The Basic Helix-Loop-Helix Transcription Factor Family in the Pea Aphid, Acyrthosiphon pisum

Authors: Dang, Chun-Wang, Wang, Yong, Chen, Ke-Ping, Yao, Qing, Zhang, De-Bao, et al.

Source: Journal of Insect Science, 11(84) : 1-11

Published By: Entomological Society of America

URL: https://doi.org/10.1673/031.011.8401

Journal of Insect Science
www.insectscience.org

The basic helix-loop-helix transcription factor family in the pea aphid, *Acyrthosiphon pisum*

Chun-Wang Dang[1a], Yong Wang[2b*], Ke-Ping Chen[1c], Qing Yao[1d], De-Bao Zhang[1e], Min Guo[1f]

[1]Institute of Life Sciences, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, P. R. China
[2]School of Food and Biological Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, P. R. China

Abstract

The basic helix-loop-helix (bHLH) proteins play essential roles in a wide range of developmental processes in higher organisms. bHLH family members have been identified in over 20 organisms, including fruit fly, zebrafish, and human. This study identified 54 bHLH family members in the pea aphid, *Acyrthosiphon pisum* (Harris) (Hemiptera: Aphididae), genome. Phylogenetic analyses revealed that they belong to 37 bHLH families with 21, 13, 9, 1, 9, and 1 members in group A, B, C, D, E, and F, respectively. Through in-group phylogenetic analyses, all of the identified *A. pisum* bHLH members were assigned into their correspondent bHLH families with confidence, among which 51 were defined according to phylogenetic analyses with orthologs from *Drosophila melanogaster* Meigen (Diptera: Drosophilidae), and 3 of them were defined according to phylogenetic analyses with orthologs from *Bombyx mori* L. (Lepidoptera: Bombycidae) and *Tribolium castaneum* (Herbst) (Coleoptera: Tenebrionidae). Analyses on genomic coding regions revealed that the number and average length of introns in *A. pisum* bHLH motifs are higher than those in other insects. The present study provides useful background information for future studies on structure and function of bHLH proteins in the regulation of *A. pisum* development.

Keywords: blast search, orthologous family, phylogenetic analysis
Abbreviations: **Ap**, *Acyrthosiphon pisum*; **Am**, *Apis mellifera*; **Bm**, *Bombyx mori* ; **Tc**, *Tribolium castaneum*
Correspondence: [a] chunwang521123@163.com, [b*] ywang@ujs.edu.cn, [c] kpchen@ujs.edu.cn, [d] yaoqin@ujs.edu.cn, [e] ff_1113a@126.com, [f] guomin20042008@126.com, *Corresponding author
Editor: Igor Sharakhov was Editor of this paper
Received: 25 November 2010, **Accepted:** 20 December 2010

ISSN: 1536-2442 | Vol. 11, Number 84

Cite this paper as:
Dang C-W, Wang Y, Chen K-P, Zhang D-B, Guo M. 2011. The basic helix-loop-helix transcription factor family in the pea aphid, *Acyrthosiphon pisum*. *Journal of Insect Science* 11:84 available online: insectscience.org/11.84

Introduction

The basic helix-loop-helix (bHLH) proteins form a large superfamily of transcription factors that play important roles in a wide range of developmental processes including neurogenesis, myogenesis, hematopoiesis, sex determination, and gut development. The bHLH domain is approximately 60 amino acids long and comprises a DNA-binding basic region (b) and two helices separated by a variable loop region (HLH) (Massari and Murre 2000). The HLH domain promotes dimerization, allowing the formation of homodimeric or heterodimeric complexes between different family members. The two basic domains which are brought together through dimerization bind specific hexanucleotide sequences.

Since the first characterization of the murine bHLH transcription factors E12 and E47 (Murre et al. 1989), Atchley et al. (1999) developed a predictive motif for the bHLH domains based on amino acid frequencies at all positions of 242 bHLH proteins, among which 19 sites were highly conserved in all the organisms. With the completion of genome sequencing projects for an increased number of organisms, over one thousand bHLH family members have been identified in organisms whose genome sequences were available. These include 8 bHLH genes in *Saccharomyces cerevisiae*, 16 in *Amphimedon queenslandica*, 33 in *Hydra magnipapillata*, 33 in *Caenorhabditis elegans*, 104 in *Gallus gallus*, 46 in *Ciona intestinalis*, 50 in *Strongylocentrotus purpuratus*, 51 in *Apis mellifera*, 52 in *Bombyx mori*, 57 in *Daphia pulex*, 59 in *Drosophila melanogaster*, 63 in *Lottia gigantea*, 64 in *Capitella* sp 1, 68 in *Nematodtella vectensis*, 78 in *Branchiostoma floridae*, 87 in *Tetraodon nigroviridis*, 114 in *Mus musculus*, 118 in *Homo sapiens*, 139 in *Brachydanio rerio*, 147 in *Arabidopsis*, and 167 in *Oryza sativa* (Zheng et al. 2009; Li et al. 2006; Satou et al. 2003; Simionato et al. 2007; Toledo-Ortiz et al. 2003; Wang et al. 2007, 2008, 2009).

Based on phylogenetic analyses to the available bHLH proteins, Ledent and Vervoort (2001) defined 44 orthologous families and 6 higher-order groups for bHLH proteins, among which 36 include bHLH from animals only, two have representatives in both yeasts and animals, two are present only in yeast, and four are present only in plants. They named the 44 families according to their first reported names, common abbreviations, or their best-known members of the family. And the higher-order groups were named A, B, C, D, E, and F based on their different DNA-binding properties of these groups. Group A and B include bHLH proteins that bind hexameric DNA sequences referred to as "E boxes" (CANNTG), in which group A binds to CACCTG or CAGCTG and group B binds to CACGTG or CATGTTG (Murre et al. 1989; Van Doren et al. 1991; Dang et al. 1992). Group C corresponds to the family of bHLH proteins known as bHLH-PAS which is about 260–310 amino acids long (Crews 1998). bHLH-PAS proteins bind the core sequence of ACGTG or GCGTG. Group D corresponds to HLH proteins that lack a basic domain. They form inactive heterodimers with group A proteins. Group E corresponds to the family of bHLH proteins which bind preferentially to sequences typical of N boxes (CACGCG or CACGAG). They also contain one additional Orange domain and one WRPW peptide in their carboxyl terminus. Group F corresponds to the family of bHLH proteins that have the COE domain which has

an additional domain involved in both dimerization and DNA binding (Ledent and Vervoort 2001).

Ledent et al. (2002) defined 44 families for bHLH proteins from animals only, among which 20, 12, 7, 1, 3, and 1 families are included in groups A, B, C, D, E, and F, respectively. In 2007, it was found that the MyoR family could be expanded into three families, i.e. MyoRa , MyoRb, and Delilah, and the originally separated families, Hairy and E(spl), needed to be combined into one family, H/E(spl), due to insufficient evidence from the phylogenetic analyses (Simionato et al. 2007). Therefore, at present, animal bHLH proteins are classified into 45 families.

The pea aphid, *Acyrthosiphon pisum* (Harris) (Hemiptera: Aphididae), is the primary aphid species used in laboratory and genetic studies. *A. pisum* has been intensively studied as a model for understanding bacterial endo-symbiosis, phenotypic plasticity, clonal vs. sexual reproduction, and the development of resistance to pesticides (Wilson et al. 2010; Srinivasan et al. 2010). bHLH proteins are important transcription factors with regulatory functions in various developmental processes in eukaryotes. Identification of bHLH protein members encoded in the *A. pisum* genome will facilitate studies on gene structure and function involved in regulation of *A. pisum* development. However, there have been no reports on identification and characterization of *bHLH* genes in *A. pisum*. In this study, amino acid sequences of 59 *D. melanogaster* Meigen (Diptera: Drosophilidae) bHLH motifs were used to conduct tblastn searches against *A. pisum* genome sequences (http://www.ncbi.nlm.nih.gov/genomeprj/13646) to obtain candidate bHLH members in *A. pisum*. Subsequent examination and analyses led to successful identification of 54 bHLH members in *A. pisum* and definition of orthologous families for them with sufficient confidence. Moreover, it was found that the number and average length of introns in *A. pisum* bHLH motifs are higher than those in other insects. These results provide useful background information for future studies on structure and function of bHLH proteins in the regulation of *A. pisum* development.

Materials and Methods

Tblastn searches

Amino acid sequences of 59 *D. melanogaster* bHLH motifs were obtained from the additional files of previous reports (Ledent and Vervoort 2001; Simionato et al. 2007). Each sequence was used as query sequence to perform tblastn searches against the *A. pisum* genome sequences. The expected value (*E*) was set at 10 in order to obtain all bHLH related sequences. The obtained subject sequences were manually examined to keep only one sequence for those that have the same contig number, reading frame, and coding regions; to add the missing amino acids to corresponding sites by EditSeq program (version 5.01) of the DNAStar package; and to find introns within the bHLH motifs. Intron analysis was done using NetGene2 application online (http://www.cbs.dtu.dk/services/NetGene2/).

Sequence alignment

All sequences that had been improved by the above methods were aligned using MEGA4 (Tamura et al. 2007) built-in ClustalW program (version 4.0) with default settings. Each sequence was examined for their amino acid residues at the 19 conserved sites by manual checking. Sequences with less than nine variations were regarded as potential ApbHLH (*A. pisum* bHLH) members. The sequences which have less than ten

conservations were discarded and the rest sequences were aligned again using ClustalW. The aligned ApbHLH motifs were shaded in GeneDoc Multiple Sequence Alignment Editor and Shading Utility (Version 2.6.02) (Nicholas et al.1997) and copied to rich text file for further annotation.

Phylogenetic analyses

Phylogenetic analyses to all the identified ApbHLH members were carried out in two steps. First, all obtained ApbHLH motif sequences were used to build neighbor-joining (NJ) distance tree with the 59 *D. melanogaster* bHLH motif sequences using PAUP 4.0 Beta 10 (Swofford 1998) based on a step matrix constructed from Dayhoff PAM 250 distance matrix by R. K. Kuzoff (http://paup.csit.fsu.edu/). Then, each ApbHLH motif sequence was used to conduct in-group phylogenetic analyses (Wang et al. 2007) with *D. melanogaster* bHLH motif sequences. That is, each amino acid sequence of *A. pisum* bHLH motifs was used to construct NJ, maximum parsimony (MP), and maximum likelihood (ML) phylogenetic trees with *D. melanogaster* bHLH family members of the corresponding group, respectively. The NJ trees were bootstrapped with 1000 replicates to provide information about their statistical reliability. MP analysis was performed using heuristic searches and bootstrapped with 100 replicates. ML trees were constructed using TreePuzzle 5.2 (Schmidt et al. 2002) with quartet-puzzling tree-search procedure and 25,000 puzzling steps. Model of substitution was set to the Jones-Taylor-Thornton (Jones et al. 1992). Other parameters were set to default values.

Results and Discussion

Identification of ApbHLH members

The tblastn searches, sequence alignment, and examination of the 19 conserved amino acid sites revealed that there were 54 *bHLH* genes in *A. pisum* genome. The alignment of all 54 ApbHLH members is shown in Figure 1 and the phylogenetic tree constructed using amino acids from 54 ApbHLH motifs and 59 *D. melanogaster* bHLH motifs is shown in Figure 2. Figure 1 and 2 show that there were 21, 13, 9, 1, 9, and 1 ApbHLH members in group A, B, C, D, E, and F, respectively. In Figure 1, there are two most conserved sites located at sites 24 and 51 of the bHLH motif, respectively. Besides these, there are seven other sites that are also conserved (indicated with asterisks on top of Figure 1). Because the phylogenetic analyses have provided sufficient bootstrap support, the identified ApbHLH motifs were named according to nomenclature used by *D. melanogaster* bHLH sequences. In the case where one *D. melanogaster* bHLH sequence has two or more *A. pisum* homologues, the researchers used 'a', 'b', and 'c' or '1', '2', and '3' etc to number them. For instance, two homologues of the *D. melanogaster Mist, Bmx* and *Stich1*, genes were found in *A. pisum*. Therefore, these *ApbHLH* genes were named *ApMist1* and *ApMist2*, *ApBmx1* and *ApBmx2*, and *ApStich1a* and *ApStich1b*, respectively. Fifty-four ApbHLHs were named in accordance with the corresponding *D. melanogaster* and other insect homologues as listed in Table 1.

Identification of orthologous families

Ortholog identification has been very uncertain since there is no absolute criterion that can be used to decide whether two genes are orthologous (Ledent and Vervoort 2001). However, in previous studies (Wang et al. 2007, 2008) in-group phylogenetic analysis was adopted to identify homologues for the unknown sequences that would form a monophyletic clade among themselves. So a

more certain criterion was used based on the criterion used by Ledent et al. (Ledent and Vervoort 2001; Ledent et al. 2002): If an unknown single *A. pisum* bHLH forms a monophyletic clade with another bHLH of known family in phylogenetic trees constructed with different methods, and all the bootstrap values exceed 50 then the known member will be regarded as a homologue of the unknown sequence. Figure 3, as an example here, shows NJ, MP, and ML phylogenetic trees constructed with one *A. pisum* bHLH member (ApDa) and seven group A bHLH members from *D. melanogaster*. In all three trees, ApDa formed monophyletic clade with Da (daughterless) specimens of *D. melanogaster* with all bootstrap values as 100. Therefore, ApDa was considered an ortholog of Da *D. melanogaster*. Similar in-group phylogenetic analyses were conducted for each of the identified *A. pisum* bHLH members. All the bootstrap values of constructed NJ, MP, and ML trees for each of the identified *A. pisum* bHLH members were listed in Table 1 without showing the correspondent constructed trees. Table 1 showed that the orthology of *A. pisum* bHLH members with *D. melanogaster* and other insect species can be divided into the following categories:

First, among all the 54 *A. pisum* bHLH members: 32 ApbHLH members have all the bootstrap values over 50 (54 ≦ ! bootstrap values ≦ !100) in constructed NJ, MP, and ML trees except *ApMax3* of which the bootstrap value of the MP tree is 42. These 32 ApbHLHs are *ApDa*, *ApMistr1, ApMistr2, ApOli, ApNet, ApMyoR, ApDel, ApTwi, ApFer1, ApFer3, ApHand, ApSCL, ApNSCL, ApMnt, ApMax1, ApMax2, ApMax3, ApCrp, ApMLX, ApSREBP, ApTai, ApClk, ApDys, ApSs, ApSim, ApTrh, ApSima, ApTgo, ApEmc, ApStich1a, ApSide*, and *ApKn(col)*. The researchers have sufficient confidence to define the orthology of these ApbHLH motifs as corresponding to *D. melanogaster* bHLH orthologs.

Second, 5 ApbHLH members (namely *ApTap*, *ApFer2*, *ApDm*, *ApUSF*, and *ApBmx2*) have bootstrap values ranging from 77 to 99 in NJ and MP trees, except *ApDm* of which the bootstrap value of the MP tree is 45. In NJ and MP trees, each of them formed a monophyletic clade with the same *D. melanogaster* bHLH orthologue. However, they formed monophyletic clades (bootstrap value:58≦bootstrap values≦89) with other *D. melanogaster* bHLH members in ML trees. Specifically, the orthologue of *ApTap* was *tap* of *D. melanogaster* in NJ and MP trees, but was *cato* in ML trees. The orthologue of *ApFer2* was *Fer2* of *D. melanogaster* in NJ and MP trees, but was *Pxs* in ML trees. The orthologues of *ApDm*, *ApUSF*, and *ApBmx2* were *dm*, *USF*, and *bmx* of *D. melanogaster,* respectively, in NJ and MP trees, but all were *SREBP* in ML trees. The orthology for these 5 ApbHLH members has been defined according to the statistical support from NJ and MP trees.

Third, 7 ApbHLH members (namely *ApAto*, *ApSage*, *ApPxs*, *ApBmx1*, *ApHey*, *ApStich1b*, and *ApH*) formed monophyletic clades with bootstrap values ranging from 52 to 100 in NJ and MP trees, but did not form any monophyletic groups with any single bHLH sequence in ML trees (marked with n/m* or n/m in Table 1). Four other ApbHLH members (namely *ApCato*, *ApRst(1)JH*, *ApCyc*, and *ApDpn*) formed monophyletic clades with bootstrap values ranging from 45 to 96 in one of the NJ, MP, and ML trees, but did not form any monophyletic clades in the other two trees. Although these 11 ApbHLH members did not have sufficient bootstrap

support, the orthologs were defined because they each have one or two bootstrap supports to testify to their orthology to the correspondent *D. melanogaster* ortholog. This phylogenetic divergence of bHLH motif sequences between *A. pisum* and *D. melanogaster* probably means that these two insect species have evolved in quite different circumstances.

Finally, there are 6 ApbHLH sequences which did not form monophyletic clade with any *D. melanogaster* bHLH sequence in all constructed phylogenetic trees. They are *ApASCb*, *ApAtonal1*, *ApMad*, *ApHES1*, *ApHES2*, and *ApHES3* (marked with [a] or [b] in Table 1 and Figure 2). Each of them was used to conduct in-group phylogenetic analyses with corresponding sequences from 3 other insect species, namely *A. mellifera*, *B. mori*, and *Tribolium castaneum*. For example, Figure 4 shows that ApASCb formed a monophyletic clade with TcASCb with bootstrap values ranging from 78 to 99. Therefore, it was considered an ortholog of TcASCb. Similarly, ApMad was found to be an ortholog of TcMad with all bootstrap values at 100 (Table 1). Orthology of *ApHES1* could also be defined, although the bootstrap values were not sufficiently high (35 ≦! bootstrap values ≦!44) and no monophyletic calde was formed in two phylogenetic trees constructed. Orthology of *ApHES2*, *ApHES3*, and *ApAtonal1* were the least clear. It was evident that *ApHES2* and *ApHES3* belonged to the H/E(spl) family. *ApAtonal1* was clearly a member of the Atonal family. Therefore, they have been named numerically (Table 1).

Identification of protein sequences and genomic contigs

Protein sequence accession numbers for all the identified ApbHLH motifs are listed in Table 1. There are 3 ApbHLH motifs, of which, protein sequence accession numbers were not found in any protein databases. They are ApSREBP, ApDys, and ApFer2, respectively. Protein sequence accession numbers for 14 ApbHLH motifs were only found in the '*Ab initio* protein' database in which all protein sequences were predicted from their corresponding genomic sequences. ApCyc protein sequence accession number was found in 'RefSeq protein' database. The rest of the ApbHLH protein sequences accession numbers were found in 'Non-RefSeq protein' database.

The coding regions and intron analysis for 54 *A. pisum* bHLH motifs are listed in Table 2. These data indicate that there are 26 ApbHLH members with introns in their bHLH motifs, and the total number of intron is 34. Eighteen ApbHLH members have one intron, among which *ApDa, ApClk, ApTgo, ApCyc, ApStich2*, and *ApHES1* have introns in the basic region; *ApMistr1, ApMistr2*, and *ApPxs* have introns in helix 1 region; *ApASCb, ApUSF, ApCrp, ApBmx1*, and *ApSREBP* have introns in the loop region; and *ApSage, ApSCL, ApMnt*, and *ApBmx2* have introns in helix 2 region. Eight ApbHLH members have two introns, among which *ApH, ApDpn, ApSide1, ApSide2, ApHES3*, and *ApKn(col)* have introns in the basic and loop regions, *ApTai* has introns in the basic and helix 2 regions, and *ApMad* has introns in the loop and helix 2 regions. The longest intron in the *A. pisum* bHLH motif is 30,718 bp (base pairs), and the average length of intron is 4193 bp. Compared with other insect species, the number and length of introns are remarkably higher in *A. pisum.* For instance, in the *B. mori* and *Apis mellifera* bHLH motifs, there are only 12 and 9 introns with the longest introns being 7083 bp and 4460 bp, and the average length of introns being 1352 bp and 1326 bp, respectively. Also, 8 ApbHLH

motifs have two introns, while no bHLH motif has been found to have two introns in *Bombyx mori* and *A. mellifera* (Wang et al. 2007, 2008).

Conclusion

Our study identified 54 bHLH members in the *A. pisum* genome. All ApbHLH members have been defined by their names and families according to various phylogenetic analyses with bHLH homologues of *D. melanogaster*, *A. mellifera*, *B. mori*, and *T. castaneum*. Among all ApbHLH members, 48 ApbHLH members have homologues in *D. melanogaster*, and 3 ApbHLH members have homologues in *B. mori* and *T. castaneum*. Three ApbHLH motifs' protein sequence accession numbers were not found in any protein database. The researchers also found that the number and average length of introns in ApbHLH motifs are higher than those in other insect species, which is quite possibly the consequence of the insertion of increased numbers of transposable elements in the coding regions of ApbHLH proteins as revealed by the International Aphid Genomics Consortium (2010). The above results would provide useful background information for future studies on functions of bHLH proteins in the regulation of *A. pisum* development.

Acknowledgements

The authors are thankful to professor Bin Chen of Jiangsu University and to two anonymous reviewers for their constructive suggestions and comments on the manuscript. This work was supported by Scientific Research Promotion Fund for the Talents of Jiangsu University (No. 09JDG029) and Jiangsu Sci-Tech Support Project - Agriculture (No. BE2008379).

References

Atchley WR, Terhalle W, Dress A. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. *Journal of Molecular Evolution* 48: 501-516.

Crews ST. 1998. Control of cell lineage-specific development and transcription by bHLH-PAS proteins. *Genes & Development* 12: 607–620.

Dang CV, Dolde D, Gillison ML, Kato GJ. 1992. Discrimination between related DNA sites by a single amino acid residue of myc-related basic-helix-loop-helix proteins. *Proceedings of the National Academy of Science USA* 89: 599–602.

Ferre-D'Amare AR, Prendergast GC, Ziff EB, Burley SK. 1993. Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* 363: 38–45.

International Aphid Genomics Consortium 2010 Genome Sequence of the Pea Aphid *Acyrthosiphon pisum*. *PLoS Biol* 8(2): e1000313. doi:10.1371/journal.pbio.1000313.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8: 275-282.

Li X, Duan X, Jiang H, Sun Y, Tang Y, Yuan Z, Guo J, Liang W, Chen L, Yin J, Ma H, Wang J, Zhang D. 2006. Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and *Arabidopsis*. *Plant Physiology* 141: 1167-1184.

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

Ledent V, Paquet O, Vervoort M. 2002. Phylogenetic analysis of the human basic helix-loop-helix proteins. *Genome Biology* 3: R30.

Murre C, Mc Caw PS, Vaessin H, Caudy M, Jan LY, Cabrera CV, Buskin JN, Hauschka SD, Lassar AB, Weintraub H, et al. 1989. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell* 58: 537–544.

Massari ME, Murre C. 2000. Helix-Loop-Helix Proteins: Regulators of Transcription in Eucaryotic Organisms. *Molecular and Cellular Biology* 20: 429-440.

Nicholas KB, Nicholas-Jr HB, Deerfield-II DW. 1997. GeneDoc: Analysis and Visualization of Genetic Variation. *Embnet News* 4: 14.

Swofford DL. 1998. *PAUP**. *Phylogenetic Analysis Using Parsimony*, Version 4. Sinauer Associates.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18: 502-504.

Satou Y, Imai KS, Levine M, Kohara Y, Rokhsar D, Satoh N. 2003. Agenomewide survey of developmentally relevant genes in Ciona intestinalis. I. Genes for bHLH transcription factors. *Development Genes and Evolution* 213: 5–60213-221.

Simionato E, Ledent V, Richards G, Thomas-Chollier M, Kerner P, Coornaert D, Degnan BM, Vervoort M. 2007. Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. *BMC Evolutionary Biology* 7: 33.

Srinivasan DG, Fenton B, Jaubert-Possamai S, Jaouannet M. 2010.Analysis of meiosis and cell cycle genes of the facultatively asexual pea aphid, *Acyrthosiphon pisum* (Hemiptera: Aphididae). *Insect Molecular Biology* 2: 229-239.

Toledo-Ortiz G, Huq E, Quail PH. 2003. The Arabidopsis basic/helix-loop-helix transcription factor family. *Plant Cell* 15: 1749-1770.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecolar Biology and Evolution* 24: 1596–1599.

Van Doren M, Ellis HM, Posakony JW. 1991. The *Drosophila* Extramacrochaetae protein antagonizes sequence-specific DNA binding by Daughterless/Achaete-Scute protein complexes. *Development* 113: 245–255.

Wang Y, Chen KP, Yao Q, Wang W, Zhu Z. 2007. The basic helix-loop-helix transcription factor family in *Bombyx mori*. *Development Genes and Evolution* 217(10): 715–723.

Wang Y, Chen KP, Yao Q, Wang WB, Zhu Z. 2008. The basic helix-loop-helix transcription factor family in the honeybee, *Apis mellifera*. *Journal of Insect Science* 8: 40. Available online at http://www.insectscience.org/8.40/.

Wang Y, Chen KP, Yao Q, Zheng XD, Yang Zhe. 2009. Phylogenetic Analysis of Zebrafish Basic Helix-Loop-Helix Transcription Factors. *Journal of Molecular Evolution* 68(10): 629-640.

Wilson AC, Ashton PD, Calevro F, Charles H, Colella S, Febvay G, Jander G, Kushlan PF,

Macdonald SJ, Schwartz JF, Thomas GH, Douglas AE. 2010. Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. *Insect Molecular Biology* 2: 249-258.

Zheng X, Wang Y, Yao Q, Yang Z, Chen K. 2009. A genome-wide survey on basic helix-loop-helix transcription factors in rat and mouse. *Mammalian Genome* 20(10): 236-246.

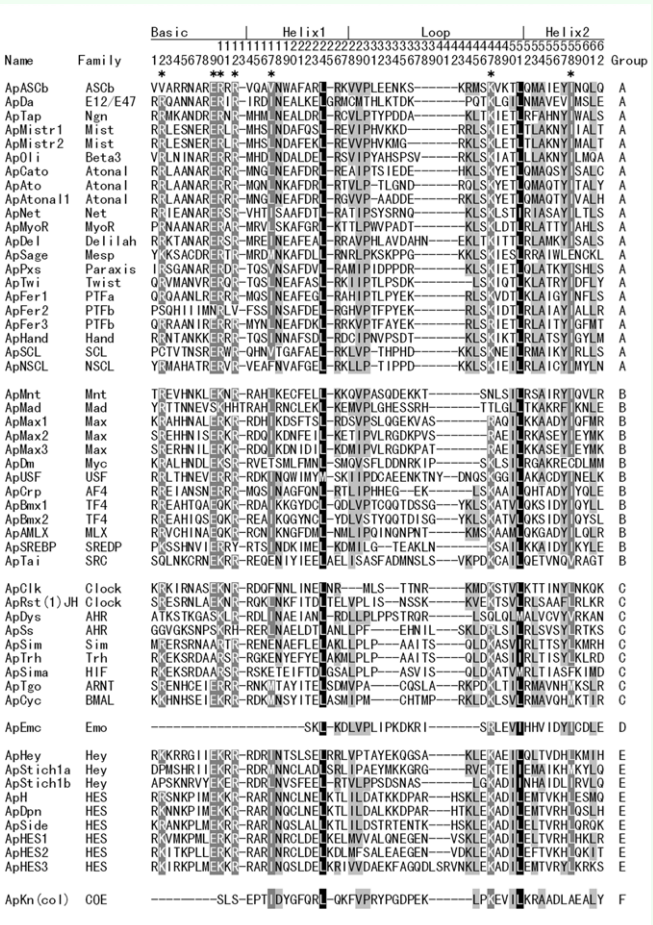

Figure 1. Alignment of 54 ApbHLH members. Designation of basic, helix 1, loop, and helix 2 follows Ferre-D'Amare et al. (1993). The family names and high-order groups have been organized according to Table 1 in Ledent et al. (2002). Highly conserved sites are indicated with asterisks on the top. High quality figures are available online.

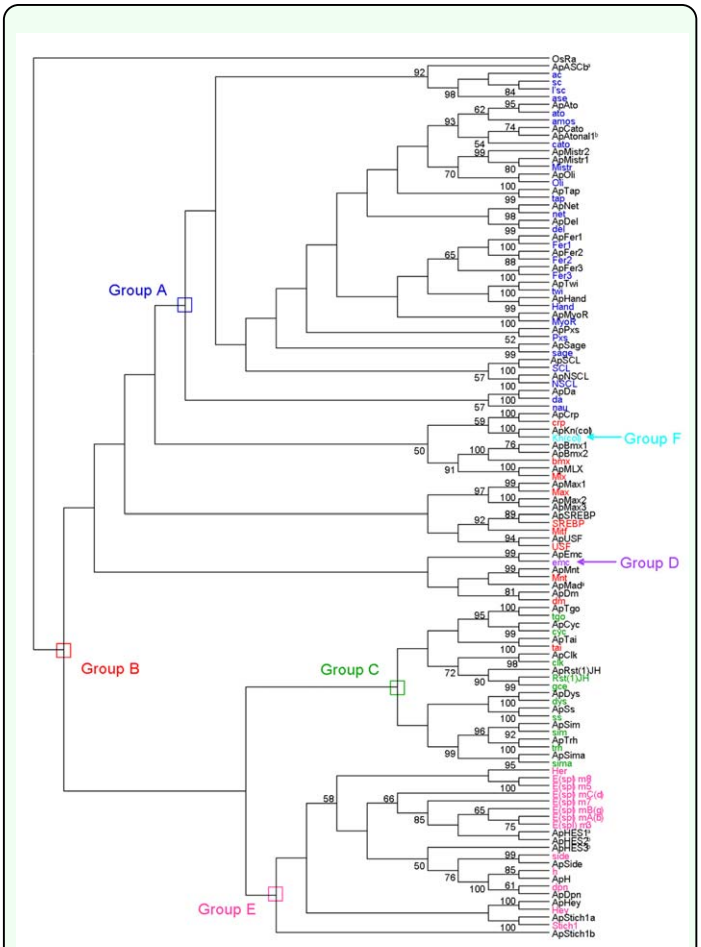

Figure 2. Phylogenetic relationship of 54 ApbHLH members with 59 *Drosophila melanogaster* bHLH members. A neighbor-joining (NJ) tree is shown. Bootstrap values less than 50 are not shown. The higher-order group labels are in accordance with Ledent et al. (2002). ApbHLH member marked with [a] or [b] meant that it did not form a monophyletic clade with any single *D. melanogaster* bHLH member and was subject to separate phylogenteic analyses with bHLH members from *Apis mellifera*, *Bombyx mori*, and *Tribolium castaneum*. High quality figures are available online.

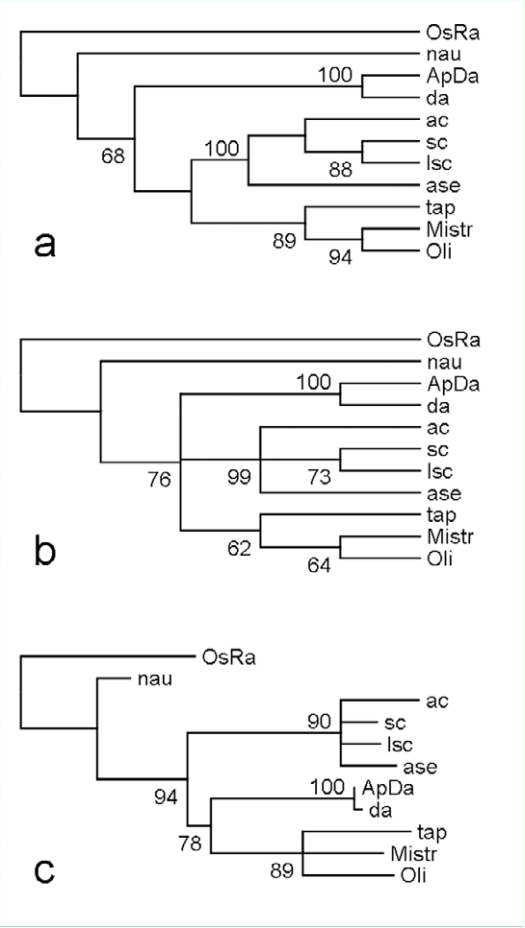

Figure 3. In-group phylogenetic analyses of ApDa. (a), (b), and (c) are NJ, MP, and ML trees, respectively, constructed with one *Acyrthosiphon pisum* bHLH member (ApDa) and seven group A bHLH members from *Drosophila melanogaster*. In all trees, OsRa (the rice bHLH motif sequence of R family) was used as the outgroup. High quality figures are available

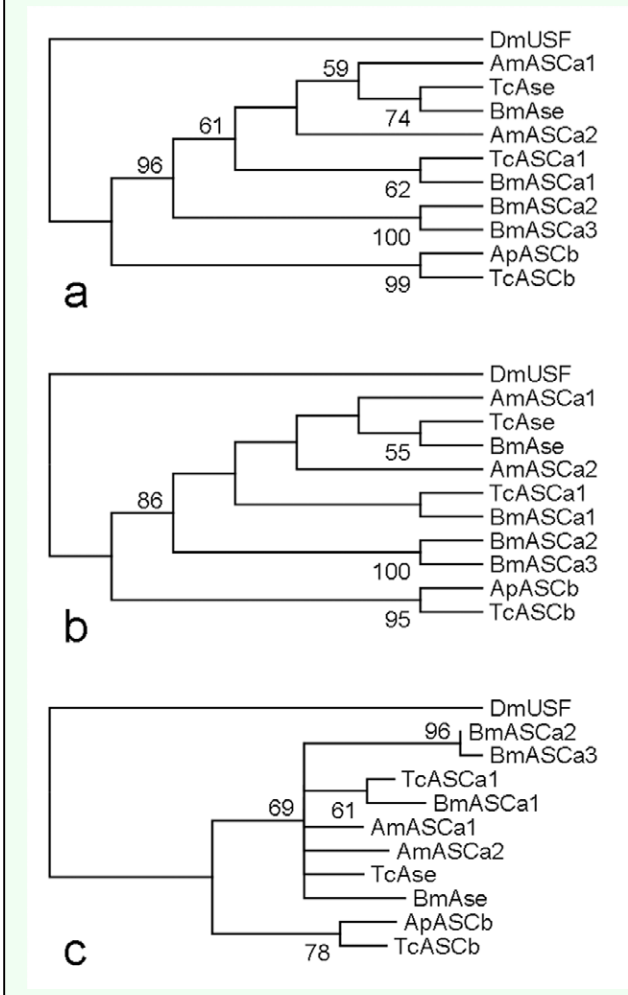

Figure 4. In-group phylogenetic analyses of ApASCb. (a), (b), and (c) are NJ, MP, and ML trees, respectively, constructed with one *Acyrthosiphon pisum* bHLH member (ApASCb) and nine ASC family members from *Apis mellifera*, *Bombyx mori*, and *Tribolium castaneum*. In all trees, bHLH motif sequence of DmUSF (*Drosophila melanogaster* upstream stimulation factor) was used as the outgroup. High quality figures are available online.

Table 1. A complete list of *Acyrthosiphon pisum* bHLH genes.

No.	Gene name	Family	Fruit fly homolog	Bootstrap values			Protein accession No.
				NJ	MP	ML	
1	*ApASCb* [a]	ASCb	*TcASCb*	99	95	78	XP_001949172.1
2	*ApDa*	E12/E47	*da*	100	100	100	XP_001950085.1
3	*ApTap*	Ngn	*tap*	99	93	58(*cato*)	hmm145914
4	*ApMistr1*	Mist	*Mistr*	100	98	87	XP_001944687.1
5	*ApMistr2*	Mist	*Mistr*	99	95	60	hmm401334
6	*ApOli*	Beta3	*Oli*	100	100	98	XP_001950802.1
7	*ApCato*	Atonal	*cato*	79	n/m*	n/m	hmm31924
8	*ApAto*	Atonal	*ato*	99	88	n/m*	hmm125654
9	*ApAtonal1* [b]	Atonal	*?*	n/m*	n/m*	n/m*	hmm61024
10	*ApNet*	Net	*net*	99	92	74	hmm79024
11	*ApMyoR*	MyoR	*MyoR*	100	99	85	XP_001948616.1
12	*ApDel*	Delilah	*dle*	99	92	78	XP_001945346.1
13	*ApSage*	Mesp	*sage*	100	99	n/m	XP_001948879.1
14	*ApPxs*	Paraxis	*Pxs*	61	52	n/m	hmm169244
15	*ApTwi*	Twist	*twi*	100	100	93	XP_001946602.1
16	*ApFer1*	PTFa	*Fer1*	100	94	89	hmm95774
17	*ApFer3*	PTFb	*Fer3*	100	100	96(*Pxs*)	hmm242594
18	*ApFer2*	PTFb	*Fer2*	94	77	72	Not available
19	*ApHand*	Hand	*Hand*	99	96	66	XP_001945320.1
20	*ApSCL*	SCL	*SCL*	100	99	75	NP_001156144.1
21	*ApNSCL*	NSCL	*NSCL*	100	100	69	XP_001951616.1
22	*ApMnt*	Mnt	*Mnt*	99	93	69	XP_001947496.1
23	*ApMad* [a]	Mad	*TcMad*	100	100	100	XP_001944077.1
24	*ApMax1*	Max	*Max*	100	96	92	XP_001942656.1
25	*ApMax2*	Max	*Max*	90	54	72	hmm160354
26	*ApMax3*	Max	*Max*	82	42	55	hmm30794
27	*ApDm*	Myc	*dm*	79	45	72(*SREBP*)	hmm384
28	*ApUSF*	USF	*USF*	98	84	68(*SREBP*)	XP_001947444.1
29	*ApCrp*	AP4	*crp*	100	97	97	XP_001945298.1
30	*ApBmx1*	TF4	*bmx*	100	94	n/m	XP_001947371.1
31	*ApBmx2*	TF4	*bmx*	98	87	89(*SREBP*)	XP_001951901.1
32	*ApMLX*	MLX	*MLX*	100	100	63	XP_001950231.1
33	*ApSREBP*	SREBP	*SREBP*	94	82	77	Not available
34	*ApTai*	SRC	*tai*	100	100	63	XP_001944363.1
35	*ApClk*	Clock	*clk*	100	93	74	XP_001944549.1
36	*ApRst(1)JH*	Clock	*Ret(1)JH*	n/m*	n/m*	59	hmm126914
37	*ApDys*	AHR	*dys*	100	100	93	Not available
38	*ApSs*	AHR	*ss*	100	100	87	XP_001946523.1
39	*ApSim*	Sim	*sim*	93	74	66	XP_001944204.1
40	*ApTrh*	Trh	*trh*	100	94	90	XP_001949586.1
41	*ApSima*	HIF	*sima*	96	94	56	XP_001951675.1
42	*ApTgo*	ARNT	*tgo*	100	100	91	XP_001945040.1
43	*ApCyc*	BMAL	*cyc*	96	n/m	n/m	NP_001164574.1
44	*ApEmc*	Emc	*emc*	99	98	95	XP_001947113.1
45	*ApHey*	Hey	*Hey*	100	98	n/m*	XP_001944649.1
46	*ApStich1a*	Hey	*stich1*	100	100	77	hmm38594
47	*ApStich1b*	Hey	*stich1*	55	52	n/m*	XP_001945126.1
48	*ApH*	H/E(spl)	*h*	96	95	n/m*	XP_001949685.1
49	*ApDpn*	H/E(spl)	*dpn*	45	n/m*	n/m*	XP_001947900.1
50	*ApSide*	H/E(spl)	*side*	99	97	95	XP_001945055.1
51	*ApHES1* [a]	H/E(spl)	*BmHES1*	n/m*	50	n/m*	XP_001946911.1
			TcHES1	44	50	58	
52	*ApHES2* [b]	H/E(spl)	*?*	n/m*	n/m*	n/m*	XP_001949270.1
53	*ApHES3* [b]	H/E(spl)	*?*	n/m*	n/m*	n/m*	XP_001943580.1
54	*ApKn (col)*	COE	*Kn(col)*	100	100	86	XP_001946640.1

ApbHLH genes were named according to their *D. melanogaster* homologues. Bootstrap values were from in-group phylogenetic analyses with *D. melanogaster* bHLH motif sequences using NJ, MP, and ML algorithms, respectively. OsRa (the rice bHLH motif sequence of R family) was used as the outgroup in every constructed tree except those for ApASCb, ApCato2, ApMad and ApHES1 which used separate outgroup sequence. n/m means that a ApbHLH does not form a monophyletic group with any other single bHLH motif sequence. n/m* means that a ApbHLH does not form a monophyletic clade with any specific bHLH motif sequence but forms a monophyletic clade with other bHLH proteins of the same family. a means that the gene's orthology was defined by in-group phylogenetic analyses with bHLH orthologs from *Bombyx mori*, *Tribolium castaneum* and/or *Apis mellifera*. b means that the gene was merely named numerically due to lack of orthologs in other insect species. The accession numbers are from different protein resources. Those labeled as "NP", "XP"and "hmm" are from 'RefSeq protein', 'Non-RefSeq protein' and 'Ab initio protein' databases, respectively.

Table 2. Table 2. Coding regions, intron location and length of 54 ApbHLH motifs.

Family	Gene name	Genomic coding sequence(s)			Intron (location, length)	Group
		Contig No.	Frame	Coding region(s)		
ASCb	*ApASCb*	NW_001917183.1	3	18063-18160	Loop: 4563bp	A
			3	22724-22787		
E12/E47	*ApDa*	NW_001932971.1	-3	5307-5278	Basic: 390bp	A
			-3	4842-4714		
Ngn	*ApTap*	NW_001924998.1	2	88937-89095		A
Mist	*ApMistr1*	NW_001938180.1	2	23615-23677	Helix 1: 7882bp	A
			2	31601-31696		
Mist	*ApMistr2*	NW_001918733.1	-2	30557-30495	Helix 1: 2304bp	A
			-2	28190-28095		
Beta3	*ApOli*	NW_001917515.1	-1	154686-154522		A
Atonal	*ApCato*	NW_001925016.1	-3	50328-50167-		A
Atonal	*ApAto*	NW_001922225.1	-1	195743-195585		A
Atonal	*ApAtonal1*	NW_001938652.1	1	55849-56007		A
Net	*ApNet*	NW_001938652.1	-1	187017-186859		A
MyoR	*ApMyoR*	NW_001936417.1	-3	27044_26886		A
Delilah	*ApDel*	NW_001921951.1	1	32101-32277		A
Mesp	*ApSage*	NW_001923944.1	1	16921-17052	Helix 2: 7107bp	A
			1	24160-24189		
Paraxis	*ApPxs*	NW_001917684.1	-2	26779-26736	Helix 1: 109bp	A
			-3	26626- 26512		
Twist	*ApTwi*	NW_001935314.1	1	46567-46722		A
PTFa	*ApFer1*	NW_001923357.1	2	22925-23083		A
PTFb	*ApFer2*	NW_001934059.1	2	40763-40921		A
PTFb	*ApFer3*	NW_001934211.1	1	51178-51336		A
Hand	*ApHand*	NW_001935894.1	-1	59779-59621		A
SCL	*ApSCL*	NW_001924455.1	-3	44157-44018	Helix 2: 8156bp	A
				35862-25844		
NSCL	*ApNSCL*	NW_001916472.1	-3	91988-91830		A
Mnt	*ApMnt*	NW_001919193.1	3	78591-78740	Helix 2: 5087bp	B
			2	83828-83836		
Mad	*ApMad*	NW_001931419.1	-1	220800-220763	Loop: 3918bp	B
			-1	216844-216730	Helix 2: 30718bp	
			-2	186011-186003		
Max	*ApMax1*	NW_001918063.1	-2	90852-90694		B
Max	*ApMax2*	NW_001931491.1	-3	3315-3157		B
Max	*ApMax3*	NW_001935958.1	1	29788-29946		B
Myc	*ApDm*	NW_001931984.1	-1	110536-110375		B
USF	*ApUSF*	NW_001917134.1	-2	24663-24541	Loop: 1629bp	B
			-2	22911-22861		
AP4	*ApCrp*	NW_001935115.1	1	4765-4875	Loop: 23289bp	B
			1	28165-28209		
TF4	*ApBmx1*	NW_001935304.1	-1	27426-27265	Loop: 370bp	B
			-2	26894-26886		
TF4	*ApBmx2*	NW_001920521.1	1	5059-5220,	Helix 2: 628bp	B
			2	5849-5857		
MLX	*ApMLX*	NW_001917260.1	-3	5328-5164		B
SREBP	*ApSREBP*	NW_001919193.1	-3	91249-91151	Loop: 71bp	B
			-2	91079-91026		
SRC	*ApTai*	NW_001935890.1	1	65269-65276	Basic: 7719bp	B
			1	72996-73158	Helix 2: 2082bp	
			1	75241-75243		
Clock	*ApClk*	NW_001927661.1	3	18999-19003	Basic: 74bp	C
			2	190078-190225		
Clock	*ApRst(1)JH*	NW_001937540.1	3	103248-103409		C
AHR	*ApDys*	NW_001938087.1	-3	33505-33344		C
AHR	*ApSs*	NW_001933871.1	3	88950-89111		C
Sim	*ApSim*	NW_001938176.1	2	52175-52336		C
Trh	*ApTrh*	NW_001932608.1	2	11282-11443		C
HIF	*ApSima*	NW_001935860.1	3	95013-95174	No	C
ARNT	*ApTgo*	NW_001927816.1	2	127112-127113	Basic: 11117bp	C
			1	138231-138390		
BMAL	*ApCyc*	NW_001922094.1	-2	169013 -169017	Basic: 518bp	C
			-1	168498-168342		
Emc	*ApEmc*	NW_001924511.1	-1	29584-29486		D
Hey	*ApHey*	NW_001922769.1	-3	108188-108021		E
Hey	*ApStich1a*	NW_001934199.1	2	183026-183193		E
Hey	*ApStich1b*	NW_001918065.1	2	100964- 100971	Basic: 62bp	E
			1	101034-101185		
H/E(spl)	*ApH*	NW_001932152.1	-1	6752-6747	Basic: 1857bp	E
			-1	4889-4776	Loop: 74bp	
			-3	4701-4648		
H/E(spl)	*ApDpn*	NW_001917026.1	1	7633-7638	Basic: 913bp	E
			2	8552-8647	Loop: 3832bp	
			3	12480-12551		
H/E(spl)	*ApSide*	NW_001936436.1	2	73498-73504	Basic: 13384bp	E
			3	86889-86984	Loop: 73bp	
			1	87058-87129		
H/E(spl)	*ApHES1*	NW_001920856.1	2	34958-34963	Basic: 604bp	E
			3	35565-35732		
H/E(spl)	*ApHES2*	NW_001918124.1	-1	40930-40925	Basic: 1069bp	E
			-2	39855-39766	Loop: 93bp	
			-2	39672-39595		
H/E(spl)	*ApHES3*	NW_001923890.1	1	11104-11223	Basic: 192bp	E
			1	11416-11508	Loop: 1567bp	
			2	13076-13159		
COE	*ApKn (col)*	NW_001916783.1	1	59578-50578	Basic: 194bp	F
			3	50773-50861	Loop: 914bp	
			2	51776-51820		