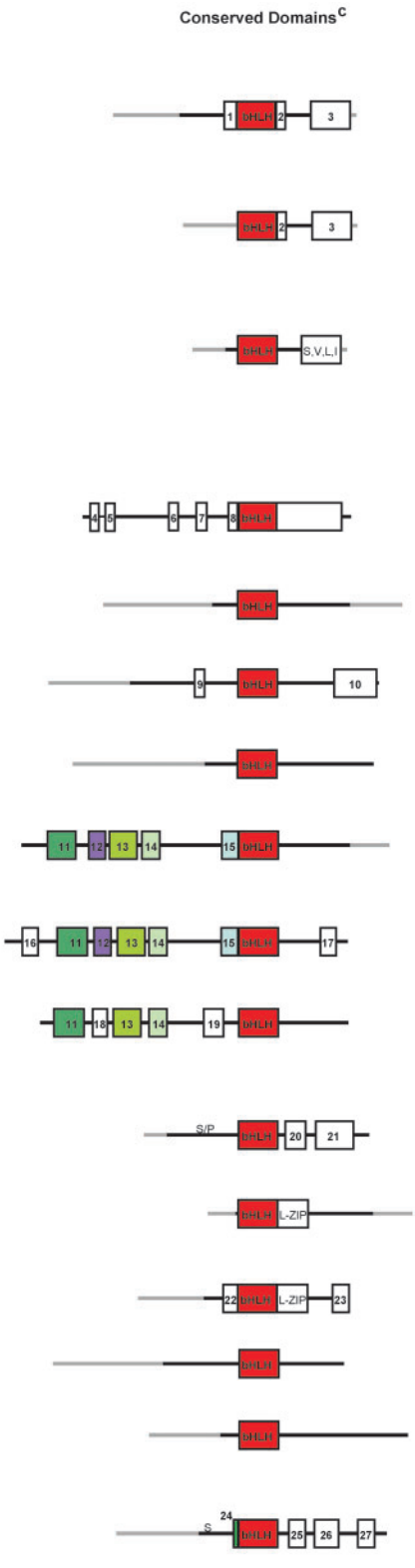
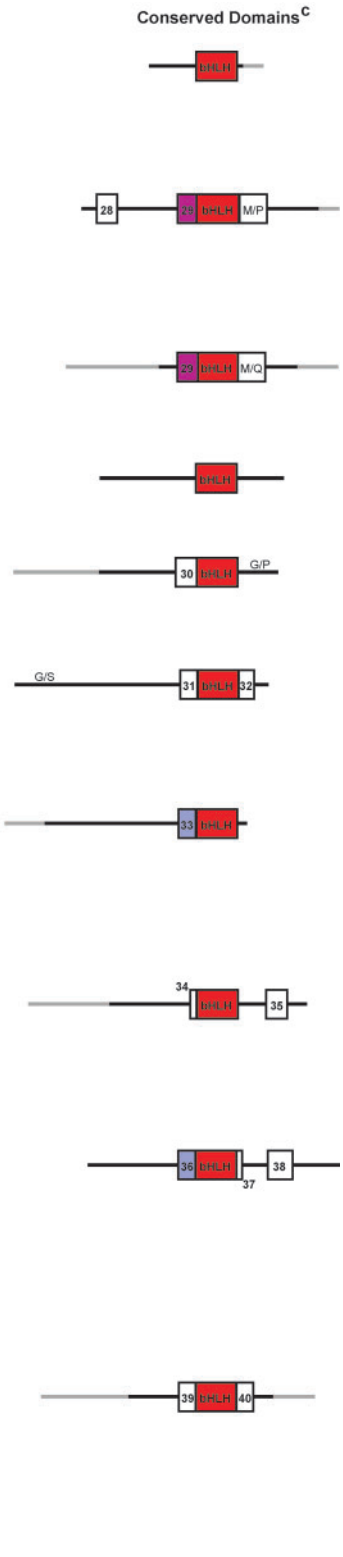


FIG. 4.—Subdivision of the *A. thaliana* bHLH gene family into groups and subgroups based on structural similarities. *bHLH is described by more than one author with different names (AtbHLH002/Atmyc-2/EGL1; AtbHLH006/RAP-1/rd22BP1; AtbHLH012/ATMYC-1/MYC1; AtbHLH026/HFR1/REP1/FBI1). n.a. indicates that there is no available GenBank accession for the cDNA sequence. [a]Red and blue asterisks indicate genes that may be functionally redundant due to genome duplication events. red: duplicated genes at unlinked loci; blue: duplicated genes with adjacent genomic location. For a graphical view of bHLH clusters on the *A. thaliana* chromosomes, we recommend the ChromosomeMap tool at TAIR (http://

Group	Generic Name[a]	Synonym[b]	AGI genecode	Accession	Exon No.	5-9-13 Motif
VI	AtbHLH109		At1g68240	n.a.	5	R-E-R
	AtbHLH108		At1g25310	n.a.	4	D-K-R
VIIa	AtbHLH065 *		At3g59060	AF488598	5	H-E-R
	AtbHLH009 *	PIF4	At2g43010	AF251694	6	H-E-R
	AtbHLH008	PIF3	At1g09530	AF251693	6	H-E-R
	AtbHLH056 *		At4g28800	n.a.	8	H-E-R
	AtbHLH127 *		At4g28815	n.a.	6	H-E-R
	AtbHLH119 *		At4g28811	AJ519811	7	H-E-R
	AtbHLH023 **		At4g28790	AF488566	6	H-E-R
	AtbHLH015 *		At2g20180	AF488560	7	H-E-R
	AtbHLH124 *	PIL1	At2g46970	AB090873	6	H-E-R
	AtbHLH132 *	PIL2	At3g62090	AB090874	5	Y-E-R
VIIb	AtbHLH026	HFR1*	At1g02340	AF488568	5	K-R-D
	AtbHLH072		At5g61270	AF488604	5	H-E-R
	AtbHLH016		At4g00050	AF488561	5	H-E-R
	AtbHLH073	ALCATRAZ	At5g67110	AF488605	5	H-E-R
	AtbHLH024	SPATULA	At4g36930	AF319540	7	H-E-R
VIIIa	AtbHLH117		At3g22100	n.a.	1	N-I-R
	AtbHLH052 *		At1g30670	AF488587	2	Q-A-R
	AtbHLH053 *		At2g34820	AF488588	2	Q-A-R
VIIIb	AtbHLH087		At3g21330	AF488617	1	Q-A-R
	AtbHLH088 *		At5g67060	AF488618	1	Q-A-R
	AtbHLH037 *		At3g50330	n.a.	1	Q-A-R
	AtbHLH043		At5g09750	n.a.	1	Q-A-R
	AtbHLH040		At4g00120	AF488578	1	Q-A-R
VIIIc	AtbHLH054		At1g27740	AY085436	4	Q-A-R
	AtbHLH084		At2g14760	n.a.	5	Q-A-R
	AtbHLH085		At4g33880	AF488616	5	Q-A-R
	AtbHLH083 *		At1g66470	AF488615	5	Q-A-R
	AtbHLH086 *		At5g37800	n.a.	5	Q-A-R
IX	AtbHLH130		At2g42280	n.a.	6	R-E-R
	AtbHLH128		At1g05805	AY045907	6	R-E-R
	AtbHLH129		At2g43140	AU237473	5	R-E-R
	AtbHLH122		At1g51140	AY063120	6	R-E-R
	AtbHLH081		At4g09180	AF488613	5	R-E-R
	AtbHLH080		At1g35460	AF488612	5	R-E-R
X	AtbHLH113		At3g19500	AF488631	5	S-A-K
	AtbHLH123		At3g20640	AU238908	7	P-F-K
	AtbHLH068		At4g29100	AF488634	7	S-L-K
	AtbHLH133		At2g20095	n.a.	7	S-L-K
	AtbHLH110		At1g27660	n.a.	7	S-F-K
	AtbHLH103		At4g21340	AY065362	6	H-F-K
	AtbHLH114		At4g05170	n.a.	6	P-F-K
	AtbHLH112		At1g61660	AF488630	6	P-F-K
	AtbHLH111		At1g31050	AA395190	6	L-K-K
XI	AtbHLH069 *		At4g30980	AF488601	6	H-E-R
	AtbHLH066 *		At2g24260	AF488599	6	H-E-R
	AtbHLH082		At5g58010	AF488614	5	H-E-R
	AtbHLH007 *		At1g03040	AF251692	6	H-E-R
	AtbHLH059 *		At4g02590	AF488592	6	H-E-R
XII	AtbHLH075		At1g25330	AF488607	6	H-E-R
	AtbHLH050 *	BEE3	At1g73830	AF488585	6	H-E-R
	AtbHLH044 *	BEE1	At1g18400	AF488579	6	H-E-R
	AtbHLH063		At4g34530	AF488596	6	H-E-R
	AtbHLH049 *		At1g68920	AF488584	8	H-E-R
	AtbHLH076 *		At1g26260	AF488608	8	H-E-R
	AtbHLH078		At5g48560	AF488610	7	H-E-R
	AtbHLH062		At3g07340	AF488595	8	H-E-R
	AtbHLH074		At1g10120	AF488606	8	H-E-R
	AtbHLH077		At3g23690	AF488609	7	H-E-R
	AtbHLH048 *		At2g42300	AF488583	6	H-E-R
	AtbHLH060 *		At3g57800	AF488593	6	H-E-R
	AtbHLH031	ZCW32	At1g59640	AB028232	5	H-E-R
	AtbHLH079		At5g62610	AF488611	6	H-E-R
	AtbHLH058 *	BEE2	At4g36540	AF488591	5	H-E-R
	AtbHLH064 *		At2g18300	AF488597	6	H-E-R

Conserved Domains[c]



arabidopsis.org/jsp/ChromosomeMap/tool.jsp). [b] References not mentioned in text: MYC7E, PIL2 and ZCW32 sequences were directly submitted to GenBank. AtbHLH002/Atmyc-2/EGL1 (*enhancer of glabra3*) data were submitted by Alan M. Lloyd and F. Zhang (1997) directly to GenBank as Atmyc-2 and published as EGL1 (Bernhardt et al. 2001; Zhang et al. 2001). [c] bHLH motif displayed as red box; other regions conserved among members of groups are highlighted in different colors.

It is not clear whether R-like bHLH proteins (group III) bind DNA in the bHLH-MYB complex or whether DNA binding occurs only via the MYB partner; to date only C1 has been shown to bind *cis*-acting elements in the promoter of the *Z. mays A1* (*DFR*) gene (Sainz, Grotewold, and Chandler 1997). Within bHLH group III, B is most similar to *A. thaliana* AtbHLH012/MYC1, a bHLH protein from *A. thaliana* identified by Urao et al. (1996).

The *Z. mays Intensifier1* gene product (In1: Burr et al. 1996) also belongs to group III and is more closely related to PhAN1 (ANTHOCYANIN1 of *Petunia hybrida*: Spelt et al. 2000) and AtbHLH042/TT8 (Nesi et al. 2000) than it is to the R-like proteins AtbHLH012/MYC1, PhJAF13 (Quattrocchio et al. 1998) or AmDEL (DELILA of *Antirrhinum majus*: Goodrich, Carpenter, and Coen 1992; de Pater et al. 1997). Most group III members with known functions act as TFs regulating genes of flavonoid metabolism. In contrast, *ZmIn1* encodes a repressor that regulates the expression of the *White pollen1* (*Whp1*) gene, one of the *Z. mays* genes encoding chalcone synthase (CHS: Franken et al. 1991) and also the *Bronze1* gene, which encodes UDP-flavonoid 3-*O*-glucosyl transferase and catalyzes one of the last steps in the anthocyanin biosynthetic pathway (Klein and Nelson 1983). The predicted structure of In1 is similar to that of AtbHLH042/TT8 and PhAN1, but many *ZmIn1* transcripts are miss-spliced (Burr et al. 1996). This may result in premature termination of translation and, more specifically, in a protein lacking the bHLH domain. The means by which truncated In1 protein negatively regulates *Whp1* and *Bz1* gene expression are not yet clear.

Specific functions for other genes belonging to group III are known, and in contrast to the R-like proteins, the corresponding gene products have been shown to bind DNA. The *Phaseolus vulgaris* bHLH protein PG1 binds to a G-box motif in the β-*phaseolin* gene promoter that is important for positive transcriptional regulation (Kawagoe and Murai 1996), and the *A. thaliana* rd22BP1 protein (AtbHLH006/rd22BP1/RAP-1) binds to a region of the *rd22* gene promoter that is sufficient for dehydration- and ABA-induced expression of the *rd22* gene (Abe et al. 1997; de Pater et al. 1997). The tomato *Fer* gene—structurally closely related to *A. thaliana* group III bHLH genes—has been shown recently to have a role in iron uptake in roots (Ling et al. 2002), and so the diverse functional nature of this group of bHLH proteins in *A. thaliana* is emerging.

Members of group IV have a highly conserved leucine zipper (ZIP) motif adjacent to the second helix of the bHLH motif. The ZIP domain is predicted to adopt a coiled-coil structure that permits dimerization between proteins (Lupas 1996). Some animal proteins also contain a ZIP domain in the same position as the plant bHLH-ZIP proteins, and a few members of group III also have a partially conserved ZIP motif. In animal bHLH-ZIP proteins, there is evidence that the ZIP motif stabilizes protein dimers (Brandt-Rauf et al. 1989; Bresnick and Felsenfeld 1994) and that particular residues within the ZIP domain determine dimerization specificity (Marchetti et al. 1995). The functions of the plant group IV bHLH-ZIP proteins are unknown.

Recently, three members of group XII, AtbHLH044/BEE1, AtbHLH058/BEE2, and AtbHLH050/BEE3 (*BR Enhanced Expression*) from *A. thaliana* have been linked to multiple pathways regulating plant growth and development (Friedrichsen et al. 2002). These closely related bHLHs act redundantly as positive regulators in the early Brassinosteroid (BR) signaling pathway and they also affect signaling by abscisic acid (ABA), a known antagonist of BR.

One might anticipate that proteins clustered on the tree have similar biological functions. However the genes that make up subgroup IIIf are involved in very different processes; flavonoid/anthocyanin biosynthesis (*AtbHLH0012/MYC1* and *AtbHLH042/TT8*) and trichome initiation (*AtbHLH001/GL3*). The maize gene, *Lc*, encodes a member of the *R/B* family and is orthologous to the *A. thaliana* gene *AtbHLH012/MYC1*. Overexpression of Lc in wild-type *A. thaliana* results in additional trichomes and an elevated level of anthocyanins compared to control plants (Lloyd, Walbot, and Davis 1992). However, in maize, members of the *R* gene family have no influence on trichome formation. The bHLH protein bHLH001/GL3, closely related to Lc (Payne, Zhang, and Lloyd 2000), belongs to a cascade of transcriptional regulators that control the differentiation of trichome progenitor cells in shoot epidermis. Although the maize *Lc* gene was overexpressed in a heterologous context, the observation that it can induce trichome formation in *A. thaliana* suggests that closely related genes retain the potential to cross over and regulate parallel processes (requiring similar transcriptional activators) when expressed at artificially high levels. This may reflect the importance that patterns of expression and strength of expression have on the specific function(s) carried out by transcription factors in wild-type tissues.

Evolution of bHLH Protein Structure and Function

In their phylogenetic classification of animal bHLH protein sequences, Atchley and Fitch (1997) identified more than 24 protein lineages containing proteins with related function. They found that these lineages could be organized further into five groups (A to E) that relate to the preferred DNA binding sequence of the proteins within each lineage. Within each group proteins bound to a variation of the palindromic hexanucleotide sequence CANNTG known as the E-box motif. Patterns of amino acids were found at three positions within the basic region of the bHLH motif (the 5–9–13 configuration) that defined the five different subgroups very accurately and suggested that the amino acid sequence motif of group B, with an H-E-R configuration, was the ancestral sequence (Atchley and Fitch 1997).

In *A. thaliana*, three proteins, namely PvPG1, AtbHLH006/rd22BP1/RAP-1 (both group III), and more recently AtbHLH008/PIF3 (group VII; Martinez-Garcia, Huq, and Quail 2000) have been shown to bind a sequence identical or very similar to the B variant of the animal E-box motif, which is also identical to the G-box core motif (CACGTG), a ubiquitous regulatory DNA element found in plants that is also bound by some bZIP TFs (Menkens,

Schindler, and Cashmore 1995). Most (87 of 133) *A. thaliana* genes have the H-E-R configuration within the bHLH domain that is found in members of group B from animals (fig. 4), suggesting that plant bHLH proteins evolved from one or several members of group B that were present in early eukaryotic lineages. In this context it is also interesting that seven plant bHLH proteins in group IV contain their ZIP motif in the identical position to animal bHLH-ZIP proteins, which all—with one exception—belong to the class B proteins. Some plant proteins, notably members of groups VIII and X, have a variation of the H-E-R configuration and contain helix-breaking prolines in the basic region (fig. 4), characteristics that may interfere with affinity for DNA. These variations may enable these proteins to act as negative regulators, retaining the ability to dimerize with other bHLH proteins but lacking the ability to bind DNA. This would make group X members candidates for reclassification into group D according to Atchley and Fitch (1997).

From comparative studies it is apparent that the bHLH motif was present in early eukaryotes and that the subfamilies of genes distinguishable today have emerged following the separation of plants, animals, and fungi (Ledent and Vervoort 2001). *A. thaliana* possesses more *bHLH* genes than are found in *Caenorhabditis elegans* or *D. melanogaster* (39 and 59 genes, respectively), but is comparable to the numbers obtained for human and mouse, 125 and 102, respectively (Ledent, Paquet, and Vervoort 2002). Although no striking similarities in gene structure were detected between plants and these other lineages, except for similarities to animal group B proteins, it is still possible that some plant proteins have retained functions possessed by the ancestral eukaryotic proteins such as the developmental control of cell lineages. For example, MyoD and related proteins are involved in the determination of muscle cell lineages, and expression is confined to these cells. They interact with the ubiquitously expressed E12 family of bHLH proteins to produce active heterodimers (French et al. 1991). Furthermore, E12 and related proteins are involved in the determination and maintenance of other cell types through their ability to dimerize with other proteins that are expressed in specific cell lineages (Littlewood and Evan 1998, pp. 27–35). Recently, two members of the plant group VII have been shown to be important in the development of floral structures. Mutants in the gene *AtbHLH024*/*SPATULA* are defective specifically in the development of the style and stigma and, with strong mutant alleles, cells comprising the pollen-transmitting tract tissue do not form (Alvarez and Smyth 1999; Heisler et al. 2001). However, expression is not confined to these tissues, indicating that SPATULA may have additional roles in other aspects of growth and development. The other member of group VII, AtbHLH073, is the most closely related *A. thaliana* protein to SPATULA and is responsible for the differentiation of a nonlignified cell layer necessary for cell separation during dehiscence (Rajani and Sundaresan 2001).

Mediating responses to the environment may be another ancient function retained by plant bHLH proteins, for example, controlling responses to light and interacting with components of the circadian clock. The PIF-like 1 protein (AtbHLH124/PIL1; group VII) has been shown to interact with APRR1, a factor that has been implicated in circadian rhythms (Makino et al. 2002) and that belongs to a class of pseudo-response regulators that are components of signal transduction systems. In addition, AtbHLH008/PIF3 (group VII) can regulate expression of the circadian clock genes *CCA1* and *LHY*, suggesting that this protein provides an entry point to phytochrome regulation of the circadian clock (Martinez-Garcia, Huq, and Quail 2000). Studies indicate that light regulation and control of biorhythms are functionally interrelated and possibly have a common evolutionary origin (Kay 1997). Photoreceptor proteins and protein components of the circadian clock from species belonging to several kingdoms of life contain PAS (PER-ARNT-SIM) domains (Nambu et al. 1991) that have been shown to mediate protein–protein interactions. These include the *D. melanogaster* PER protein that interacts via its PAS domain with Timeless to regulate the circadian clock (Zeng et al. 1996), and the White Collar 2 protein (WC2) from *Neurospora crassa* that is involved in the blue light regulation of gene expression (Linden and Macino 1997) and has also been shown to be a clock component (Crosthwaite, Dunlap, and Loros 1997). A PAS-like domain has been identified in PIF3 and a related protein, PIF4 (Huq and Quail 2002), but in both cases the domain shows only modest similarity to animal PAS domains, although the region encompassing the PAS domain in PIF3 is important for the interaction of PhyB (Ni et al. 1996).

Protein–Protein Interactions and Their Importance in Gene Regulation

In mammals and *D. melanogaster* there are many instances in which bHLH proteins have been shown to homodimerize or heterodimerize as a means of controlling transcription (Littlewood and Evan 1998). The ability to heterodimerize allows a large number of potential DNA binding complexes that may have different biochemical properties—including variable DNA binding characteristics and activation or repression potentials—thereby creating a lot of opportunities for regulating different genes or sets of genes. In plants, dimerization has been shown for two related members of group VII, AtbHLH008/PIF3, and AtbHLH026/HFR1/REP1, with proposed dominant roles in phytochrome B and phytochrome A signaling, respectively (Halliday et al. 1999; Fairchild, Schumaker, and Quail 2000; Soh et al. 2000). In yeast-two-hybrid experiments, PIF3 binds preferentially to PhyB rather than PhyA, as well as being able to interact with itself (Zhu et al. 2000); AtbHLH026/HFR1/REP1 binds neither to phyA nor to phyB, but it does bind to PIF3 in vitro, suggesting that it is likely to operate as a heterodimer in regulating transcription from target promoters (Fairchild, Schumaker, and Quail 2000). Recently the AtbHLH009/PIF4 protein, which is closely related to PIF3, has been shown to have a role in phyB-regulated de-etiolation as a negative regulator (Huq and Quail 2002). PIF4 and PIF3 have similar biochemical properties, but unlike PIF3, PIF4 is unable to bind to the

active form of PhyB when bound to the G-box element. In addition, PIF4 cannot regulate G-box–containing gene promoters including those of the circadian clock genes *CCA1* and *LHY*, which are targets of PIF3 (Martinez-Garcia, Huq, and Quail 2000). Structurally, the most similar gene to *PIF4* in *A. thaliana*, *AtbHLH065* (fig. 4) appears on a duplicated section of the *A. thaliana* genome, so it may have functions similar if not identical to those of PIF4. Subgroup VIIa contains several more genes encoding factors which are related to PIF3, PIL1, and PIF4 that have not yet been studied (AtbHLH015, AtbHLH023, AtbHLH056, AtbHLH119, and AtbHLH127) but that have (due to their structural similarity) the potential to be involved in aspects of phytochrome signaling. These five genes may have redundant functions, as two are found in duplicated sections of the genome and one is part of a cluster of four structurally similar genes, indicating that there have been recent duplications of these *bHLH* genes in *A. thaliana* (fig. 4).

Interactions can also occur between different classes of TFs. For example, members of bHLH group III associate with R2R3-MYB proteins. In a similar way to the *Z. mays* bHLH protein B, which interacts with the R2R3-type MYB protein C1 to create an active transcription complex (Goff, Cone, and Chandler 1992), bHLH TFs related to B from other species also behave as co-activators of MYB proteins to induce anthocyanin biosynthetic genes (PhJAF13; Quattrocchio et al. 1998). AtMYC1 is the most closely related protein in *A. thaliana* to these R-like proteins, but it is phylogenetically distinct from two other regulators of anthocyanin biosynthesis, ANTHOCYANIN1 (AN1) from *Petunia* (Spelt et al. 2000) and TT8 from *A. thaliana* (Nesi et al. 2000). Together with a MYB protein (AN2 in *P. hybrida*, TT2 in *A. thaliana*), these petunia bHLH proteins can regulate a specific part of the anthocyanin pathway, downstream of dihydroflavonol biosynthesis. In transfection assays, co-expression of PhAN1 or PhJAF13 with PhAN2 (a *P. hybrida* R2R3-MYB) resulted in much stronger activation of the *DFR* target gene than if either PhAN1 or PhJAF13 were expressed alone (Spelt et al. 2000). Interestingly, when PhAN1 was co-expressed with ZmP, a R2R3-MYB protein that activates a branch of flavonoid biosynthesis in *Z. mays* maternal seed tissues, expression was enhanced 10-fold more than if ZmP was expressed alone; this did not happen if PhJAF13 was co-expressed with ZmP (Spelt et al. 2000). This experiment suggests that, at least under certain conditions, specific bHLH factors differ in their affinities for particular MYB proteins, and this finding is a reflection of their structural differences.

The AtbHLH001/GL3 protein that controls trichome development in *A. thaliana* interacts with the R2R3-MYB protein AtMYB0/GLABROUS1 (GL1) in yeast-2-hybrid experiments (Payne, Zhang, and Lloyd 2000). GL1 and GL3 positively control expression of the homeodomain transcription factor GLABRA2 (GL2; Rerie, Feldmann, and Marks 1994) but it is not known how GL2 then induces the cellular requirements for the directional shaping of a developing trichome cell. Interestingly, GL1 and WEREWOLF (WER) belong to the same sub-group of R2R3-MYB proteins (subgroup 15; Stracke, Werber, and Weisshaar 2001) but control opposite epidermal cell fates: WER is expressed specifically in non-hair cells of root epidermis to prevent hair formation, whereas GL1 promotes trichome formation that is expressed only in shoot epidermis. The biochemical function of the WER protein is likely to be identical to that of GL1, because each gene can complement the mutant phenotype of the other when expressed from the reciprocal promoter (Lee and Schiefelbein 2001). In this case function is dependent on cell-specific expression and the other (co-localized) TFs with which each protein can interact. Some specificity might come through the differential expression of bHLH genes with which WER might interact, as GL3 does not affect root hair formation.

There is also evidence that another member of group III, AtbHLH006/rd22BP1/RAP-1, interacts with the R2R3-MYB protein AtMYB2 to activate gene expression from a dehydration and ABA-responsive promoter element (Abe et al. 1997). In conclusion, bHLH/MYB interactions are probably conserved for group III bHLH proteins and for members of MYB subgroups 5, 6, and 15 (Stracke, Werber, and Weisshaar 2001), but it remains to be clarified whether interactions are confined to the protein products of these particular gene clades or are more widespread. The N-terminal region conserved in group III bHLH proteins may be involved in bHLH/MYB interactions and could provide the specificity for particular protein–protein interactions (Goff, Cone, and Chandler 1992).

Future Perspectives

It is likely that the main groups of *bHLH* genes identified here are present in all higher plants and have evolved specific functions with different biochemical properties. The current challenge is to assign functions to all these subgroups and the genes within them. Although the recent work on group VII proteins illustrates that functions can be diverse within a subgroup, in other cases conservation of function between proteins of the same subgroup may be high. One reason could be that the *A. thaliana* genome contains sections of sequence duplication covering at least 60% of the entire genome which probably have arisen from a tetraploidization event (Blanc et al. 2000). By searching for *bHLH* genes within these regions (MIPS Redundancy Viewer: mips.gsf.de/proj/thal/db/gv/rv/), we estimate that duplication events account for 51 (40%) of the 133 genes (fig. 4). As expected, each duplicated gene pair detected falls within the same major clade of bHLH genes. In seven instances there are up to four bHLH genes arranged in tandem and often only one gene of the cluster is present on a duplicated section. This may be because clusters have arisen by gene duplication after a major segmental duplication event, or because genes have been lost from clusters after a major segmental duplication event. However, the clustering analysis shows that pairs of structurally very similar genes exist (data not shown) that are not in corresponding duplicated regions of the genome nor in tandem repeats. These genes may have arisen following large segmental duplications but have been translocated subsequently to other regions of the

genome, in which case they are still likely to share similar functions.

Major segmental duplication events appear to have occurred at different times during angiosperm evolution, possibly including a polyploidization event (Blanc et al. 2000; Vision, Brown, and Tanksley 2000), and the more recent duplication events may have contributed significantly to the phenomenon of gene redundancy (Pickett and Meeks-Wagner 1995). Combining the data on inferred gene duplications with clustering analysis of genes from one species (fig. 4) provides a framework for visualizing the likely functional diversity within groups of related genes such as those encoding bHLH proteins. From these data, educated guesses can be made concerning gene redundancy, which will be important information for estimating the number of knockout mutations required for observing altered phenotypes for closely related genes. From structural similarities we estimate that the bHLH family includes at least 33 discernible functions in transcriptional regulation. An automated interrogatory gene tree founded on the Arabidopsis genes would be a useful aid for providing clues to the functions for bHLH proteins in other model and crop species and would allow subsequent experiments on these genes to be designed more effectively. Knowledge about the structure of the *AtbHLH* gene family, and the option to integrate new *bHLH* genes from other plant species into this frame, will further facilitate the identification of the functions of *bHLH* genes in plants.

Acknowledgments

The work for this study was financed in part from the European Union's REGIA project granted to C.M. and B.W. (Regulatory Gene Initiative in Arabidopsis; QLG2-CT-1999–000876), and from the Deutsche Forschungsgemeinschaft to B.W. and M.H. (''Graduiertenkolleg für Molekulare Analysen von Entwicklungsprozessen'' der Universität zu Köln; IIIGK-GRK 306/1). The Ph.D. work of M.W. was funded by Norddeutsche Pflanzenzucht (NPZ), Hohenlieth, Germany. Further financial support to B.W. was from GABI (German plant genomics program ''Genomanalyse im biologischen System Pflanze''). We thank Ute Tartler and Doris Falkenhahn for technical support, Heiko Schoof (Munich Institute for Protein Sequences [MIPS], München, Germany) for help with AGI gene identifiers, Guenter Theissen and Gieta Dewal for critical reading of the manuscript, and the Max-Planck-Institute for Plant Breeding Research DNA core facility ''ADIS'' for their sequencing work.

Literature Cited

Abe, H., K. Yamaguchi-Shinozaki, T. Urao, T. Iwasaki, D. Hosokawa, and K. Shinizaki. 1997. Role of Arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. Plant Cell **9**:1859–1868.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–402.

Alvarez, J., and D. R. Smyth. 1999. Crabs Claw and Spatula, two Arabidopsis genes that control carpel development in parallel with AGAMOUS. Development **126**:2377–2386.

Atchley, W. R., and W. M. Fitch. 1997. A natural classification of the basic helix–loop–helix class of transcription factors. Proc. Natl. Acad. Sci. USA **94**:5172–5176.

Atchley, W. R., W. Terhalle, and A. Dress. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. J. Mol. Evol. **48**:501–16.

Bailey, T. L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, 28–36. AAAI Press, Menlo Park, Calif.

Becker, A., K.-U. Winter, B. Meyer, H. Saedler, and G. Theissen. 2000. MADS-Box gene diversity in seed plants 300 million years ago. Mol. Biol. Evol. **17**:1425–1434.

Berben, G., M. Legrain, V. Gilliquet, and F. Hilger. 1990. The yeast regulatory gene Pho4 encodes a helix-loop-helix motif. Yeast **6**:451–454.

Bernhardt, C., A. Gonzalez, F. Zhang, A. M. Lloyd, and J. W. Schiefelbein. 2001. Two bHLH genes are involved in Arabidopsis root epidermis cell specification. 12th International Conference on Arabidopsis Research. Wisconsin Union Conference Services, Madison, Wis.

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and M. Delseny. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12**:1093–1101.

Brandt-Rauf, P. W., M. R. Pincus, J. M. Chen, and G. Lee. 1989. Conformational energy analysis of the leucine repeat regions of C/EBP, GCN4, and the proteins of the myc, jun, and fos oncogenes. J. Prot. Chem. **8**:679–688.

Bresnick, E. H., and G. Felsenfeld. 1994. The leucine zipper is necessary for stabilizing a dimer of the helix–loop–helix transcription factor USF but not for maintenance of an elongated conformation. J. Biol. Chem. **269**:21110–21116.

Brownlie, P., T. Ceska, M. Lamers, C. Romier, G. Stier, H. Teo, and D. Suck. 1997. The crystal structure of an intact human Max-DNA complex: new insights into mechanisms of transcriptional control. Structure **5**:509–520.

Burr, F. A., B. Burr, B. E. Scheffler, M. Blewitt, U. Wienand, and E. C. Matz. 1996. The maize repressor-like gene *intensifier1* shares homology with the *r1-b1* multigene family of transcription factors and exhibits missplicing. Plant Cell **8**:1249–1259.

Crosthwaite, S. K., J. C. Dunlap, and J. J. Loros. 1997. Neurospora wc-1 and wc-2: transcription, photoresponses, and the origins of circadian rhythmicity. Science **276**:763–769.

Cubas, P., N. Lauter, J. Doebley, and E. Coen. 1999. The TCP domain: a motif found in proteins regulating plant growth and development. Plant J. **18**:215–222.

de Pater, S., K. Pham, J. Memelink, and J. Kijne. 1997. RAP-1 is an *Arabidopsis* MYC-like R protein homologue that binds to G-box sequence motifs. Plant Mol. Biol. **34**:169–174.

Ellenberger, T., D. Fass, M. Arnaud, and S. C. Harrison. 1994. Crystal structure of transcription factor E47: E-box recognition by a basic region helix–loop–helix dimer. Genes Dev. **8**:970–980.

Eulgem, T., P. J. Rushton, S. Robatzek, and I. E. Somssich. 2000. The WRKY superfamily of plant transcription factors. Trends Plant Sci. **5**:199–206.

Fairchild, C. D., M. A. Schumaker, and P. H. Quail. 2000. HFR1 encodes an atypical bHLH protein that acts in phytochrome A signal transduction. Genes Dev. **14**:2377–2391.

Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Version 3.54c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Ferré-D'Amaré, A. R., P. Pognonec, R. G. Roeder, and S. K. Burley. 1994. Structure and function of the b/HLH/Z domain of USF. EMBO J. **13**:180–189.

Ferré-D'Amaré, A. R., G. C. Prendergast, E. B. Ziff, and S. K. Burley. 1993. Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature **363**:38–45.

Franken, P., U. Niesbach-Klösgen, U. Weydermann, L. Maréchal-Drouard, H. Saedler, and U. Wienand. 1991. The duplicated chalcone synthase genes *C2* and *Whp* (*white pollen*) of *Zea mays* are independently regulated; evidence for translational control of *Whp* expression by the anthocyanin intensifying gene *in*. EMBO J. **10**:2605–2612.

French, B. A., K. L. Chow, E. N. Olson, and R. J. Schwartz. 1991. Heterodimers of myogenic helix–loop–helix regulatory factors and E12 bind a complex element governing myogenic induction of the avian cardiac alpha-actin promoter. Mol. Cell. Biol. **11**:2439–2450.

Friedrichsen, D. M., J. Nemhauser, T. Muramitsu, J. N. Maloof, J. Alonso, J. R. Ecker, M. Furuya, and J. Chory. 2002. Three redundant brassinosteroid early response genes encode putative bhlh transcription factors required for normal growth. Genetics **162**:1445–1456.

Frohman, M. A., M. K. Dush, and G. R. Martin. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA **85**:8998–9002.

Goff, S. A., K. C. Cone, and V. L. Chandler. 1992. Functional analysis of the transcriptional activator encoded by the maize *B* gene: evidence for a direct functional interaction between two classes of regulatory proteins. Genes Dev. **6**:864–675.

Goodrich, J., R. Carpenter, and E. S. Coen. 1992. A common gene regulates pigmentation pattern in diverse plant species. Cell **68**:955–964.

Grotewold, E., M. B. Sainz, L. Tagliani, J. M. Hernandez, B. Bowen, and V. L. Chandler. 2000. Identification of the residues in the Myb domain of maize C1 that specify the interaction with the bHLH cofactor R. Proc. Natl. Acad. Sci. USA **97**:13579–13584.

Halliday, K. J., M. Hudson, M. Ni, M. Qin, and P. H. Quail. 1999. poc1: an Arabidopsis mutant perturbed in phytochrome signaling because of a T DNA insertion in the promoter of PIF3, a gene encoding a phytochrome-interacting bHLH protein. Proc. Natl. Acad. Sci. USA **96**:5832–5837.

Heisler, M. G., A. Atkinson, Y. H. Bylstra, R. Walsh, and D. R. Smyth. 2001. Spatula, a gene that controls development of carpel margin tissues in Arabidopsis, encodes a bHLH protein. Development **128**:1089–1098.

Huq, E., and P. H. Quail. 2002. PIF4, a phytochrome-interacting bHLH factor, functions as a negative regulator of phytochrome B signaling in Arabidopsis. EMBO J. **21**:2441–2450.

Jakoby, M., W. Droege-Laser, T. Kroj, J. Tiedemann, J. Vicente-Carbajosa, B. Weisshaar, and F. Parcy. 2002. The bZIP family of transcription factors in *Arabidopsis thaliana*. Trends Plant Sci. **7**:106–111.

Kawagoe, Y., and N. Murai. 1996. A novel basic region/helix–loop–helix protein binds to a G-box motif CACGTG of the bean seed storage protein beta-phaseolin gene. Plant Sci. **116**:47–57.

Kay, S. A. 1997. PAS, present, and future: clues to the origins of circadian clocks. Science **276**:753–754.

Klein, A. S., and O. E. Nelson, Jr. 1983. The allelic state at *intensifier* influences accumulation of UDP glucose: flavonoid 3-*O*-glucosyltransferase in maize. Phytochemistry **22**:2612–2634.

Kosugi, S., and Y. Ohashi. 1997. PCF1 and PCF2 specifically bind to cis elements in the rice proliferating cell nuclear antigen gene. Plant Cell **9**:1607–1619.

Kranz, H. D., M. Denekamp, R. Greco et al. (14 co-authors). 1998. Towards functional characterisation of the members of the *R2R3-MYB* gene family from *Arabidopsis thaliana*. Plant J. **16**:263–276.

Ledent, V., O. Paquet, and M. Vervoort. 2002. Phylogenetic analysis of the human basic helix–loop–helix proteins. Genome Biol. **3**(6):1–18.

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

Lee, M. M., and J. Schiefelbein. 2001. Developmentally distinct MYB genes encode functionally equivalent proteins in Arabidopsis. Development **128**:1539–1546.

Linden, H., and G. Macino. 1997. White collar 2, a partner in blue-light signal transduction, controlling expression of light-regulated genes in Neurospora crassa. EMBO J. **16**:98–109.

Ling, H.-Q., P. Bauer, Z. Bereczky, B. Keller, and M. W. Ganal. 2002. The tomato fer gene encoding a bHLH protein controls iron-uptake responses in roots. Proc. Natl. Acad. Sci. USA **99**:13938–13943.

Littlewood, T. D., and G. I. Evan. 1998. Basic helix–loop–helix transcription factors. Oxford University Press, Oxford.

Lloyd, A. M., V. Walbot, and R. W. Davis. 1992. *Arabidopsis* and *Nicotiana* anthocyanin production activated by maize regulators R and C1. Science **258**:1773–1775.

Ludwig, S. R., L. F. Habera, S. L. Dellaporta, and S. R. Wessler. 1989. *Lc*, a member of the maize *R* gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the *myc*-homology region. Proc. Natl. Acad. Sci. USA **86**:7092–7096.

Lupas, A. 1996. Coiled coils: new structures and new functions. Trends Biochem. Sci. **21**:375–382.

Makino, S., A. Matsushika, M. Kojima, T. Yamashino, and T. Mizuno. 2002. The APRR1/TOC1 quintet implicated in circadian rhythms of *Arabidopsis thaliana*: I. Characterization with APRR1-overexpressing plants. Plant Cell Physiol. **43**:58–69.

Marchetti, A., M. Abril-Marti, B. Illi, G. Cesareni, and S. Nasi. 1995. Analysis of the Myc and Max interaction specificity with lambda repressor-HLH domain fusions. J. Mol. Biol. **248**:541–550.

Martinez-Garcia, J., E. Huq, and P. Quail. 2000. Direct targeting of light signals to a promoter element-bound transcription factor. Science **288**:859–863.

Menkens, A. E., U. Schindler, and A. R. Cashmore. 1995. The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins. Trends Biochem. Sci. **20**:506–510.

Murre, C., G. Bain, M. A. van Dijk, I. Engel, B. A. Furnari, M. E. Massari, J. R. Matthews, M. W. Quong, R. R. Rivera, and M. H. Stuiver. 1994. Structure and function of helix–loop–helix proteins. Biochim. Biophys. Acta **1218**:129–135.

Murre, C., P. S. McCaw, and D. Baltimore. 1989. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, *daughterless, MyoD*, and *myc* proteins. Cell **56**:777–783.

Nambu, J. R., J. O. Lewis, K. A. Wharton Jr., and S. T. Crews. 1991. The Drosophila single-minded gene encodes a helix–loop–helix protein that acts as a master regulator of CNS midline development. Cell **67**:1157–1167.

Nesi, N., I. Debeaujon, C. Jond, G. Pelletier, M. Caboche, and L. Lepiniec. 2000. The TT8 gene encodes a basic helix–loop–helix domain protein required for expression of DFR and BAN genes in arabidopsis siliques. Plant Cell **12**:1863–1878.

Neuffer, M. G., E. H. Coe Jr., and S. R. Wessler. 1997. Mutants of maize. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Ni, W., T. Fahrendorf, G. M. Ballance, C. Lamb, and R. A. Dixon. 1996. Stress responses in alfalfa (Medicago sativa L.). XX. Transcriptional activation of phenylpropanoid pathway genes in elicitor-induced cell suspension cultures. Plant Mol. Biol. **30**:427–438.

Ni, M., J. M. Tepperman, and P. H. Quail. 1998. PIF3, a phytochrome-interacting factor necessary for normal photoinduced signal transduction, is a novel basic helix–loop–helix protein. Cell **95**:657–667.

Payne, C. T., F. Zhang, and A. M. Lloyd. 2000. *GL3* encodes a bHLH protein that regulates trichome development in Arabidopsis through interaction with GL1 and TTG1. Genetics **156**:1349–1362.

Petroni, K., E. Cominelli, G. Consonni, G. Gusmaroli, G. Gavazzi, and C. Tonelli. 2000. The developmental expression of the maize regulatory gene Hopi determines germination-dependent anthocyanin accumulation. Genetics **155**:323–336.

Pickett, F. B., and D. R. Meeks-Wagner. 1995. Seeing double: appreciating genetic redundancy. Plant Cell **7**:1347–1356.

Ptashne, M. 1988. How eukaryotic transcriptional activators work. Nature **335**:683–689.

Quattrocchio, F., J. F. Wing, K. van der Woude, J. N. M. Mol, and R. Koes. 1998. Analysis of bHLH and MYB domain proteins: species specific regulatory differences are caused by divergent evolution of target anthocyanin genes. Plant J. **13**:475–488.

Quong, M. W., M. E. Massari, R. Zwart, and C. Murre. 1993. A new transcriptional-activation motif restricted to a class of helix–loop–helix proteins is functionally conserved in both yeast and mammalian cells. Mol. Cell. Biol. **13**:792–800.

Radicella, J. P., D. Turks, and V. L. Chandler. 1991. Cloning and nucleotide sequence of a cDNA encoding B-Peru, a regulatory protein of the anthocyanin pathway in maize. Plant Mol. Biol. **17**:127–130.

Rajani, S., and V. Sundaresan. 2001. The Arabidopsis myc/bHLH gene ALCATRAZ enables cell separation in fruit dehiscence. Curr. Biol. **11**:1914–1922.

Rerie, W. G., K. A. Feldmann, and M. D. Marks. 1994. The *GLABRA2* gene encodes a homeo domain protein required for normal trichome development in *Arabidopsis*. Genes Dev. **8**:1388–1399.

Riechmann, J. L., and E. M. Meyerowitz. 1998. The AP2/EREBP family of plant transcription factors. Biol. Chem. **379**:633–646.

Riechmann, J. L., and O. J. Ratcliffe. 2000. A genomic perspective on plant transcription factors. Curr. Opin. Plant Biol. **3**:423–434.

Romero, I., A. Fuertes, M. J. Benito, J. M. Malpical, A. Leyva, and J. Paz-Ares. 1998. More than 80 *R2R3-MYB* regulatory genes in the genome of *Arabidopsis thaliana*. Plant J. **14**:273–284.

Sainz, M. B., E. Grotewold, and V. L. Chandler. 1997. Evidence for direct activation of an anthocyanin promoter by the maize C1 protein and comparison of DNA binding by related Myb domain proteins. Plant Cell **9**:611–625.

Smolen, G. A., L. Pawlowski, S. E. Wilensky, and J. Bender. 2002. Dominant alleles of the basic helix–loop–helix transcription factor ATR2 activate stress-responsive genes in Arabidopsis. Genetics **161**:1235–1246.

Soh, M. S., Y. M. Kim, S. J. Han, and P. S. Song. 2000. REP1, a basic helix–loop–helix protein, is required for a branch pathway of phytochrome A signaling in arabidopsis. Plant Cell **12**:2061–2074.

Spelt, C., F. Quattrocchio, J. N. M. Mol, and R. Koes. 2000. Anthocyanin1 of petunia encodes a basic helix–loop–helix protein that directly activates transcription of structural anthocyanin genes. Plant Cell **12**:1619–1631.

Stracke, R., M. Werber, and B. Weisshaar. 2001. The R2R3-MYB gene family in Arabidopsis thaliana. Curr. Opin. Plant Biol. **4**:447–456.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. Clustal_W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Trezzini, G. F., A. Horrichs, and I. E. Somssich. 1993. Isolation of putative defense-related genes from *Arabidopsis thaliana* and expression in fungal elicitor-treated cells. Plant Mol. Biol. **21**:385–389.

Urao, T., K. Yamaguchishinozaki, N. Mitsukawa, D. Shibata, and K. Shinozaki. 1996. Molecular cloning and characterization of a gene that encodes a MYC-related protein in Arabidopsis. Plant Mol. Biol. **32**:571–576.

Vision, T. J., D. G. Brown, and S. D. Tanksley. 2000. The origins of genomic duplications in Arabidopsis. Science **290**:2114–2117.

Zeng, H., Z. Qian, M. P. Myers, and M. Rosbash. 1996. A light-entrainment mechanism for the Drosophila circadian clock. Nature **380**:129–135.

Zhang, F., A. Gonzalez, M. Zhao, and A. M. Lloyd. 2001. An enhancer of Glabra3 defines roles for bHLH proteins in all TTG1-dependent pathways. 12th International Conference on Arabidopsis Research, Madison, Wis.

Zhu, Y., J. M. Tepperman, C. D. Fairchild, and P. H. Quail. 2000. Phytochrome B binds with greater apparent affinity than phytochrome A to the basic helix–loop–helix factor PIF3 in a reaction requiring the PAS domain of PIF3. Proc. Natl. Acad. Sci. USA **97**:13419–13424.

William Taylor, Associate Editor

Accepted December 23, 2002