A Genome-Wide Identification of Basic Helix-Loop-Helix Motifs in Pediculus humanus corporis (Phthiraptera: Pediculidae)

Authors: Wang, Xu-Hua, Wang, Yong, Zhang, De-Bao, Liu, A-Ke, Yao, Qin, et al.

RESEARCH

A Genome-Wide Identification of Basic Helix-Loop-Helix Motifs in *Pediculus humanus corporis* (Phthiraptera: Pediculidae)

Xu-Hua Wang,[1] Yong Wang,[2,3] De-Bao Zhang,[1] A-Ke Liu,[1] Qin Yao,[1] and Ke-Ping Chen[1]

[1]Institute of Life Sciences, Jiangsu University, 301 Xuefu Rd., Zhenjiang 212013, China
[2]School of Food and Biological Engineering, Jiangsu University, 301 Xuefu Rd., Zhenjiang 212013, China
[3]Corresponding author, e-mail: ywang@ujs.edu.cn

ABSTRACT. Basic helix-loop-helix (bHLH) proteins comprise a large superfamily of transcription factors, which are involved in the regulation of various developmental processes. bHLH family members are widely distributed in various eukaryotes including yeast, fruit fly, zebrafish, mouse, and human. In this study, we identified 55 bHLH motifs encoded in genome sequence of the human body louse, *Pediculus humanus corporis* (Phthiraptera: Pediculidae). Phylogenetic analyses of the identified *P. humanus corporis* bHLH (PhcbHLH) motifs revealed that there are 23, 11, 9, 1, 10, and 1 member(s) in groups A, B, C, D, E, and F, respectively. Examination to GenBank annotations of the 55 PhcbHLH members indicated that 29 PhcbHLH proteins were annotated in consistence with our analytical result, 8 were annotated different with our analytical result, 12 were merely annotated as hypothetical protein, and the rest 6 were not deposited in GenBank. A comparison on insect bHLH gene composition revealed that human body louse possibly has more *hairy* and *E(spl)* genes than other insect species. Because *hairy* and *E(spl)* genes have been found to negatively regulate the differentiation of insect preneural cells, it is suggested that the existence of additional *hairy* and *E(spl)* genes in human body louse is probably the consequence of its long period adaptation to the relatively dark and stable environment. These data provide good references for further studies on regulatory functions of bHLH proteins in the growth and development of human body louse.

Key Words: basic helix-loop-helix, phylogenetic analysis, ortholog, *Pediculus humanus corporis*, BLAST search

The basic helix-loop-helix (bHLH) proteins form a superfamily of transcription factors involved in a wide range of eukaryotic developmental and biochemical processes including neurogenesis, myogenesis, sex determination, and environmental response (Massari and Murre 2000, Jones 2004, Castillon et al. 2007). These proteins are characterized by their bHLH motif, which is about 60 amino acids in length. The basic region is located at the N-terminal of bHLH motif. It is primarily responsible for binding to DNA with the assistance of certain basic residues such as R (arginine), K (lysine), and H (histidine). The HLH region is composed of two helices and a loop structure with variable length. It facilitates the formation of homodimeric or heterodimeric complexes between different family members through dimerization (Murre et al. 1989, Kadesh 1993, Massari and Murre 2000).

All eukaryotic bHLH transcription factors were first classified into 27 bHLH families and 4 higher order groups by means of phylogenetic analysis (Atchley and Fitch 1997). Two decades later, animal bHLH proteins were expanded to 45 bHLH families and 6 higher order groups. The 6 higher order groups (A, B, C, D, E, and F) were found to have 22, 12, 7, 1, 2, and 1 bHLH families, respectively, based on evolutionary relevance, structural, and functional properties (Simionato et al. 2007). Group A proteins mainly regulate neurogenesis, myogenesis, and mesoderm formation. They recognize and bind to E-box sequence typical of CAGCTG or CACCTG. Group B proteins mainly control cell proliferation and differentiation, sterol metabolism, adipocyte formation, and expression of glucose-responsive genes. They recognize and bind to E-box sequence typical of CACGTG or CATGTTG. Group C proteins usually contain a conserved Per-Arnt-Sim homolog (PAS) domain in addition to the bHLH motif. PAS domain promotes dimerization with another protein containing PAS domain. They are mainly involved in the regulation of midline development, tracheal development, and circadian rhythms, and in the activation of gene transcription in response to environmental toxins. Group C proteins recognize and bind to DNA core sequence as of ACGTG or GCGTG. Group D proteins serve as antagonist of group A protein for lack of the basic region. Group E proteins bind to CACGCG or CACGAG and usually contain two particular peptides named "Orange" and "WRPW" at the carboxyl terminus. Group F corresponds to the Col/Olf-1/EBF (COE) proteins, which lack a basic domain and are characterized by the presence of COE domain involved in both dimerization and DNA binding (Atchley and Fitch 1997, Crews 1998, Ledent and Vervoort 2001, Ledent 2002).

With the rapid expansion of nucleotide and protein databases available to public, it is becoming more and more convenient for any researchers to conduct surveys on bHLH proteins of any organisms whose genomes are sequenced and released online. It would not only benefit researchers who are dedicated to study structures and functions of individual bHLH proteins but also enable a quick growth of organism list with identified bHLH repertoire. Up to now, over 1,000 bHLH family members have been identified including 8 bHLH members in *Saccharomyces cerevisiae*, 16 in *Amphimedon queenslandica*, 33 in *Hydra magnipapillata*, 45 in *Caenorhabditis elegans*, 46 in *Ciona intestinalis*, 50 in *Strongylocentrotus purpuratus*, 50 in *Tribolium castaneum*, 51 in *Apis mellifera*, 52 in *Bombyx mori*, 54 in *Acyrthosiphon pisum*, 57 in *Daphia pulex*, 57 in *Harpegnathos saltator*, 59 in *Drosophila melanogaster*, 63 in *Lottia gigantea*, 64 in *Capitella* sp1, 68 in *Nematodtella vectensis*, 70 in *Acropora digitifera*, 78 in *Branchiostoma floridae*, 86 in *Taeniopygia guttata*, 87 in *Tetraodon nigroviridi*s, 104 in *Gallus gallus*, 107 in *Ailuropoda melanoleuca*, 114 in *Rattus norvegicus*, 114 in *Mus musculus*, 117 in *Homo sapiens*, 139 in *Brachydanio rerio*, 147 in *Arabidopsis*, and 167 in *Oryza sativa* (Robinson and Lopes 2000; Bailey et al. 2002; Ledent et al. 2002; Li et al. 2003; Satou et al. 2003; Simionato et al. 2007; Wang et al. 2007, 2008, 2009; Bitra et al. 2009; Zheng et al. 2009; Pires and Dolan 2010; Dang et al. 2011a,b; Gyoja et al. 2012; Liu et al. 2012).

The human body louse, *Pediculus humanus corporis* (Phthiraptera: Pediculidae), causes the cutaneous disease named pediculosis vestimenti by laying their eggs in the seams of clothing. It is the primary

vector of human diseases including relapsing fever, trench fever, and epidemic typhus. Human body louse diverged from human head louse (*Pediculus humanus capitis*) at ~100,000 years ago, dovetailing with the origin of clothing (Toups et al. 2011). The body louse has a long evolutionary association with human, which has been considered in medical and healthcare practice (James et al. 2011). During its long period adaptation to human parasitism, certain physiological and biochemical features could have been remarkably changed. However, previous studies have been mainly focused on the parasitic relationship between human and body louse, and the development processes of body lice to prevent and treat pediculosis (Levot 2000, Pedra et al. 2003, Toups et al. 2011). A comprehensive identification of bHLH proteins of the human body louse would facilitate studies on the emergence and underlying mechanism of specific physiological and biochemical features in human body louse.

Therefore, in this study, we conducted a genome-wide survey to genome sequence database of human body louse (Kirkness et al. 2010) and successfully identified 55 bHLH motifs encoded in the genome of human body louse. Further phylogenetic analyses enabled us to define orthology of the 55 identified *P. humanus corporis* bHLH (PhcbHLH) members by using known bHLH members from fruit fly and other insect species. It was found that 29 of the 55 PhcbHLH proteins have been annotated in consistency with our analytical result, 20 were either annotated different with our analytical result or were merely annotated as hypothetical proteins, and the rest 6 were not found in current GenBank databases. Besides, human body louse possibly has more *hairy* and *E(spl)* genes than other insect species, which is probably the result of its long period parasitism on human. Our present work establishes a good basis for further studies on regulatory functions of bHLH proteins in the growth and development of human body louse.

Materials and Methods

BLAST Searches and Manual Examination. First, with both 59 *D. melanogaster* bHLH (DmbHLH) and the 45 representative bHLH motifs obtained from the additional files of previous reports (Ledent and Vervoort 2001, Simionato et al. 2007) as query sequences, tBLASTn searches were performed against the RefSeq genomic and trace-whole-genome shotgun sequence databases of human body louse (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&BLAST_SPEC=OGP__121224__16222) to retrieve all potential bHLH sequences. All query sequences were not filtered to obtain coding regions covering the full bHLH range. Other parameters for the search were of default values. The retrieved sequences were manually checked to discard redundant ones having the same contig number, the same reading frame, and the same coding regions. In case where the retrieved amino acids did not cover the full bHLH range, we retrieved the corresponding nucleotide sequences from GenBank nucleotide database and translated them into amino acids by using EditSeq program of DNAStar package (version 5.01) to supplement the absent amino acids. Intron splice sites, which separated bHLH coding sequences into more than one region, were assessed by NetGene2 online (http://www.cbs.dtu.dk/services/NetGene2/).

Each of the above sequences was manually examined to see how many conserved amino acids existed in the 19 highly conserved sites (Atchley et al. 1999). If more than 10 conserved amino acids were present in the bHLH motif (Toledo-Ortiz et al. 2003), it was regarded as a candidate bHLH motif and was subject to further analyses. The bHLH motifs of Emc and COE families are relatively shorter, having 35 and 50 amino acids, respectively. Therefore, if more than five and eight conserved amino acids were present in potential Emc and COE sequences, the sequences were subject to further analyses as well.

To check whether there are protein sequences corresponding to the candidate motifs, BLASTp searches were performed against the RefSeq protein database of human body louse (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&BLAST_SPEC=OGP__121224__16222) using the above obtained candidate bHLH motif sequences.

Multiple Sequence Alignment. All the above bHLH motif sequences were aligned by ClustalW program implemented in MEGA 5 (Tamura et al. 2011) with default settings. We then obtained a rich text file using GeneDoc Multiple Sequence Alignment Editor and Shading Utility (version 2.6.02) (Nicholas et al. 1997), in which the conserved sites of aligned PhcbHLH motifs were shaded with different gray depths.

Phylogenetic Analysis. Evolutionary relationships among all identified PhcbHLH motifs were analyzed using three different algorithms: distance neighbor-joining (NJ), maximum parsimony (MP), and maximum likelihood (ML). NJ phylogenetic analyses (Saitou and Nei 1987) were performed online (http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=bionj) using BioNJ algorithm (Gascuel 1997). MP phylogenetic analyses were conducted using PAUP 4.0 Beta 10 (Swofford 1998) based on the step matrix constructed from Dayhoff PAM 250 distance matrix by R. K. Kuzoff (http://paup.csit.fsu.edu/data/pam250.nex). NJ distance tree was bootstrapped with 1,000 replicates to provide information about the statistical reliability. MP tree was generated using heuristic searches and bootstrapped with 500 replicates. ML trees were constructed using PhyML program online (http://www.atgc-montpellier.fr/phyml/) (Guindon et al. 2010) with the following parameter settings: BioNJ starting tree, 500 bootstrap steps, and LG (Le and Gascuel et al. 2008) substitution model. Other parameters such as proportion of invariable sites or gamma-shape parameter were optimized by ProtTest (Abascal et al. 2005).

Phylogenetic analyses of PhcbHLH motifs were carried out in two steps. First, all the candidate PhcbHLH motifs were used to build ML trees with 59 DmbHLH, 114 *M. musculus* bHLH (MmbHLH), and 70 *Acr. digitifera* bHLH (AdibHLH) motif sequences (from Supp Figs. S1–S6 [online only]). These trees clearly displayed to which higher order group a candidate PhcbHLH sequence belonged. Then, each candidate PhcbHLH motif was used to conduct in-group phylogenetic analysis with DmbHLH motif sequences. That is, a single PhcbHLH sequence was used to construct NJ, MP, and ML phylogenetic trees with known DmbHLH members of the same group (Wang et al. 2007, 2008). When in-group phylogenetic analysis using DmbHLH members could not yield evolutionary trees with sufficient bootstrap support, bHLH sequences from *Anopheles gambiae*, *A. mellifera*, *Acy. pisum*, or *T. castaneum* were then used to do the in-group analysis till sufficient bootstrap support was obtained for orthology assignment. Criterion for orthology assignment was as follows: if a PhcbHLH sequence formed a monophyletic clade with one DmbHLH or other insect bHLH sequence with bootstrap support >50% in various phylogenetic trees, the known DmbHLH or other insect bHLH member was regarded as an ortholog of the PhcbHLH sequence.

Protein Functional Domain Prediction. To further assess the reliability of our classification to the identified motifs and to examine whether the full-length protein sequences contain additional characteristic domains such as PAS, we carried out prediction of protein domain architectures with simple modular architecture research tool (SMART, http://smart.embl.de/) online.

Results and Discussion

Identification of PhcbHLH Members. Through BLAST searches, manual examinations and phylogenetic analyses, we identified 55 bHLH motifs encoded in the genome of *P. humanus corporis*. The alignment of all 55 PhcbHLH motifs is shown in Fig. 1. In our study, we named PhcbHLH genes according to the family names they belong to, which will facilitate further studies on structural and functional comparison with other organisms. Meanwhile, we added "1", "2," and "3," etc. to names of some PhcbHLHs, which belong to one single bHLH

```
                    Basic   |  Helix 1  |     Loop      |   Helix 2    |
                    1111111111222222222233333333334444444444555555555566666666667 7
Name      Family    12345678901234567890123456789012345678901234567890123456789 01
                    **  *  *     *     *                        ***  *     *
PhcASCa1    ASCa     SVARRNARERNVKQVNNGFATLR-SHIP--ASVAAALTSEKSSPRNSSASKLS VETLKMAVEYIRSLQ   A
PhcASCa2    ASCa     SWNSRCER-RNVKQVNEGFATLR-KRIP--EYISEAYNEQKG--RKS--AKLS VETLRMAVDYIRHLE   A
PhcASCb     ASCb     SIARRNARERNVKQVNEGFATLR-NRIP--EYISEAYNEEKG--RKS--AKLS VETLRMAVDYIRHLE   A
PhcE12/E47  E12/E47  RRQANNARERIRDINEALKELG-RMCM--THLKT------------DKPQTLGLLNMAVEVIMQLE       A
PhcMyoD     MyoD     RRKAATLRERRRLKVNEAFEVLK-RRTC--SN-PN---------QRLPVEILRNAIEYIESLN         A
PhcNgn      Ngn      RRMKANDRERNRMHLINEALDRLR-CVIP--TYPTD-----------TKILTIETLRFAHNYIWALS     A
PhcMist1    Mist     RRLESNERERMMHSINDAFQSLR-EVIP--HVKKE----------RKLSIETLTLAKNYILALT         A
PhcMist2    Mist     RRLESNERERMMHSINDAFEQLR-EVIP--HVKME----------RKLSIETLTLAKNYIMALT         A
PhcBeta3    Beta3    VRLNINARERRMHDINDALDELR-SVIP--YAHSPS--------VRKLSIATLLLAKNYILMQA         A
PhcAtonal1  Atonal   RRLAANARERRRMNGINEAFDRLR-EVIP--SLGAD---------HKLSIFETLQMAQSYIHALC        A
PhcAtonal2  Atonal   RRLAANARERMENLNKAFDRLR-THIP--SLGSD---------RQLSIYETLQMAQTYISALC          A
PhcAtonal3  Atonal   RRLAANARERKMNSLNDAFDRLR-EWIP--SLGND---------RKLSIYETLQMAQTYISALY         A
PhcNet      Net      RRIEANARERTLVHTISAAFDTLR-HSIP--AYSH----------NQKLSLSVLRIACSYIMTLS        A
PhcMyoRa    MyoRa    QRNAANARERARMRILSKAFSRLK-TTIP--WVPAD---------TKILSLDTLRLATSYIAHLR        A
PhcDeliah   Delilah  RRKTANARERNRMREINVAFETLR-RVIP--DISNSVST-----GNEKLSITTLRLAMKYIQTLR        A
PhcMesp     Mesp     YKKSACDRERTMRDLNRAFDLRAKLPV-TKSPG---------KKLSIEAIRLSIRYIRHLQ           A
PhcTwist    Twist    QRVMANVRERQRTQSLNEAFTSLR-KIIP--TLPSD---------K-SLSIIQTLKLASRYIDFLY       A
PhcPTFa     PTFa     QRQAANLRERKMQSLNDAFEGLR-AHIP--TLPYE---------KRLSIVDTLKLAIGYINFLS         A
PhcPTFb1    PTFb     QRHAANIRERKMLSINSAFDELR-IHIP--TFPYE---------KRLSIIDTLRLAIAYIALLR         A
PhcPTFb2    PTFb     QRRAANIRERRMFNLNEAFDKLR-RKIP--TFAYE---------KRLSIIETLRLAITYISFMA         A
PhcHand     Hand     RRNTANKKERRRTQSLNNAFADLR-DCIP--NVPAD---------TKILSIKTLRLATSYIGYLM        A
PhcSCL      SCL      KKIFTNSRERWRQQNVSGAFAELR-KLYP--THPPD---------KKLSINEIRMAIKYINLLT         A
PhcNSCL     NSCL     YRTAHATRERIRVEAFNVAFSDLR-KLIP--TLPPD---------KKLSIEIRLAICYIAYLN          A

PhcMnt      Mnt      TREVHNKLEKNRRAHLKECFELVK-KQLP--ASQE------------DKKASNLNILHSALRFIQTLR     B
PhcMax      Max      KRAHHNALERKRDHIKDSFSSLR-DSVP--ALHG---------EKVASLAQLLKKAAEYIQYMR        B
PhcMyc      Myc      KRTIHNSMERQRRIDINRNAFEELR-KLYP--FTAGN---------EKAPKVVILSYGADFCNTLT       B
PhcUSF      USF      RRATHNEVERRRDKLNNWIAKLS-KIIP--DCSSMDN------NKSVQSGGILAKACEYIMELR         B
PhcMITF1    MITF     KKDNHNMIERRRRFNINDRIKELG-TLLP--KNNDP---------DIRPNGTILKSSVDYIKILK        B
PhcMITF2    MITF     KKDNHNMIERRRRYNINDRIKELA-DLLP--KTNDPHYDLV---RDNRQNGAILKSSVDYIKLLK        B
PhcAP4      AP4      RREIANSNERRRMQSLNAGFQSLR-TLLP--SHEG----------ELLSVKSLRLRRKNECTYV        B
PhcTF4      TF4      RREAHTQAEQRRRDAIKKGYDTLQ-ELYPTCQQTDNSG------YKLSIAIVLQRSIDYIQFLH        B
PhcMLX      MLX      RRVCHINAEQKRYNIKNGFDVLH-SLIP--YLTLNPN-------VKLSIATMLQKGAEYIRQLQ        B
PhcSREBP    SREBP    IFLAHNAIERRYRTSINEKIIELK-NIIV-G------------KEAKINSLIIKKAIDYIKFLK        B
PhcSRC      SRC      CFSNKCLNERREQENIYIEELA-ELISASFADMSS--------LSVIPDCAILQETVNQIRHIK        B

PhcClock1   Clock    KRRSRNLSEKKRRDQFNMLINELS-SMISTNNR----------KIDISTVLKSTIAFLESHN          C
PhcClock2   Clock    SRMARNLAEKMRDKLNTHINELS-TLVPMTAGSSK--------KIDITSILRLSATYIRMNK           C
PhcAHR1     AHR      SSKSTKGASKLRDLINAEISNLR-DLLP--LPPSTR--------QRLSQLQLMALVCVYIRKAN        C
PhcAHR2     AHR      DGVTKSNPSRHRERLNAELDTLA-SLLPFEQNILS--------KLDLSLIRLSVSYIRTKS           C
PhcSim      Sim      MKEKSKNAARNREKENAEFLELA-KLLPLPSAITS--------QLDIASIDRLTTSYIKMRH          C
PhcTrh      Trh      RKEKSRDAARSRGKENFEFYELA-KMLPLPAAITS--------QLDIASIDRLTISYIKLRD          C
PhcHIF      HIF      RKEKSRDAARCRRSRETEIFTDLA-NGLPLKQSVIA--------QLDKASVRLVISYIKVRH          C
PhcARNT     ARNT     ARENHCEIERRRRNKITAYITELS-DMYPTCSALAR-------KPDKLTIRMAVAHIKALR           C
PhcBMAL     BMAL     FRQNHSEIEKRRRDKMNTYITELA-NMVPICHANPR-------KLDKLTVLRMAVQHIRLLL          C

PhcEmc      Emc      ----------------SKILK-DLVP--FMPKNK---------KLTKLEVIQYVIDYICDLQ          D

PhcHey1     Hey      RKKRRGIIEKRRRDRINSSLSELR-RLYPTAFEKQG-------SAKIEIAEILQMTVDHIKNLH        E
PhcHey2     Hey      DPMSHRIIEKRRRDRINNCLADLS-RLIPPEYLKKG-------RGTIEITEIIEMAIKHIGHLQ        E
PhcHES1     H/E(spl) RRSNKPIMEKRRRARINNSLNELR-TLILDAMKKDP-------ARHSKLEKADILEMTVKYIEALQ      E
PhcHES2     H/E(spl) RKSNKPIMEKRRRARINHCLNELR-TLILDAMKKDP-------ARHSKLEKADILEMTVKHIQQIQ      E
PhcHES3     H/E(spl) FQSNKPIMEKRRRARINNSLNELR-TLILDAMKKDP-------ARHSKLEKADILEMTVKYIEALQ      E
PhcHES4     H/E(spl) FQSNKPIMEKRRRARINNSLNELR-TPILDAMKKDP-------ARHSKLEKADILEMTGKYIEALQ      E
PhcHES5     H/E(spl) RRINKPLMEKRRRARINQSLAILK-ALILDSTKTEQNN-----TKHSKLEKADILELTVRHIQRQK      E
PhcHES6     H/E(spl) RKVMKPMLEKRRRARINRCLDELK-ELIVAALQAEG-------ENVSKLEKADILELTVSHIHKLR      E
PhcHES7     H/E(spl) KKVMKPMLEKRRRARINRCLDELK-ELIVAALQAEG-------ENVSKLEKADILELTVDHIGKLR      E
PhcHES8     H/E(spl) LKVMKPMLERKRRARINKCLDELK-ELIVAALQAEG-------ENVSKLEKADILELTVSHIHKLR      E

PhcCOE      COE      ------ALS-EPTIDYG--FQRIQ-KLIP--RHPGDP-------EKLPKEIILKRAADLAEALY        F
```

Fig. 1. Multiple alignment of the 55 PhcbHLH motifs. Highly conserved sites are marked with asterisks and hyphens denote gaps. The family names and high-order groups have been organized according to Table 1.

family. For example, there are two human body louse bHLH genes in family Mist, which were named *PhcMist1* and *PhcMist2*, respectively. Detailed information of the 55 PhcbHLH genes, including names, bootstrap values from phylogenetic analyses, and GenBank annotations are listed in Table 1. Figure 1 and Table 1 led us to conclude that there were 23, 11, 9, 1, 10, and 1 PhcbHLH members in groups A, B, C, D, E, and F, respectively.

Determination of PhcbHLH Orthology. Orthology determination tells whether similar genes in different organisms are orthologous. Although orthology determination has confronted with certain difficulty because no absolute criterion can be used to determine whether two genes are orthologous (Ledent and Vervoort 2001), the in-group phylogenetic analysis has proved to be reliable for identifying orthologous sequences in our previous studies (Wang et al. 2007, 2008).

Therefore, in this study, we also used in-group phylogenetic analysis to define orthology for the identified PhcbHLH motifs.

Based on the overall ML trees constructed using amino acids of 55 PhcbHLH motifs and bHLH motifs from *D. melanogaster*, *M. musculus*, and *Acr. digitifera* (Supp Figs. S1–S6 [online only]), in-group phylogenetic analysis was conducted to determine orthology of each PhcbHLH member. For example, Supp Fig. S3 [online only] showed that PhcSim formed a large evolutionary clade with other group C bHLH members. Therefore, it was used to construct NJ, MP, and ML phylogenetic trees with 10 group C bHLH members from *D. melanogaster* (Fig. 2). As a result, PhcSim formed monophyletic clade with sim (single minded) of *D. melanogaster* with high bootstrap values. Therefore, we considered PhcSim as an ortholog of fruit fly sim. Similarly, in-group phylogenetic analysis was conducted to each of the

Table 1. A complete list of PhcbHLH genes

Family	Fruit fly gene name	PhcbHLH name	Bootstrap values			Protein accession no.	Annotation in GenBank
			NJ	MP	ML		
ASCa	*Ase*	*PhcASCa1*[Ag]	75	70	86	XP_002430851.1	Hypothetical protein
		PhcASCa2[Am]	98	100	96	Not available	Not available
ASCb	*ASCb*	*PhcASCb*[Ap]	68	56	76	XP_002430850.1	**ASCb-like**
E12/E47	*da*	*PhcE12/E47*	95	100	100	XP_002428289.1	**E12/E47-like**
MyoD	*nau*	*PhcMyoD*	67	100	99	XP_002429203.1	**MyoD-like**
Ngn	*tap (bp)*	*PhcNgn*	74	82	97	XP_002432820.1	Hypothetical protein
Mist	*Mistr*	*PhcMist1*	99	100	99	XP_002432057.1	*Class B bHLH protein*
		PhcMist2	n/m	100	99	XP_002432605.1	**Mist1-like**
Beta3	*Olig*	*PhcBeta3*	99	100	100	XP_002431018.1	**Oli-like**
Atonal	*cato*	*PhcAtonal1*[Ap]	n/m	56	72	XP_002425799.1	Hypothetical protein
	ato	*PhcAtonal2*	78	81	99	XP_002427609.1	Hypothetical protein
	amos	*PhcAtonal3*[Tc]	44	53	60	XP_002425798.1	Hypothetical protein
Net	*net*	*PhcNet*	94	99	100	XP_002424436.1	**Net-like**
MyoRa	*MyoR*	*PhcMyoRa*	95	99	100	XP_002427093.1	*NeuroD-like*
Delilah	*del*	*PhcDelilah*	n/m	98	84	Not available	Not available
Mesp	*sage*	*PhcMesp*	84	99	100	XP_002429125.1	Hypothetical protein
Twist	*twi*	*PhcTwist*	99	99	100	XP_002428670.1	**Twist-like**
PTFa	*Fer1*	*PhcPTFa*	61	79	94	XP_002430763.1	**PTFa-like**
PTFb	*Fer2*	*PhcPTFb1*	83	88	91	XP_002430775.1	Hypothetical protein
	Fer3	*PhcPTFb2*	96	100	100	XP_002424449.1	*Atonal-like*
Hand	*Hand*	*PhcHand*	94	98	100	XP_002423846.1	**Hand-like**
SCL	*SCL*	*PhcSCL*	96	99	100	XP_002423128.1	*NSCL-like*
NSCL	*NSCL*	*PhcNSCL*	89	99	100	XP_002425962.1	Hypothetical protein
Mnt	*Mnt*	*PhcMnt*	91	93	100	XP_002432975.1	**Mnt-like**
Max	*max*	*PhcMax*	79	96	99	XP_002425776.1	**Max-like**
Myc	*dm*	*PhcMyc*	87	89	98	XP_002432048.1	**Myc-like**
USF	*USF*	*PhcUSF*	71	79	97	XP_002426967.1	**USF-like**
MITF	*Mitf*	*PhcMITF1*	82	100	100	XP_002433027.1	**MITF-like**
	Mitf	*PhcMITF2*	88	99	100	XP_002433291.1	**MITF-like**
AP4	*crp*	*PhcAP4*	78	95	100	XP_002432850.1	*Atonal-like*
TF4	*bmx*	*PhcTF4*	76	77	100	XP_002426560.1	**TF4-like**
MLX	*MLX*	*PhcMLX*	96	98	100	XP_002428196.1	**MLX-like**
SREBP	*SREBP*	*PhcSREBP*	78	64	90	XP_002424615.1	**SREBP-like**
SRC	*tai*	*PhcSRC*	99	99	100	XP_002430185.1	*Clock-like*
Clock	*clk*	*PhcClock1*	89	100	100	XP_002429283.1	**Clock-like**
	gce	*PhcClock2*[Tc]	66	97	98	XP_002430841.1	*HIF-like*
AHR	*dys*	*PhcAHR1*	99	100	100	XP_002424108.1	Hypothetical protein
	ss	*PhcAHR2*	96	100	100	XP_002427145.1	**AHR-like**
Sim	*sim*	*PhcSim*	70	92	72	XP_002423706.1	**Sim-like**
Trh	*trh*	*PhcTrh*	51	95	100	XP_002433081.1	**Trh-like**
HIF	*sima*	*PhcHIF*	78	85	94	XP_002431762.1	**HIF-like**
ARNT	*tgo*	*PhcARNT*	95	100	100	XP_002430960.1	Hypothetical protein
BMAL	*cyc*	*PhcBMAL*	53	84	76	XP_002432327.1	Hypothetical protein
Emc	*emc*	*PhcEmc*	96	100	100	XP_002427092.1	**Emc-like**
Hey	*Hey*	*PhcHey1*	83	95	95	XP_002429873.1	**Hey-like**
	Stich1	*PhcHey2*	97	100	100	XP_002424531.1	*Class B bHLH protein*
H/E(spl)	*h*	*PhcHES1*	58	84	93	XP_002425040.1	**H/E(spl)-like**
		PhcHES2[Ag]	66	84	80	XP_002423249.1	**H/E(spl)-like**
		PhcHES3	50	81	94	Not available	Not available
		PhcHES4	54	79	92	Not available	Not available
	side	*PhcHES5*	98	93	100	XP_002432526.1	Hypothetical protein
	E(spl)mB(g)	*PhcHES6*[Ap]	90	94	100	XP_002428314.1	**H/E(spl)-like**
		PhcHES7[Ap]	90	88	98	Not available	Not available
		PhcHES8[Ap]	90	96	98	Not available	Not available
COE	*kn (col)*	*PhcCOE*	99	100	100	XP_002424210.1	**COE-like**

PhcbHLH genes were named according to the correspondent families they belong to. Bootstrap values were obtained from in-group phylogenetic analyses with OsRa as outgroup in each constructed tree. n/m, a PhcbHLH does not form a monophyletic group with any other single bHLH motif sequence. Superscript letters Ag, Am, Ap, and Tc mean that orthology of the gene was defined through in-group phylogenetic analyses with bHLH orthologs from *An. gambiae*, *A. mellifera*, *Acy. pisum*, and *T. castaneum*, respectively. In last column, words in bold face indicate identical or alternative names annotated in GenBank, which are consistent with our phylogenetic analytical result. Those in italic type indicate GenBank annotations different with our analytical result, and those in normal type indicate hypothetical protein or absence of annotation information in GenBank.

identified PhcbHLH members. All the bootstrap values of constructed NJ, MP, and ML trees for each identified PhcbHLH member were listed in Table 1 without displaying correspondent constructed trees. It was found that the orthology of PhcbHLH members with *D. melanogaster* and other insect species could be divided into the following categories.

First, 43 PhcbHLH motifs formed monophyletic clades with DmbHLH sequences with all the bootstrap values over 50 in constructed NJ, MP, and ML trees. They are PhcE12/E47, PhcMyoD, PhcNgn, PhcMist1, PhcBeta3, PhcAtonal2, PhcNet, PhcMyoR, PhcMesp, PhcTwist, PhcPTFa, PhcPTFb1, PhcPTFb2, PhcHand, PhcSCL, PhcNSCL, PhcMnt, PhcMax, PhcMyc, PhcUSF, PhcMITF1, PhcMITF2, PhcAP4, PhcTF4, PhcMLX, PhcSREBP, PhcSRC, PhcClock1, PhcAHR1, PhcAHR2, PhcSim, PhcTrh, PhcHIF, PhcARNT, PhcBmal, PhcEmc, PhcHey1, PhcHey2, PhcHES1, PhcHES3, PhcHES4, PhcHES5, and PhcCOE. Because of the high bootstrap values above the set criterion (50), we, therefore, confidently
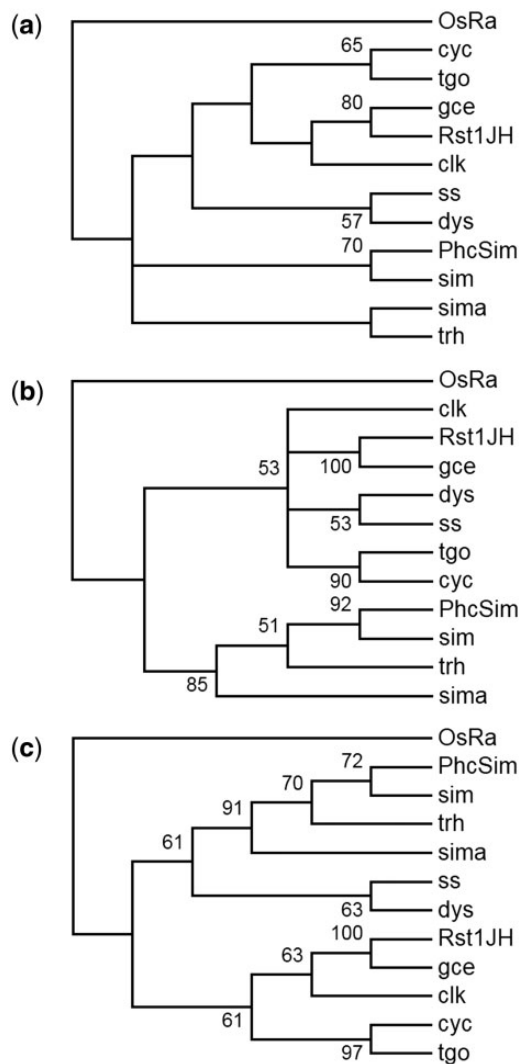
Fig. 2. In-group phylogenetic analyses of PhcSim. (a–c) are NJ, MP, and ML trees, respectively, which were constructed with one PhcSim and 10 group C bHLH members from *D. melanogaster*. In all trees, OsRa (a rice bHLH motif sequence of R family) was used as outgroup. Only bootstrap values no less than 50 are shown.

defined orthology of these PhcbHLH motifs as correspondent to DmbHLH orthologs.

Second, two PhcbHLH motifs, namely PhcMist2, and PhcDelilah, did not form monophyletic clade with bHLH sequences of *D. melanogaster* in NJ phylogenetic tree (marked with n/m in Table 1). PhcMist2 motif formed monophyletic clade in MP and ML trees with bootstrap values of 100 and 99, respectively. PhcDelilah formed monophyletic clade in MP and ML trees with bootstrap values of 98 and 84, respectively. Although we did not have sufficient bootstrap supports from all three constructed phylogenetic trees, we defined orthology for them based on the two formed monophyletic clades with bootstrap values over 50. These assignments may be modified if new data demonstrate discrepancy with our current analysis.

Finally, the rest 10 PhcbHLH motifs, namely PhcASCa1, PhcASCa2, PhcAtonal1, PhcAtonal3, PhcASCb, PhcClock2, PhcHES2, PhcHES6, PhcHES7, and PhcHES8, did not form any monophyletic clade with corresponding DmbHLH sequence in all three constructed phylogenetic trees. Therefore, we defined their orthology through constructing phylogenetic trees with corresponding bHLH members from *An. gambiae, A. mellifera, Acy. pisum,* or *T. castaneum,* respectively (marked with superscript letters Ag, Am, Ap, and Tc,

respectively in Table 1). Among them, PhcASCa1, PhcASCa2, PhcASCb, PhcClock2, PhcHES2, PhcHES6, PhcHES7, and PhcHES8 were defined with sufficient confidence because all bootstrap values were over 50 in the three constructed trees, while the rest two members, PhcAtonal1 and PhcAtonal3, had bootstrap values of over 50 only in two of the three constructed trees.

It is to be noted that three additional bHLH families, i.e., pearl and amber, which belong to group A, and peridot, which belongs to group D, have been found in *Acr. digitifera* (Gyoja et al. 2012). We have included these three sequences in both our general phylogenetic analyses (Supp Figs. S1 and S4 [online only]) and in-group phylogenetic analysis (Table 1). However, in all the constructed phylogenetic trees, no PhcbHLH sequence formed monophyletic clade with any of the three AdibHLH sequences, providing another instance of probable loss of these three genes during insect evolution.

Identification of PhcbHLH Protein Sequences. Protein sequence accession numbers of the identified PhcbHLHs are listed in Table 1. It was found that 49 PhcbHLH motifs have correspondent protein sequences deposited in GenBank (show as "XP_" plus numbers) and the rest 6 PhcbHLHs, namely PhcASCa2, PhcDelilah, PhcHES3, PhcHES4, PhcHES7, and PhcHES8, do not have correspondent protein sequences in current database. Further examination to the 49 PhcbHLH protein sequences revealed that all of them are from the annotation to genome sequences after completion of the human body louse genome sequencing project (Kirkness et al. 2010). Among them, 29 PhcbHLH proteins were annotated in consistence with our analytical result (Table 1, shown in bold face at the last column), 8 PhcbHLH proteins were annotated different with our analytical result (Table 1, shown in italics at the last column), and the rest 12 were merely annotated as hypothetical proteins (Table 1, shown in normal type at the last column). Therefore, our data provide good reference for updating annotations to the 26 PhcbHLH proteins in current GenBank database. For example, our analysis highly supports that PhcPTFb2 is a bHLH member of PTFb family but not that of Atonal family (Table 1).

Although amino acid sequences flanking the bHLH motif are generally divergent even in closely related proteins from the same species, certain conserved domains or motifs are often present within related bHLH protein groups (Jones 2004). To further determine reliability of our classification to the identified PhcbHLHs, a separate phylogenetic tree (Fig. 3) with predicted protein domain was constructed based on an alignment of all PhcbHLH motifs. As we can see, HLH domain was identified in all PhcbHLH protein sequences. In addition, group C PhcbHLHs are characterized by having two PAS and one C-terminal to PAS motif (PAC) domains with exception only on PhcAHR1. Four of the six group E PhcbHLH full-length protein sequences have an Orange domain. Apart from the common domains existing in groups C and E bHLH proteins, other structural domains were also found in individual PhcbHLH members. For example, PhcE12/E47 (group A) and PhcClock1 (group C) have a coiled coil domain, PhcUSF (group B), PhcBMAL, and PhcARNT (group C) have a transmembrane domain, PhcSRC (group B) has two PAS domains, and PhcCOE (group F) has an Immunoglobulin Plexin Transcription (IPT) domain. To sum up, our analyses indicated that protein architecture is highly conserved within specific bHLH groups, and the above data provide further support to the results of our phylogenetic analysis based on bHLH motifs (Fig. 1 and Table 1).

Genomic Coding Regions of PhcbHLH Motifs. Coding regions of the 55 PhcbHLH motifs are listed in Table 2. It was found that the coding regions of 21 PhcbHLH motifs contain one intron in basic, helix 1, loop, or helix 2 region, and those of 4 PhcbHLH motifs (i.e., PhcMITF1, PhcSREBP, PhcHES1, and PhcHES2) have 2 introns, which are located in the basic and loop region, respectively. So, totally there are 29 introns in the coding regions of all 55 PhcbHLH motifs. The longest intron in coding regions of PhcbHLH motifs is 6,723 bp (base pairs), the shortest one is only 66 bp, and the average length of the 29 introns is 616 bp. While in *Acy. pisum, H. saltator, D. melanogaster,*
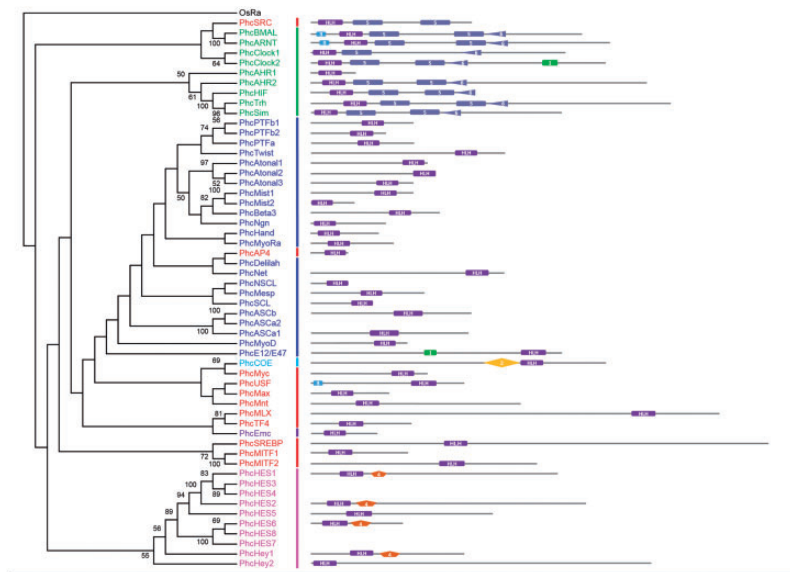
Fig. 3. Phylogenetic relationship of PhcbHLH members and architecture of PhcbHLH protein conserved domains. The left panel is an ML tree of 55 bHLHs in the human body louse with OsRa as outgroup. For simplicity, branch lengths of the tree are not proportional to distances between sequences and only bootstrap values no less than 50 are shown. PhcbHLH names of groups A–F are shown as blue, red, green, purple, magenta, and aqua characters, respectively. The right panel is the architecture of HLH and additional domains detected by SMART, shown by numbered and colored blocks, among which 1–6 stand for domains of coiled coil region, IPT, transmembrane region, Orange, PAS, and PAC, respectively. The six PhcbHLH members without full-length protein sequences (as indicated in Table 1) were excluded from SMART analysis.

and *A. mellifera*, there are 26, 22, 18, and 9 bHLH members having introns in coding regions of their bHLH motifs. The total number of their introns is 34, 26, 20, and 9, the longest intron is of 30,718, 7,943, 11,845, and 4,460 bp, the shortest one is of 62, 82, 57, and 72 bp, and the average intron length is of 4,193, 1,391, 1,082, and 1,326 bp, respectively (Liu et al. 2012). In summary, the number of PhcbHLH motifs having introns is more than many other insect species and only inferior to pea aphid. However, the average length of *PhcbHLH* introns is the least among these five insect specifies and the shortest length of *PhcbHLH* intron is just higher than those of pea aphid and fruit fly. Whether this has any evolutionary significance remains for future exploration.

It should be noted that coding regions of five PhcbHLH motifs, namely PhcASCa2, PhcHES3, PhcHES4, PhcHES7, and PhcHES8, were identified from trace-whole-genome shotgun nucleotide sequences (Table 2). We included them as bHLH members because their motif sequences are different with other identified PhcbHLH motifs. Whether they are genuine novel bHLH family members awaits further verification upon completion of genome sequence assembly with higher quality.

A Comparison on Insect bHLH Family Members. So far, bHLH repertoires have been established for 10 insect species, namely *P. humanus corporis* (Phc, *Aedes aegypti* (Aa), *An. gambiae* (Ag), *Culex quinquefasciatus* (Cq), *H. saltator* (Pa), *A. mellifera* (Am), *Acy. pisum* (Ap), *B. mori* (Bm), *D. melanogaster* (Dm), and *T. castaneum* (Tc). The numbers of bHLH family members in each of the 10 insect species are listed in Table 3. Table 3 displays that all insect species lack bHLH genes of Olig, MyoRb, and Figα families. Many families have at least one gene including E12/E47, Ngn, Mist, Beta3, Atonal, Net, MyoRa, Twist, PTFa, PTFb, Hand, SCL, NSCL, Mnt, Max, Myc, USF, AP4, TF4, SREBP, SRC, Clock, AHR, Sim, Trh, HIF, ARNT, BMAL, Emc, Hey, and H/E(spl), among which the 10 insect species have the same number of genes in 10 families, such as E12/E47, Beta3, Net, MyoRa, Hand, SCL, NSCL, Myc, SRC, HIF, and ARNT. However, we failed to identify any Paraxis family member in human body louse. Although all other nine insect species have been found to have one Paraxis family

member, the absence of it in human body louse is probably due to incompleteness of the louse genome sequences. Therefore, it is expected to find this absent bHLH member after a new and higher quality version of human body louse genome sequences is released. Similar situation is also present in *A. pisum*, which lacks ASCa, MyoD, and Microphthalmia transcription factor (MITF) family members, and in *T. castaneum,* which lacks Mesp and MLX family members, both of which should be due to incompleteness of genome sequences as well. Coding regions of these missing bHLH members are expected to be present in genome sequences of higher quality.

Table 3 also presents that the number of H/E(spl) family members varies greatly among different insect species. It ranges from 4 in mosquitoes to 11 or 12 in fruit fly. Human body louse has eight H/E(spl) family members, being second only to that of fruit fly. In insects, there are four different bHLH genes in H/E(spl) family. They are genes *H* (hairy), *Dpn* (deadpan), *Side* (similar to deadpan), and *E(spl)* (enhancer of split). A close examination to distribution of insect bHLH genes in H/E(spl) family revealed that human body louse has one or two more *H* genes than other insect species (Table 4). A phylogenetic tree constructed using all H/E(spl) family members of the 10 insect species demonstrated that the three *P. humanus corporis H* genes, i.e., *PhcHES1*, *PhcHES3*, and *PhcHES4*, were from species-specific gene duplication in louse lineage (Fig. 4). Nevertheless, the three louse *E(spl)* genes, i.e., *PhcHES6*, *PhcHES7*, and *PhcHES8*, were also originated from a species-specific gene duplication, whereas *E(spl)* genes in other insect species except mosquitoes were derived from various *E(spl)* genes already existed in common ancestor of insects (Fig. 4).

Taken together, it is possible that human body louse has more *hairy* and *E(spl)* genes than other insect species. Why does human body louse have such additional genes? It could be the consequence of long period adaptation to relatively dark and stable environment on human body. First, *hairy* gene is involved in negative regulation of insect eye development. It was found that *Drosophila* hairy negatively regulates progression of the morphogenetic furrow across the eye imaginal disc (Brown et al. 1995) and is able to restrain proneural pathways whose activation is imminent (Greenwood and Struhl 1999). Therefore, the

Table 2. Coding regions, intron location, and length of 55 PhcbHLH motifs

Family	PhcbHLH name	Genomic coding sequence(s)			Intron location and length	Group
		Contig no.	Frame	Coding region(s)		
ASCa	*PhcASCa1*	NW_002987838.1	−3	52535–52332		A
	PhcASCa2	gnl|ti|1388835101	+3	9–197		A
ASCb	*PhcASCb*	NW_002987838.1	−3	32987–32796		A
E12/E47	*PhcE12/E47*	NW_002987410.1	+3	80970–80999	Basic: 298 bp	A
			+2	81299–81430		
MyoD	*PhcMyoD*	NW_002987763.1	−2	187468–187424	Helix 1: 1121 bp	A
			−2	186304–186194		
Ngn	*PhcNgn*	NW_002987887.1	+2	325373–325531		A
Mist	*PhcMist1*	NW_002987868.1	−2	333881–333819	Helix 1: 852 bp	A
			−3	332968–332873		
	PhcMist2	NW_002987883.1	−1	511757–511695	Helix 1: 91 bp	A
			−3	511605–511510		
Beta3	*PhcBeta3*	NW_002987848.1	+2	212459–212620		A
Atonal	*PhcAtonal1*	NW_002987189.1	+3	391179–391337		A
	PhcAtonal2	NW_002987340.1	+3	463830–463988		A
	PhcAtonal3	NW_002987189.1	+1	374464–374622		A
Net	*PhcNet*	NW_002987093.1	+2	1195364–1195522		A
MyoRa	*PhcMyoRa*	NW_002987286.1	−2	238480–238322		A
Delilah	*PhcDelilah*	NW_002987276.1	−2	1247771–1247598		A
Mesp	*PhcMesp*	NW_002987756.1	+3	385254–385384	Loop: 66 bp	A
			+3	385451–385481		
Twist	*PhcTwist*	NW_002987470.1	+1	134428–134583		A
PTFa	*PhcPTFa*	NW_002987835.1	−2	292920–292857	Helix 1: 80 bp	A
			−3	292774–292680		
PTFb	*PhcPTFb1*	NW_002987835.1	+2	429542–429585	Helix 1: 91 bp	A
			+3	429677–429791		
	PhcPTFb2	NW_002987093.1	+1	1292365–1292523		A
Hand	*PhcHand*	NW_002987060.1	−3	6547–6389		A
SCL	*PhcSCL*	NW_002987011.1	−2	208837–208698	Helix 2: 70 bp	A
			−3	208627–208609		
NSCL	*PhcNSCL*	NW_002987206.1	+3	27501–27659		A
Mnt	*PhcMnt*	NW_002987887.1	−2	2506229–2506080	Helix 2: 715 bp	B
			−3	2505364–2505356		
Max	*PhcMax*	NW_002987187.1	+3	39072–39230		B
Myc	*PhcMyc*	NW_002987868.1	−1	246303–246145		B
USF	*PhcUSF*	NW_002987276.1	+1	1251733–1251852	Loop: 80 bp	B
			+3	1251933–1251983		
MITF	*PhcMITF1*	NW_002987888.1	+3	11028–11049	Basic: 91 bp	B
			+2	11142–11217	Loop: 248 bp	
			+1	11466–11529		
	PhcMITF2	NW_002987853.1	+1	67453–67632		B
AP4	*PhcAP4*	NW_002987887.1	+3	751803–751913	Loop: 120 bp	B
			+3	752934–752978		
TF4	*PhcTF4*	NW_002987246.1	+1	110821–110982	Helix 2: 97 bp	B
			+2	111080–111088		
MLX	*PhcMLX*	NW_002987392.1	−2	14639–14529	Loop: 80 bp	B
			−3	14452–14399		
SREBP	*PhcSREBP*	NW_002987101.1	+1	4135–4142	Basic: 184 bp	B
			+2	4327–4417	Loop: 167 bp	
			+1	4585–4638		
SRC	*PhcSRC*	NW_002987817.1	+1	428098–428256	Helix 2: 485 bp	B
			+3	428742–428759		
Clock	*PhcClock1*	NW_002987764.1	+1	334987–334991	Basic: 77 bp	C
			+3	335069–335216		
	PhcClock2	NW_002987837.1	+3	382617–382778		C
AHR	*PhcAHR1*	NW_002987077.1	+2	704759–704920		C
	PhcAHR2	NW_002987288.1	−2	437888–437727		C
Sim	*PhcSim*	NW_002987052.1	+1	136939–137100		C
Trh	*PhcTrh*	NW_002987890.1	+2	145844–146005		C
HIF	*PhcHIF*	NW_002987862.1	+2	295751–295912		C
ARNT	*PhcARNT*	NW_002987846.1	+1	153736–153740	Basic: 6723 bp	C
			+1	160464–160620		
BMAL	*PhcBMAL*	NW_002987878.1	−1	397722–397561		C
Emc	*PhcEmc*	NW_002987286.1	−1	213896–213798		D
Hey	*PhcHey1*	NW_002987798.1	+3	39453–39620		E
	PhcHey2	NW_002987097.1	−3	60506–60339		E
H/E(spl)	*PhcHES1*	NW_002987134.1	−2	206788–206783	Basic: 2208 bp	E
			−3	204576–204481	Loop: 105 bp	
			−1	204377–204306		
	PhcHES2	NW_002987021.1	+3	88623–88628	Basic: 497 bp	E
			+2	89126–89221	Loop: 2778 bp	
			+2	92000–92071		

(continued)

Table 2. Continued

Family	PhcbHLH name	Genomic coding sequence(s)			Intron location and length	Group
		Contig no.	Frame	Coding region(s)		
	PhcHES3	gnl|ti|1367165743	−2	637–536	Loop: 100 bp	E
			−1	437–366		
	PhcHES4	gnl|ti|1367221341	+3	192–293	Loop: 98 bp	E
			+2	392–463		
	PhcHES5	NW_002987882.1	+2	93188–93193	Basic: 148 bp	E
			+3	93342–93515		
	PhcHES6	NW_002987423.1	+1	25210-25215	Basic: 107 bp	E
			+1	25342–25509		
	PhcHES7	gnl|ti|1366788626	−3	946–773		E
	PhcHES8	gnl|ti|1382178459	+2	137–310		E
COE	*PhcCOE*	NW_002987086.1	+3	85044–85133	Loop: 76 bp	F
			+1	85210–85254		

Table 3. The bHLH family members in 10 insect species

Group	Family name	Phc	Aa	Ag	Cq	Pa	Am	Ap	Bm	Dm	Tc
A	ASCa	2	3	2	3	2	2	0	4	4	2
	ASCb	1	1(?)	0	1(?)	0	0	1	0	0	1(?)
	E12/E47	1	1	1	1	1	1	1	1	1	1
	MyoD	1	1	1	1	1	1	0	1	1	1
	Ngn	1	1	2	2	1	1	1	1	1	1
	NeuroD	0	1	0	0	1	0	0	0	0	1
	Mist	2	1	1	1	2	2	2	1	1	1
	Beta3	1	1	1	1	1	1	1	1	1	1
	Atonal	3	5	4	5	3	3	3	1	3	3
	Olig	0	0	0	0	0	0	0	0	0	0
	Net	1	1	1	1	1	1	1	1	1	1
	MyoRa	1	1	1	1	1	1	1	1	1	1
	MyoRb	0	0	0	0	0	0	0	0	0	0
	Delilah	1	1	1	1	0	0	1	1	1	2
	Mesp	1	1	1	1	1	1	1	1	1	0
	Paraxis	0	1	1	1	1	1	1	1	1	1
	Twist	1	1	1	1	2	1	1	1	1	1
	PTFa	1	1	2	1	1	1	1	1	1	1
	PTFb	2	2	2	2	2	1	2	1	2	2
	Hand	1	1	1	1	1	1	1	1	1	1
	SCL	1	1	1	1	1	1	1	1	1	1
	NSCL	1	1	1	1	1	1	1	1	1	1
B	Mnt	1	1	1	1	2	1	1	1	1	1
	Mad	0	0	0	0	0	0	1	0	0	1
	Max	1	1	1	1	2	1	3	1	1	1
	Myc	1	1	1	1	1	1	1	1	1	1
	USF	1	1	1	1	2	2	1	1	1	1
	MITF	2	1	1	1	1	1	0	1	1	1
	AP4	1	1	1	1	2	1	1	1	1	1
	TF4	1	1	1	1	1	1	2	1	1	1
	MLX	1	1	1	1	1	1	1	1	1	0
	SREBP	1	1	1	2	1	1	1	1	1	1
	Figα	0	0	0	0	0	0	0	0	0	0
	SRC	1	1	1	1	1	1	1	1	1	1
C	Clock	2	2	2	2	2	2	2	3	3	2
	AHR	2	2	2	2	3	2	2	3	2	1
	Sim	1	1	2	1	1	1	1	1	1	2
	Trh	1	1	1	2	1	1	1	1	1	1
	HIF	1	1	1	1	1	1	1	1	1	1
	ARNT	1	1	1	1	1	1	1	1	1	1
	BMAL	1	1	1	1	1	1	1	2	1	1
D	Emc	1	1	2	1	1	1	1	1	1	1
E	Hey	2	3	3	3	2	2	3	2	1(2?)	2(1?)
	H/E(spl)	8	4	4	4	6	6	6	5	11(10?)	5(6?)
F	COE	1	1	1	1	1	1	1	1	1	1
	Total	55	55	55	57	57	51	54	52	59	53

Data of Phc (*P. humanus corporis*) were obtained in this study. Those of Am (*A. mellifera*), Ap (*Acy. pisum*), Pa (*H. saltator*), and Bm (*B. mori*) were from our published data (Wang et al. 2008, 2009; Dang et al. 2011a; Liu et al. 2012) and those of Dm (*D. melanogaster*) and Tc (*T. castaneum*) were from Simionato et al. (2007) and Bitra et al. (2009), respectively. Data of Aa (*Ae. aegypti*), Ag (*An. gambiae*), and Cq (*Culex quinquefasciatus*) were from our recently published data (Zhang et al. 2013). The question mark indicates family classification with uncertainty.

Table 4. Distribution of insect bHLH genes in H/E(spl) family

Gene name	Phc	Aa	Ag	Cq	Pa	Am	Ap	Bm	Dm	Tc
H	4(2?)	1	1	1	2(1?)	1	1	1	1	1
Dpn	0(1?)	1	1	1	0(1?)	1	1	1	1	1
Side	1	1	1	1	1	1	1(2?)	0	1	2
E(spl)	3	1	1	1	3	3	3(2?)	3	8(9?)	2(3?)
Total	8	4	4	4	6	6	6	5	11(12?)	6(7?)

The question mark indicates family classification with uncertainty.
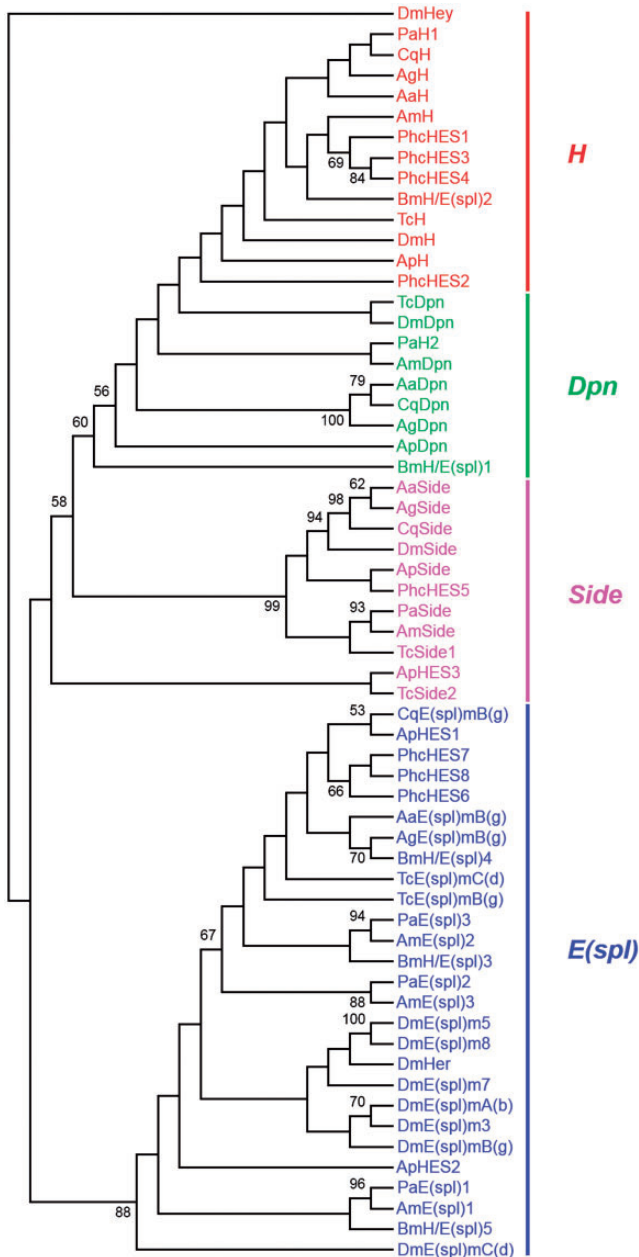
Fig. 4. Evolutionary relationship among insect H/E(spl) family members. An ML tree based on the multiple alignment that includes all members of H/E(spl) family, which has been rooted with the closely related Hey gene from *D. melanogaster,* is shown. Only bootstrap values no less than 50 are shown. *Ae. aegypti*, Aa; *An. gambiae*, Ag; *A. mellifera*, Am; *Acy. pisum*, Ap; *B. mori*, Bm; *Culex quinquefasciatus*, Cq; *D. melanogaster*, Dm; *H. saltator*, Pa; *P. humanus corporis*, Phc; and *T. castaneum*, Tc. bHLH names of *H, Dpn, Side,* and *E(spl)* genes are represented with red, green, magenta, and blue characters, respectively.

existence of multiple *hairy* genes probably means that eye development of human body louse is hindered for adaptation to dark environment. Second, *E(spl)* gene inhibits differentiation of specific preneural cells. *Drosophila* E(spl)mB(g) was found to be a potent inhibitor to prevent ectodermal cells from adopting the sensory organ precursor fate (Giagtzoglou et al. 2003). Although various *E(spl)* genes are functionally redundant (Schrons et al. 1992), the existence of three *E(spl)* genes in human body louse probably indicates that specific preneural cells are inhibited from forming functional sensory apparatus, leading to deficiency in sensory functions other than photoreception.

In this study, *P. humanus corporis* genome sequences were searched and identified to encode 55 members of the bHLH superfamily. Phylogenetic analyses revealed that the 55 PhcbHLHs are distributed in 39 bHLH families with 23, 11, 9, 1, 10, and 1 member(s) in groups A, B, C, D, E, and F, respectively. Group C and E PhcbHLH proteins were found to possess PAS/PAC and Orange domains, respectively, further verifying our classification to the identified PhcbHLH family members. Examination to GenBank annotations of the 55 PhcbHLH members indicated that 29 PhcbHLH proteins were annotated in consistence with our analytical result, 8 were annotated different with our analytical result, 12 were merely annotated as hypothetical protein, and the rest 6 were not deposited in GenBank. A comparison on insect bHLH gene composition revealed that human body louse possibly has more *hairy* and *E(spl)* genes than other insect species. Because *hairy* and *E(spl)* genes have been found to negatively regulate the differentiation of insect preneural cells, it is suggested that the existence of additional *hairy* and *E(spl)* genes in human body louse is probably the consequence of long-period adaptation to relatively dark and stable environment on human body. These data provide good reference for further studies on regulatory functions of bHLH proteins in the growth and development of human body louse.

References Cited

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21: 2104–2105.

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: 501–516.

Bailey, P. C., C. Martin, G. Toledo-Ortiz, P. H. Quail, E. Huq, M. A. Heim, M. Jakoby, M. Werber, and B. Weisshaar. 2003. Update on the basic helix-loop-helix transcription factor gene family in *Arabidopsis thaliana*. Plant Cell 15: 2497–2501.

Bitra, K., A. Tan, A. Dowling, and S. R. Palli. 2009. Functional characterization of PAS and HES family bHLH transcription factors during the metamorphosis of the red flour beetle, *Tribolium castaneum*. Gene 448: 74–87.

Brown, N. L., S. A. Sattler, S. W. Paddock, and S. B. Carroll. 1995. hairy and emc negatively regulate morphogenetic furrow progression in the developing *Drosophila* eye. Cell 80: 879–887.

Castillon, A., H. Shen, and E. Huq. 2007. Phytochrome interacting factors: central players in phytochrome-mediated light signaling networks. Trends Plant Sci. 12: 514–521.

Crews, S. T. 1998. Control of cell lineage-specific development and transcription by bHLH-PAS proteins. Genes Dev. 12: 607–620.

Dang, C. W., Y. Wang, K. P. Chen, Q. Yao, D. B. Zhang, and M. Guo. 2011a. The basic helix-loop-helix transcription factor family in the pea aphid, *Acyrthosiphon pisum*. J. Insect 11: 84.

Dang, C. W., Y. Wang, D. B. Zhang, Q. Yao, and K. P. Chen. 2011b. A genome-wide survey on basic helix-loop-helix transcription factors in giant panda. PLoS One 6: e26878.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14: 685–695.

Giagtzoglou, N., P. Alifragis, K. A. Koumbanakis, and C. Delidakis. 2003. Two modes of recruitment of E(spl) repressors onto target genes. Development 130: 259–270.

Greenwood, S., and G. Struhl. 1999. Progression of the morphogenetic furrow in the *Drosophila* eye: the roles of Hedgehog, Decapentaplegic and the Raf pathway. Development 126: 5795–5808.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59: 307–321.

Gyoja, F.,T. Kawashima, and N. Satoh. 2012. A genomewide survey of bHLH transcription factors in the coral *Acropora digitifera* identifies three novel orthologous families, pearl, amber, and peridot. Dev. Genes Evol. 222: 63–76.

James, W. D., T. G. Berger, and D. M. Elston. 2011. Andrews' diseases of the skin, 11th Edition. Saunders Elsevier, Philadelphia, USA.

Jones, S. 2004. An overview of the basic helix-loop-helix proteins. Genome Biol. 5: 226.

Kadesh, T. 1993. Consequences of heterodimeric interactions among helix-loop-helix proteins. Cell Growth Differ. 4: 49–55.

Kirkness, E. F., B. J. Haas, W. L. Sun, H. R. Braig, M. A. Perotti, J. M. Clark, S. H. Lee, H. M. Robertson, R. C. Kennedy, E. Elhaik, et al. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc. Natl. Acad. Sci. USA. 107: 12168–12173.

Le, S. Q., and O. Gascuel. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25: 307–1320.

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

Ledent, V., O. Paquet, and M. Vervoort. 2002. Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biol. 3: RESEARCH0030.

Levot, G. 2000. Resistance and control of lice on humans and production animals. Int. J. Parasitol. 30: 291–297.

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

Liu, A. K., Y. Wang, C. W. Dang, D. B. Zhang, H. F. Song, Q. Yao, and K. P. Chen. 2012. A genome-wide identification and analysis of the basic helix-loop-helix transcription factors in the ponerine ant, *Harpegnathos saltator*. BMC Evol. Biol. 12: 165.

Massari, M. E., and C. Murre. 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. Mol. Cell. Biol. 20: 429–440.

Murre, C., P. S. McCaw, H. Vaessin, M. Caudy, L. Y. Jan, Y. N. Jan, C. V. Cabrera, J. N. Buskin, S. D. Hauschka, A. B. Lassar, et al. 1989. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. Cell 58: 537–544.

Nicholas, K. B., H. B. Nicholas, Jr. and D. W., Deerfield I 1997. GeneDoc: analysis and visualization of genetic variation. Embnet News 4: 14.

Pedra, J. H. F., A. Brandt, H. M. Li, R. Westerman, J. Romero-Severson, R. J. Pollack, L. L. Murdock, and B. R. Pittendrigh 2003. Transcriptome identification of putative genes involved in protein catabolism and innate immune response in human body louse (Pediculidae: Pediculus humanus). Insect Biochem. Mol. 33: 1135–1143.

Pires, N., and L. Dolan. 2010. Origin and diversification of basic-helix-loop-helix proteins in plants. Mol. Biol. Evol. 27: 862–874

Robinson, K. A., and J. M. Lopes. 2000. Survey and summary: *Saccharomyces cerevisiae* basic helix-loop-helix proteins regulate diverse biological processes. Nucleic Acids Res. 28: 1499–1505.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–425.

Satou, Y., K. S. Imai, M. Levine, Y. Kohara, D. Rokhsar, and N. Satoh. 2003. A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. I. Genes for bHLH transcription factors. Dev. Genes Evol. 213: 213–221.

Schrons, H., E. Knust, and J. A. Campos-Ortega. 1992. The enhancer of split complex and adjacent genes in the 96F region of *Drosophila melanogaster* are required for segregation of neural and epidermal progenitor cells. Genetics 132: 481–503.

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

Swofford, D. L. 1998. PAUP. Phylogenetic analysis using parsimony, Version 4. Sinauer Associates, Massachusetts, USA.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28: 2731–2739.

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

Toups, M. A., A. Kitchen, J. E. Light, and D. L. Reed. 2011. Origin of clothing lice indicates early clothing use by anatomically modern humans in Africa. Mol. Biol. Evol. 28: 29–32.

Wang, Y., K. Chen, Q. Yao, W. Wang, and Z. Zhi. 2007. The basic helix-loop-helix transcription factor family in *Bombyx mori*. Dev. Genes Evol. 217: 715–723.

Wang, Y., K. Chen, Q. Yao, W. Wang, and Z. Zhu. 2008. The basic helix-loop-helix transcription factor family in the honey bee, *Apis mellifera*. J. Insect Sci. 8: 1–12.

Wang, Y., K. Chen, Q. Yao, X. Zheng, and Z. Yang. 2009. Phylogenetic analysis of zebrafish basic helix-loop-helix transcription factors. J. Mol. Evol. 68: 629–640.

Zhang, D. B., Y. Wang, A. K. Liu, X. H. Wang, C. W. Dang, Q. Yao, and K. P. Chen. 2013. Phylogenetic analyses of vector mosquito basic helix-loop-helix transcription factors. Insect Mol. Biol. 22: 608–621.

Zheng, X. D., Y. Wang, Q. Yao, Z. Yang, and K. P. Chen. 2009. A genome-wide survey on basic helix-loop-helix transcription factors in rat and mouse. Mamm. Genome 20: 236–246.