A genomewide survey of basic helix-loop-helix factors in Drosophila

Adrian W. Moore, Sandra Barbel, Lily Yeh Jan, and Yuh Nung Jan*

Howard Hughes Medical Institute, Departments of Physiology and Biochemistry, University of California, San Francisco, CA 94143-0725

Contributed by Yuh Nung Jan, June 29, 2000

The basic helix-loop-helix (bHLH) transcription factors play important roles in the specification of tissue type during the development of animals. We have used the information contained in the recently published genomic sequence of Drosophila melanogaster to identify 12 additional bHLH proteins. By sequence analysis we have assigned these proteins to families defined by Atonal, Hairy-Enhancer of Split, Hand, p48, Mesp, MYC/USF, and the bHLH-Per, Arnt, Sim (PAS) domain. In addition, one single protein represents a unique family of bHLH proteins. mRNA in situ analysis demonstrates that the genes encoding these proteins are expressed in several tissue types but are particularly concentrated in the developing nervous system and mesoderm.

The basic helix-loop-helix (bHLH) proteins comprise an evolutionarily ancient, important group of transcription factors in animals, plants, and fungi. Their functions range from control of cellular proliferation to tissue differentiation. They are united by conservation solely within the bHLH domain (1) and act as dimers binding the E-box site CANNTG to regulate transcription (2, 3). The bHLH domain consists of the basic domain, a run of approximately 15 aa with a high number of basic residues, followed by a loop region of variable length. Interaction between the helix regions of two different proteins leads to intermolecular dimerization, and the basic region of each partner binds to half of the E-box sequence (4–6).

In animals, bHLH proteins often are used in cascades to specify tissue identity. Furthermore, closely related families of bHLH proteins as defined by their level of identity in the bHLH domain tend to have functions in a similar tissue type. For example, in Drosophila, Twist is required for mesoderm specification. It then acts alongside (the MyoD family ortholog) Nautilus (Nau) in myogenesis. Similarly, in vertebrates the initial specification and division of the mesoderm involves the activity of the MyoD family orthologs such as E47 or E12 in vertebrates (13) to repress transcription via interaction with the non-bHLH Groucho protein (18).

Genes of the MYC family of bHLH proteins contain a leucine zipper (zip) immediately C-terminal to the bHLH domain. MYC forms a heterodimeric transcription factor with the related MAX or MAD proteins to regulate cell proliferation (19). The mammalian upstream transcription factor (USF) proteins also have a bHLH-zip structure and are proposed to act to antagonize the proliferation function of MYC (20).

A further group of bHLH proteins contains a Per, Arnt, Sim (PAS) domain C-terminal to the bHLH domain. This functions to mediate protein–protein interaction. The bHLH-PAS proteins form heterodimeric transcription factors with roles in development [e.g., Tango (21)], toxin metabolism [e.g., Ral1(1)H (22)], and the regulation of circadian rhythm-expressed genes (23).

The recent publication of the entire heterochromatinic sequence of Drosophila melanogaster (24) presents an opportunity to analyze those members of the bHLH family as yet unstudied. We have used the information contained in the bHLH transcription factors act in a cell-autonomous manner. Hence, the in situ data (Table 1 and Fig. 3) along with the sequence analysis (Figs. 1 and 2) presented in this paper give clues about the developmental roles of these newly identified proteins in Drosophila. The proteins identified in this study also may facilitate the search for their orthologs in vertebrate species. Finally, using a genomewide survey we can ask whether all tissue types in the developing Drosophila embryo have bHLH cascades associated with their genesis.

Materials and Methods

Database Search. We used the BLAST search (25) and PATTERN search programs provided by The National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) and Berkeley Drosophila Genome Project (http://www.fruitfly.org) to search the recently published sequence and predicted protein information from the entire Drosophila genome (24). As query sequence, we used the amino acid sequences of the bHLH domains of several known bHLH proteins and sequences based on the predicted degenerate sequence for all bHLH domains (26). We restricted our search to those proteins with an intact basic domain.

Tree Building and Sequence Lineups. We constructed a database consisting of the amino acid sequences of the bHLH domains of the 12 proteins reported in this study and the closest related proteins to these (identified by BLAST search). In addition, we added all known Drosophila bHLH-containing proteins and representative members of all of the different bHLH domain-
containing families (27). The sequences were aligned and
the tree was constructed by using the CLUSTALX program (28).
Gaps in the sequence alignment were ignored in the tree-building
process. To increase the predictive sequence available to the
CLUSTALX program, the genes of the AS complex (AS-C) were
not included in the data set for tree drawing. Ten trees were
constructed by using different randomizations of sequence input
order and bootstrapped 1,000 times to give an indication of the
predictive strength of each node.

In Situ Analysis. Analysis of mRNA distribution in
Drosophila embryos was carried out as described in ref. 29. As hybridization
probes, we used digoxigenin-labeled PCR product amplified by
using primers designed to the coding region of each of these
genes and either Drosophila genomic DNA or embryo-derived
cDNA as template.

Results
We have identified 12 additional proteins of the bHLH class and
estimate from surveying the genome that the total number of
bHLH genes in Drosophila with an intact basic domain is 46. We
used the neighbor-joining method to classify these 12 proteins
into families (28) (Fig. 1). The publication of the genomic
sequence of Drosophila included an initial annotation of the
sequence (24). This study identified 69 putative proteins with a
HLH dimerization domain (30) that are listed at the Genome
Annotation Database of Drosophila (GadFly) web site (http://
www.fruitfly.org/annot/). Eleven of the proteins identified in
this study are a subset of the HLH proteins identified; CG18144
was not identified in the initial genome annotation analysis.
In two cases, CG5952/Fer2 and CG17592/Dm Usf, the pre-
dicted protein sequence contained inserted sequences that dis-
rupted the basic helix1–loop–helix2 domains. Using embryonic
cDNA as a template, products spanning the region of doubt were
amplified by PCR. Subsequent sequencing of these products dem-
strated that the predicted sequence of CG17592/Dm Usf is
correct. The sequence of CG5952/Fer2 presented in this paper has
been altered from that predicted from the genomic sequence.

CG8667 (Mistr) and CG5545 (Doli), Part of the Ato-Related Family, Are
Expressed in the Developing Nervous System. CG5545 is closely
related to the vertebrate Beta 3 protein, a repressor molecule
(31) (96% sequence identity in the bHLH domain), and the Olig
proteins involved in oligodendritic precursor formation (32, 33)
(Fig. 2a). We suggest that this protein should be named Doli
(Drosophila Olig family). CG8667 has closest sequence identity
to the vertebrate Mist1 protein (34), a negative regulatory factor
of MyoD activity (78% identical over the entire bHLH domain
and 92% identical in the basic domain alone). We propose that
this protein should be named Mistr (Mist 1-related protein).
As with the other proteins of the Ato-related family, the genes

Table 1. Summary of gene expression using the expression code of Hartenstein and Jan (48)

<table>
<thead>
<tr>
<th>Gadfly no.</th>
<th>Name</th>
<th>mRNA expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG8667</td>
<td>Mistr</td>
<td>N7+ring gland:9</td>
</tr>
<tr>
<td>CG5545</td>
<td>Doli</td>
<td>N1:9</td>
</tr>
<tr>
<td>CG10066</td>
<td>Fer1</td>
<td>Epidermis:15</td>
</tr>
<tr>
<td>CG5952</td>
<td>Fer2</td>
<td>N5:11</td>
</tr>
<tr>
<td>CG6913</td>
<td>Fer3</td>
<td>M:11</td>
</tr>
<tr>
<td>CG18144</td>
<td>Dm Hand</td>
<td>N5:13 M1:10 DU2:10</td>
</tr>
<tr>
<td>CG10446</td>
<td>Side</td>
<td>N5:12</td>
</tr>
<tr>
<td>CG5927</td>
<td>Her</td>
<td>Ubiquitous</td>
</tr>
<tr>
<td>CG12952</td>
<td>Sage</td>
<td>SH3:10</td>
</tr>
<tr>
<td>CG6211</td>
<td>Gce</td>
<td>GO:9</td>
</tr>
<tr>
<td>CG11450</td>
<td>Shout</td>
<td>SM1:5 AS:5</td>
</tr>
<tr>
<td>CG17592</td>
<td>Dm Usf</td>
<td>N7+proventriculus:14</td>
</tr>
</tbody>
</table>

Bold type indicates expression pattern code; normal type indicates the stage at which expression is initiated.
encoding these proteins are expressed in the developing Drosophila nervous system. CG5545/doli is expressed first in a subset of cells in both the ventral nerve cord (VNC) and the procephalic region at stage 9. The number of cells in these regions expressing the gene increases to a peak at stage 11 (Fig. 3a). By stage 14, levels of expression have fallen such that CG5545/doli is expressed only in a few cells per hemisegment on the ventral surface of the VNC.

There is a strong maternal contribution of CG8667/mistr mRNA. Zygotic transcription is initiated at stage 14. It is expressed in bilateral domains in the cephalic region, which, as development proceeds, fuse into a U shape forming part of the ring gland (Fig. 3b). Concomitant expression of CG8667/mistr also begins in the CNS. By stage 17, CG8667/mistr is in clusters of cells at the anterior and posterior of the VNC and bilaterally in two lateral cells per hemisegment in the VNC.

CG10066 (Fer1), CG5952 (Fer2), and CG6913 (Fer3) Are Related to Mammalian p48. Three new bHLH proteins are most closely related to the bHLH domain of the p48 subunit of PTF1, a pancreatic, exocrine cell-specific transcription factor in the mouse (35), and represent a new bHLH family in Drosophila. We name these proteins Fer for Drosophila (Fig. 3a). By stage 14, levels of expression have fallen such that CG5545/doli is expressed only in a few cells per hemisegment on the ventral surface of the VNC.

CG6913/fer3 is expressed at stage 11 in part of the posterior midgut primordia and stage 12 in part of the anterior midgut primordia. At later stages, we have detected expression in several unidentified cells scattered throughout the embryo.

CG10446 (Side) and CG5927 (Her) Are in the HES Family. CG10446 is most closely related to Dpn (76% identity in the basic and 62% in the entire bHLH domain). We name this protein Side (similar to Deadpan). CG5927 is most closely related to the proteins of the Enhancer of split [E(spl)] complex, such as HLHm17 (76% identity in the basic and 51% identity in the entire bHLH domain) (Fig. 2d). We name CG5927 Her (HES-related).

There is a strong maternal contribution of CG10446/side mRNA. Zygotic transcription of the gene begins at stage 12 in a subset of cells in the CNS (Fig. 3e). CG5927/her has a low level of maternal mRNA contribution and then is expressed ubiquitously throughout embryogenesis.

CG12592 (Sage) Is Distantly Related to the Mesp Family and Expressed in the Salivary Gland. CG12592 represents a protein with little sequence similarity to other known proteins. In the neighboring treelike, it is placed in the same family as the vertebrate Mesp proteins, which are necessary for mesoderm segmentation initiation (53% identity in the bHLH domains) (Fig. 2f) (36). CG12592 has a strong maternal mRNA contribution in early embryogenesis. Its zygotic expression begins in the salivary gland anlage at stage 10 and persists until stage 15 (Fig. 3f). We name CG12592 Sage (salivary gland-expressed bHLH).

CG17592 (Dm Usf) Is the Ortholog of the Mammalian USF Proteins. CG17592 is the single Drosophila sequence homologue of the vertebrate USF proteins that are involved in cell proliferation control (92% identical in the basic domain) (Fig. 2f) (20). We term this protein Dm Usf. Both vertebrate and Drosophila USF
are bHLH-zip proteins. Dm Usf has a loop and a second helix region, high in serines, which is greatly diverged from that of mouse and human and, hence, may have lost its ability to dimerize (Fig. 2h). There is a weak maternal contribution of Dm usf mRNA. At stage 7, Dm usf is expressed in bilateral domains in the ventral cephalic furrow. In later stages (15 onward) of development, Dm usf expression is confined to the proventriculus and a subset of cells in the CNS (Fig. 3g). This specific expression pattern differs from the ubiquitous USF expression pattern reported in vertebrates.

CG6211 (Gce) Is Closely Related to the bHLH-PAS Rst(1)JH Protein. CG6211 is closely related to the Rst(1)JH protein, a bHLH-PAS protein (78% identity in the bHLH, 68% in the PAS-A, and 86% in the PAS-B)
in the PAS-B domains) (Fig. 2f). Rst(1)JH originally was isolated in a screen to find Drosophila resistant to the Juvenile Hormone Analog insecticide Methoprene. CG6211 transcript is expressed strongly as a maternally supplied message and then later in a subset of the germ cells of the developing embryo (Fig. 3h). We suggest that this protein should be named Gce (germ cell-expressed bHLH-PAS).

CG11450 (shout) Is Expressed During Mesoderm Formation and in Myoblasts. CG11450 represents a member of a new bHLH family (Figs. 1 and 2g). It is expressed first in the dorsal and ventral cellular blastoderm. In the ventral region of the embryo, the gene is expressed continually in the presumptive mesoderm throughout gastrulation and then in a segmented pattern in the ventral mesoderm layer at the extended germ-band stage. It is expressed in the myoblast cells that then migrate dorsally from this layer (Fig. 3i–k). The expression pattern of CG11450 overlaps with that of the bHLH transcription factor twist, suggesting that it may be playing a role in the same mesoderm specification and myogenic pathways; therefore, we term this gene shout after “Twist and Shout” by Lennon and McCartney (1963).

CG18144 (Dm Hand) Is the Drosophila Ortholog of the Vertebrate Hand Proteins. CG18144 (Dm Hand) represents the single Drosophila ortholog of the vertebrate dHand (76% identity) and eHand (69% identity in the bHLH domain) proteins involved in heart formation (37) (Fig. 2c). Dm hand expression begins at stage 10 of embryonic development in bilateral stripes in the ventral mesoderm. It continues to be expressed in two tissues derived from this mesoderm, the dorsal vessel (heart) and the circular visceral musculature. In addition, at stage 13 Dm hand mRNA appears in a small subset of cells in the CNS (Fig. 3l–o).

Discussion

In this study we have identified 12 additional bHLH proteins and examined the expression pattern of the genes that encode them. The majority of the proteins we identified are members of previously identified families. In these cases, the available knowledge may be known members of these families, together with the expression pattern of these newly identified genes, allows one to make predictions concerning their function. For example, all members of the HES proteins mediate transcription repression via their interaction with Groucho. CG10446/Side and CG5952/Her have the WRPW domain required for this interaction, implying that they are highly likely to act via the same mechanism. CG10446/side is expressed solely in the CNS at a stage at which cell differentiation is occurring. We hypothesize that it may play a role in antagonizing the function of transcription factors involved in the later stages of CNS differentiation.

However, sequence homology between species does not always imply functional homology. For example, CG8667/Mistr is a Drosophila sequence ortholog of the mammalian Mist1 protein. It is expressed solely in the developing nervous system, whereas Mist1 is expressed not in the nervous system but in gut, pancreas, submandibular gland, lung, and skeletal muscle (34). In this case, differences in expression pattern of the genes encoding these proteins argue against any conservation of developmental role.

Two developmental processes in which bHLH gene function has been extensively studied in both Drosophila and other model systems are neurogenesis and mesoderm specification/myogenesis. We found potential new components in both of these processes in Drosophila.

Some of the genes we uncovered were expressed solely in the nervous system: CG8667/mistr, CG5545/doli, CG10446/side, and CG5952/her. Although several bHLH orthologs in Drosophila and vertebrates are known to control the formation and differentiation of neurons, similar bHLH factors in the determination of glia have not been identified. The HLF38 bHLH gene is expressed solely in the Drosophila midline glia (A.W.M., unpublished data); however, this gene is the sequence homologue of mammalian hematopoietic stem cell leukemia factor, which has not been described in gliogenesis in mammals (38). In this study, CG5545/Doli is particularly interesting because it is related to the Olig proteins. The Olig genes are expressed in the zone from which oligodendrocyte precursors arise and then in the precursors themselves throughout the CNS in mice and rats (32, 33). Ectopic expression of these genes promotes the expression of some oligodendritic precursor markers. From the mRNA in situ analysis we have carried out for CG5545/doli it is not clear in which cell types in the nervous system this gene is expressed; however, it may have a conserved role in gliogenesis if further experiments demonstrate its expression in a glial cell type.

CG11450/shout is the founding member of a new bHLH family. It is expressed in the ventral-most cells of the blastoderm-stage embryo that are fated to become mesoderm. This expression domain overlaps with that of twist (7). All members of the SHH/Hand family are expressed in the ventral midline and CG11450/shout continue to be expressed in the presumptive mesoderm during gastrulation. At the extended germ-band stage, both twist and CG11450/shout are expressed in alternating high and low levels along the length of the mesoderm. These alternating expression levels of twist are required for the specification of muscle derived from this tissue (39). The pattern of CG11450/shout expression in the ventral mesoderm implies that it could have a similar role to twist in specification of mesoderm derivatives. In Drosophila, Twist activates Snail and other downstream, mesoderm-specific regulators such as Tinman, Bagpipe, and Mef2 (7); all of these proteins have orthologs in vertebrates implicated in mesoderm development. Hence, CG11450/Shout represents a good candidate for both sequence and function conservation across species.

In both Drosophila and vertebrates, the heart (dorsal vessel in Drosophila) assembles at the midline from bilaterally symmetrical ventral mesoderm precursors. In vertebrates, the homeodomain-containing transcription factor Nkx2.5 is the earliest known marker for the cardiac lineage. The d and k orthologs additionally are expressed in the early cardiac progenitors. The Mef2C transcription factor is expressed in the heart myoblasts. In Nkx2.5, dHand, and Mef2C null mice, the early heart tube forms but it fails to undergo looping and express some heart-specific molecules (40–42), implying that these three genes may lie in the same pathway. tinman (tin) is the Drosophila ortholog of Nkx2.5. It is expressed first uniformly in the presumptive mesoderm before gastrulation and in the ventral mesoderm after this process. It is required for the formation of the heart and circular visceral musculature from this mesoderm (43). The mef2 gene in Drosophila is a direct transcription target of Tin and is required for the correct differentiation of the heart myoblasts (44–46). In this study, we identified Drosophila hand and showed that it is expressed in the early heart mesoderm, mirroring the situation in vertebrates. All these data imply that tin, mef2, and Dm hand may lie in a pathway in Drosophila homologous to that proposed in vertebrates. In addition, we also detected Dm hand expression in the forming circular visceral musculature of the Drosophila embryo. Because tin and mef2 also are required for this structure, the same interaction may be conserved in this tissue.

It has been proposed that all regions of the body plan are specified by a hox gene code (47). Do all tissue types in the embryo have a bHLH component to their specification? The answer seems to be no. The expression of factors such as Da, which forms heterodimers with tissue-specific bHLH genes such as those of the Ato-related or MyoD families, is ubiquitous. However, some tissue types such as the fat body (fly...
evidence suggests that bHLH proteins are not involved in the function in all tissues in which it is expressed; loss of Da equivalent of the liver do not express any other tissue-specific bHLH genes with which this gene can act. In fact, Da may not be


We thank Sarah Goulding and members of the Jan lab for useful discussions and Sarah Meadows and Mike Rothenberg for reagents. A.W.M. is supported by a Wellcome Prize Traveling Research Fellowship. L.Y.J. and Y.N.J. are Howard Hughes Investigators.