The Plant Cell, Vol. 15, 2497–2501, November 2003, www.plantcell.org © 2003 American Society of Plant Biologists

LETTER TO THE EDITOR

Update on the Basic Helix-Loop-Helix Transcription Factor Gene Family in *Arabidopsis thaliana*

Basic helix-loop-helix (bHLH) transcription factors represent a family of proteins that contain a bHLH domain, a motif involved in binding DNA. Recently, two groups independently analyzed the *BHLH* gene family of *Arabidopsis thaliana* (Heim et al., 2003; Toledo-Ortiz et al., 2003). These analyses revealed that this family is one of the largest transcription factor gene families in *Arabidopsis thaliana*. Although both analyses intended to give complete overviews of *AtBHLH* genes, some discrepancies were detected when the data sets were compared. After careful re-examination, we have resolved these discrepancies. In Table 1, we provide a uniform nomenclature for all of the genes that are mentioned in our two articles, and we encourage the use of this nomenclature in future reports concerning bHLH domain transcription factors (e.g., *AtBHLH042*/*TT8*).

Cross-referencing between the two data sets and further analysis have extended the total number of detected *AtBHLH* genes to 162 (Table 1). We assume that this count is very close to the final number of *AtBHLH* genes present in the *Arabidopsis thaliana* genome, but clearly, corrections or additions to the "complete" *Arabidopsis thaliana* genome sequence in the future still may cause this number to change. During examination and comparison of the data sets, we observed some common problems that contributed to the discrepancies. These problems arise commonly during the handling of large data sets and are discussed here to aid future attempts at gene family annotation. The main reasons for discrepancies were as follows.

(1) Differences between TIGR (www.tigr.org) or TAIR (www.arabidopsis.org) and MIPS (MAtDB; mips.gsf.de/projects/plants). Such differences are not easy to avoid, despite the best efforts of the database providers. Most problematic are differences in Arabidopsis Genome Initiative

Table 1. Summary of the *AtBHLH* Genes Detected

Species[a]	Generic Name	AGI Gene Code	Entry Number[b]	Synonym(s)	Accession Number[c]	Reference[d]
At	BHLH001	At5g41315	31	GL3	AF246291	Payne et al., 2000
At	BHLH002	At1g63650	30	EGL1/EGL3/AtMYC146	AF027732	Zhang et al., 2003
At	BHLH003	At4g16430	34		AF251688	
At	BHLH004	At4g17880	37	AtMYC4	AF251689	Abe et al., 2003
At	BHLH005	At5g46760	36	ATR2/AtMYC3	AF251690	Smolen et al., 2002
At	BHLH006	At1g32640	38	AtMYC2/RAP1	X99548	Abe et al., 2003
At	BHLH007	At1g03040	92		AF251692	
At	BHLH008	At1g09530	100	PIF3	AF251693	Ni et al., 1998
At	BHLH009	At2g43010	102	PIF4	AF251694	Huq and Quail, 2002
At	BHLH010	At2g31220	23		AF251695	
At	BHLH011	At4g36060	137		AF251696	
At	BHLH012	At4g00480	58	AtMYC1	AF251697	Urao et al., 1996
At	BHLH013	At1g01260	39	Myc7E	AY120752	GenBank entry[e]
At	BHLH014	At4g00870	33		AJ519812	
At	BHLH015	At2g20180	101	PIL5	AF488560	Yamashino et al., 2003
At	BHLH016	At4g00050	108		AF488561	
At	BHLH017	At2g46510	35		AY094399	
At	BHLH018	At2g22750	28		AF488562	
At	BHLH019	At2g22760	26		AF488563	
At	BHLH020	At2g22770	27		AF488564	
At	BHLH021	At2g16910	48	AMS	AF488565	Sorensen et al., 2003
At	BHLH022	At4g21330	49		NM_118253	
At	BHLH023	At4g28790	107		AF488566	
At	BHLH024	At4g36930	99	SPATULA	AF319540	Heisler et al., 2001
At	BHLH025	At4g37850	29		AF488567	
At	BHLH026	At1g02340	68	HFR1	AF488568	Fairchild et al., 2000
At	BHLH027	At4g29930	42		AF488569	
At	BHLH028	At5g46830	40		AF252636	
At	BHLH029	At2g28160	43		AF488570	
At	BHLH030	At1g68810	53		AY072161	
At	BHLH031	At1g59640	88	ZCW32	AB028232	GenBank entry[e]
At	BHLH032	At3g25710	54		AF488571	
At	BHLH033	At1g12860	44		AF488572	
At	BHLH034	At3g23210	135		AF488573	
At	BHLH035	At5g57150	41		AF488574	
At	BHLH036	At5g51780	6		AF488575	
At	BHLH037	At3g50330	117		NM_114893	
At	BHLH038	At3g56970	8	ORG2	AF488576	Kang et al., 2003
At	BHLH039	At3g56980	9	ORG3	AF488577	Kang et al., 2003
At	BHLH040	At4g00120	120		AF488578	
At	BHLH041	At5g56960	51		NM_125078	
At	BHLH042	At4g09820	32	TT8	AJ277509	Nesi et al., 2000
At	BHLH043	At5g09750	119		NM_121012	
At	BHLH044	At1g18400	77	BEE1	AF488579	Friedrichsen et al., 2002
At	BHLH045	At3g06120	20		AF488580	
At	BHLH046	At5g08130	126		AF488581	
At	BHLH047	At3g47640	139		AF488582	
At	BHLH048	At2g42300	97		AF488583	

Continued

LETTER TO THE EDITOR

(AGI) codes for the same gene between the different databases.

(2) Positions on pseudochromosomes that are not stable as a result of corrections in single BAC sequences that affect the entire area "downstream" of the corrected locus.

(3) BAC identifiers and BAC sequence coordinates that differ for the same gene when either the upper or the lower strand is considered. One option is to keep the gene orientation according to the direction of transcription; the other is to keep the original BAC sequence in its 5′ to 3′ arrangement. Clearly consistency is very important.

(4) Genes located at BAC borders that can result in either double entries of the same gene or failure to detect the gene as a result of the destruction of a continuous signature pattern.

(5) Sequence errors in the genome sequence that destroy open reading frames.

(6) Differences in the detailed definition of what constitutes a bHLH domain.

Both studies started with a subset of known bHLH domain transcription factors and used a consensus sequence described by Atchley et al. (1999) as a reference. However, whereas one analysis was based on bHLH proteins similar to *Zea mays* Sn (e.g., *Zm*R) that are involved in secondary metabolism and cell identity pathways (Heim et al., 2003), the other used a subset based on PHYTOCHROME-INTERACTING FACTOR3 (PIF3) as a starting point (Toledo-Ortiz et al., 2003). In addition, the set of databases used was not completely overlapping. Consequently, some genes were identified as encoding true bHLHs by one group but not by the other, and vice versa. These differences have been removed; there are now only two *BHLH* genes listed in Table 1 (*AtBHLH136*/At5g39860 and *AtBHLH160*/At1g71200) that fit the criteria of Heim et al. (2003) but not those of Toledo-Ortiz et al. (2003). A third article analyzing plant bHLH domain proteins ap-

Table 1. (continued).

Species[a]	Generic Name	AGI Gene Code	Entry Number[b]	Synonym(s)	Accession Number[c]	Reference[d]
At	BHLH049	At1g68920	82		AF488584	
At	BHLH050	At1g73830	76	BEE3	AF488585	Friedrichsen et al., 2002
At	BHLH059	At4g02590	93		AF488592	
At	BHLH060	At3g57800	91		AF488593	
At	BHLH061	At5g10570	46		AF488594	
At	BHLH062	At3g07340	85		AF488595	
At	BHLH063	At4g34530	84		AF488596	
At	BHLH064	At2g18300	79		AF488597	
At	BHLH065	At3g59060	103	PIL6	AF488598	Yamashino et al., 2003
At	BHLH066	At2g24260	95		AF488599	
At	BHLH067	At3g61950	11		AF488600	
At	BHLH068	At4g29100	60		AF488634	
At	BHLH069	At4g30980	94		AF488601	
At	BHLH070	At2g46810	13		AF488602	
At	BHLH071	At5g46690	17		AF488603	
At	BHLH072	At5g61270	109		AF488604	
At	BHLH073	At5g67110	98	ALCATRAZ	AF488605	Rajani and Sundaresan, 2001
At	BHLH074	At1g10120	90		AF488606	
At	BHLH075	At1g25330	78		AF488607	
At	BHLH076	At1g26260	83		AF488608	
At	BHLH077	At3g23690	87		AF488609	
At	BHLH078	At5g48560	86		AF488610	
At	BHLH079	At5g62610	81		AF488611	
At	BHLH080	At1g35460	71		AF488612	
At	BHLH081	At4g09180	72		AF488613	
At	BHLH082	At5g58010	96		AF488614	
At	BHLH083	At1g66470	112		AF488615	
At	BHLH084	At2g14760			AJ577584	
At	BHLH085	At4g33880	115		AF488616	
At	BHLH086	At5g37800	113		NM_123139	
At	BHLH087	At3g21330	121		AF488617	
At	BHLH088	At5g67060	118		AF488618	
At	BHLH089	At1g06170	24		AF488619	
At	BHLH090	At1g10610	50		AF488620	
At	BHLH091	At2g31210	25		AJ519809	
At	BHLH092	At5g43650	22		AY065390	
At	BHLH093	At5g65640	47		AF488621	
At	BHLH094	At1g22490	16		AF488622	
At	BHLH095	At1g49770	21		AF488623	
At	BHLH096	At1g72210	15		AJ459771	
At	BHLH097	At3g24140	14		AF488624	
At	BHLH098	At5g53210	19		NM_124700	
At	BHLH099	At5g65320	18		AF488625	
At	BHLH100	At2g41240	7		AF488626	
At	BHLH101	At5g04150	10		AJ519810	
At	BHLH102	At1g69010	125		AF488627	
At	BHLH103	At4g21340	62		AY065362	
At	BHLH104	At4g14410	136		AF488628	
At	BHLH105	At5g54680	133		AF488629	
At	BHLH106	At2g41130	56		AY074639	
At	BHLH107	At3g56770	55		NM_115536	
At	BHLH108	At1g25310	132		NM_102341	
At	BHLH109	At1g68240			AJ577585	
At	BHLH110	At1g27660	59		NM_102531	
At	BHLH111	At1g31050	66		AA395190	

Continued

peared recently (Buck and Atchley, 2003) reporting ∼118 *AtBHLH* genes. Of these, 116 correspond to those listed in Table 1. The remaining two (At1g49830 and At5g33210) do not fit the criteria used for Table 1.

Search engines have been greatly improved in the last few years, but they still often are not exact enough to identify certain motifs. This is not necessarily the result of deficiencies in the search algorithms but may result from the structure of matrices that describe known motifs (e.g., *AtBHLH125* spanned two separate BAC ends, and two separate predictions had to be fused). Even the continuous optimization of our bHLH domain matrix never resulted in the identification of all 162 *AtBHLH* genes in one search. Additionally, gene prediction tools are sometimes not flexible enough to respond to variable intron lengths and exon distribution (e.g., the prediction NM_105789 for *AtBHLH160* contains an intron that causes an overestimate of the length of the loop structure). It sounds obvious, but it is worth emphasizing that cDNA sequences, even from reverse transcriptase–mediated PCR experiments, should be deposited in GenBank (http://www.ncbi.nlm.nih.gov/) or EMBL (http://www.ebi.ac.uk/Databases/) even if the genomic sequence is already in the database, and the "metadata" of the database entry should be written with care. The most unambiguous identifier of any given gene (unless a sequence-identical duplication exists) is its DNA sequence, and only this information allows designations and identifier assignments to be checked and rechecked.

It is an interesting and critical point that even with a combination of all available BLAST (Basic Local Alignment Search Tool) tools, both groups were unable to obtain a full set of Arabidopsis bHLH domain transcription factors in their initial analyses. Both studies relied on BLAST search capabilities (TBLASTN and BLASTP) and subsequent evaluation of the hits for the respective bHLH consensus sequences. In addition, position-specific iterated BLAST was used by one of the two groups to identify remaining unidentified bHLH domain–encoding sequences. Nevertheless, several true *BHLH*

Table 1. (continued).

Species[a]	Generic Name	AGI Gene Code	Entry Number[b]	Synonym(s)	Accession Number[c]	Reference[d]
At	BHLH112	At1g61660	64		AF488630	
At	BHLH113	At3g19500	61		AF488631	
At	BHLH114	At4g05170	65		NM_116756	
At	BHLH115	At1g51070	134		AF488632	
At	BHLH116	At3g26744	45	ICE1	AY079016	Chinnusamy et al., 2003
At	BHLH117	At3g22100	140		NM_113106	
At	BHLH118	At4g25400	5		NM_118672	
At	BHLH119	At4g28811	104		AJ519811	
At	BHLH120	At5g51790	4		NM_124558	
At	BHLH121	At3g19860	138		AF488633	
At	BHLH122	At1g51140	70		AY063120	
At	BHLH123	At3g20640	63		AU238908	
At	BHLH124	At2g46970	110	PIL1	AB090873	Yamashino et al., 2003
At	BHLH125	At1g62975	2		AF506369	
At	BHLH126	At4g25410	3		Z46563	
At	BHLH127	At4g28815			AJ577586	
At	BHLH128	At1g05805	74		AY045907	
At	BHLH129	At2g43140	73		AU237473	
At	BHLH130	At2g42280	69		NM_129790	
At	BHLH131	At4g38071			AJ577587	
At	BHLH132	At3g62090	111	PIL2	AB090874	Yamashino et al., 2003
At	BHLH133	At2g20095			AJ577588	
At	BHLH134	At5g15160	52		AK118887	
At	BHLH135	At1g74500	67		AY088286	
At	BHLH136	At5g39860			AY088246	
At	BHLH137	At5g50915	89		AY087602	
At	BHLH138	At2g31215			NM_179830	
At	BHLH139	At5g43175	116		NM_148080	
At	BHLH140	At5g01310	122		NM_120209	
At	BHLH141	At5g38860	127		NM_123247	
At	BHLH142	At5g64340	128		AY062561	
At	BHLH143	At5g09460	129		BT000009	
At	BHLH144	At1g29950	130		AF361607	
At	BHLH145	At5g50010	131		BT005301	
At	BHLH146	At4g30180	141		AU237244	
At	BHLH147	At3g17100	142		NM_180270	
At	BHLH148	At3g06590	143		NM_111535	
At	BHLH149	At1g09250	144		BT003052	
At	BHLH150	At3g05800	145		NM_111454	
At	BHLH151	At2g47270	146		NM_130295	
At	BHLH152	At1g22380	147		NM_102088	
At	BHLH153	At1g05710			AJ576040	
At	BHLH154	At2g31730			AJ576041	
At	BHLH155	At2g31280			AJ576042	
At	BHLH156	At2g27230			AJ576043	
At	BHLH157	At1g64625			AJ576044	
At	BHLH158	At2g43060			AJ576045	
At	BHLH159	At4g30410			AJ576046	
At	BHLH160	At1g71200			NM_105789	
At	BHLH161	At3g47710			NM_114639	
At	BHLH162	At4g20970			NM_118215	

[a] The prefix At indicates *Arabidopsis thaliana* (see text).

[b] BHLH "entry numbers" (Toledo-Ortiz et al., 2003).

[c] GenBank accession number of the cDNA sequence representing the open reading frame used to evaluate the presence or absence of a proper bHLH domain signature.

[d] References for the synonyms that are used in the literature.

[e] The synonym was found only in a GenBank entry but not in an article.

genes were not detected. Some of these initial false negatives were found by searching for the term "helix-loop-helix" in the annotation databases (e.g., *AtBHLH134* and *AtBHLH136*). However, this search also resulted in many false positives that had to be excluded as a result of misannotations based on weak homology or of "inherited misannotation," in which a single wrong annotation text had been used as a reference during annotation. In essence, we were unable to detect slightly divergent or mispredicted *BHLH* genes. The only solution to this problem may involve systematic annotation by expert annotators, comprehensive EST data production from normalized libraries, and the generation of full-length cDNA at least for protein-coding gene sequences. A significant part of the improvement of the data set presented in Table 1 is based on the reannotation of the Arabidopsis genome by the TIGR group, which followed this approach.

We were able to improve gene annotation further by comparing closely related *BHLH* genes for their exon/intron structures. This powerful similarity-based approach (used here within a single species) led to the correction of some gene annotations and, consequently, to a further increase in the total number of *AtBHLH* genes detected. Several of the genes that escaped the initial screens by both groups contain short introns in the region that encodes the loop of the HLH region. These comparably short introns, and also short exons that are part of the bHLH open reading frame, resulted in mispredictions that were a significant cause of false negatives in our initial analyses. One example is *AtBHLH160*, for which we found a formerly unpredicted intron after comparison with the most closely related genes *AtBHLH038*/*ORG2*, *AtBHLH039*/*ORG3*, *AtBHLH100*, and *AtBHLH101*.

The combined effort of our two groups and the lessons we have learned from the comparison of the two data sets have resulted in an (almost) complete view of the *AtBHLH* transcription factor gene family, now provided with unambiguous generic names and reference to synonyms. We hope that this work will serve as a solid foundation for further investigations into the functions of the different members of this interesting gene family in plants.

Paul C. Bailey and Cathie Martin
John Innes Centre
Colney Lane
NR4 7UH Norwich, UK

Gabriela Toledo-Ortiz and Peter H. Quail
Department of Plant and
Microbial Biology
University of California
Berkeley, CA 94720
and United States Department
of Agriculture
Agricultural Research Service Plant
Gene Expression Center
Albany, CA 94710

Enamul Huq
Section of Molecular Cell and
Developmental Biology
University of Texas
1 University Station, A6700
Austin, TX 78712

Marc A. Heim, Marc Jakoby,
and Martin Werber
Max-Planck-Institute for Plant
Breeding Research,
50829 Köln, Germany

Bernd Weisshaar
Institute for Genome Research,
Bielefeld University,
33594 Bielefeld, Germany

REFERENCES

Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2003). Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. Plant Cell **15,** 63–78.

Atchley, W.R., Terhalle, W., and Dress, A. (1999). Positional dependence, cliques, and predictive motifs in the bHLH protein domain. J. Mol. Evol. **48,** 501–516.

Buck, M.J., and Atchley, W.R. (2003). Phylogenetic analysis of plant basic helix-loop-helix proteins. J. Mol. Evol. **56,** 742–750.

Chinnusamy, V., Ohta, M., Kanrar, S., Lee, B.H., Hong, X., Agarwal, M., and Zhu, J.K. (2003). ICE1: A regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis. Genes Dev. **17,** 1043–1054.

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

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

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.

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

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

Kang, H.G., Foley, R.C., Onate-Sanchez, L., Lin, C., and Singh, K.B. (2003). Target genes for OBP3, a Dof transcription factor, include novel basic helix-loop-helix domain proteins inducible by salicylic acid. Plant J. **35,** 362–372.

Nesi, N., Debeaujon, I., Jond, C., Pelletier, G., Caboche, M., and Lepiniec, L. (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.

Ni, M., Tepperman, J.M., and Quail, P.H. (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., Zhang, F., and Lloyd, A.M. (2000). *GL3* encodes a bHLH protein that regulates trichome development in Arabidopsis through interaction with GL1 and TTG1. Genetics **156,** 1349–1362.

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

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

Sorensen, A.M., Krober, S., Unte, U.S., Huijser, P., Dekker, K., and Saedler, H. (2003). The Arabidopsis ABORTED MICROSPORES (AMS) gene encodes a MYC class transcription factor. Plant J. **33,** 413–423.

Toledo-Ortiz, G., Huq, E., and Quail, P.H. (2003). The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell **15,** 1749–1770.

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

Yamashino, T., Matsushika, A., Fujimori, T., Sato, S., Kato, T., Tabata, S., and Mizuno, T. (2003). A link between circadian-controlled bHLH factors and the APRR1/TOC1 quintet in *Arabidopsis thaliana*. Plant Cell Physiol. **44,** 619–629.

Zhang, F., Gonzalez, A., Zhao, M., Payne, C.T., and Lloyd, A.M. (2003). A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. Development **130,** 4859–4869.