Gene, 172 (1996) 249–254
© 1996 Elsevier Science B.V. All rights reserved. 0378-1119/96/$15.00

GENE 09658

The *Drosophila melanogaster similar bHLH-PAS* gene encodes a protein related to human hypoxia-inducible factor 1α and *Drosophila* single-minded

(Sima; basic helix-loop-helix; transcription factor; embryogenesis)

John R. Nambu[b,*], Wei Chen[b], Song Hu[b,a] and Stephen T. Crews[a]

[a]*Department of Biochemistry and Biophysics, School of Medicine, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260, USA; and* [b]*Department of Biology, Molecular Biology Institute, UCLA, Los Angeles, CA 90024, USA*

Received by J.A. Hoch: 27 October 1995; Revised/Accepted: 15 December 1995; Received at publishers: 29 January 1996

SUMMARY

The *Drosophila melanogaster* (*Dm*) *similar* (*sima*) gene was isolated using a low-stringency hybridization screen employing a *Dm single-minded* gene basic helix-loop-helix (*bHLH*) DNA probe. *sima* is a member of the *bHLH-PAS* gene family and the conceptual protein shares a number of structural features, including a bHLH domain, PAS domain, and homopolymeric amino acid stretches. Sima is most closely related to the human hypoxia-inducible factor 1α bHLH-PAS protein. In situ hybridization experiments reveal that *sima* is transcribed in most or all cells throughout embryogenesis. It has been cytologically mapped to position 99D on the third chromosome, and is not closely linked to other known *bHLH-PAS* genes.

INTRODUCTION

The bHLH-PAS transcription factor family is comprised of a number of important developmental and physiological regulatory proteins. Developmental regulators include the *Dm* Single-minded (Sim) protein, which is required for central nervous system (CNS) midline cell development (Nambu et al., 1991), and Trachealess (Trh), which is required for tracheal formation (Isaac and Andrew, 1995; Wilk et al., 1995). One important mammalian bHLH-PAS regulatory factor is the aromatic hydrocarbon receptor complex (AHRC), which is involved in the induction of hydrocarbon metabolizing enzyme genes (reviewed in Swanson and Bradfield, 1993). AHRC consists of two bHLH-PAS proteins: the aromatic hydrocarbon receptor (Ahr) and its partner Aryl hydrocarbon receptor nuclear translocator (Arnt). The cellular response to hypoxia in humans is controlled by hypoxia-inducible factors 1α (HIF-1α) and 1β (HIF-1β). HIF-1α is related to Sim, whereas HIF-1β is Arnt (Wang et al., 1995). All of these proteins have an N-terminal bHLH domain followed by the PAS multifunctional interaction domain, which is subdivided into PAS-A and PAS-B subregions. Another *Dm* protein, Period (Per), controls rhythmic behavior, has only a PAS domain, and is a negative regulator of transcription (Rosbash, 1995). The

Correspondence to: Dr. S.T. Crews, Department of Biochemistry and Biophysics, School of Medicine, CB#7260 Faculty Laboratory Office Building, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260, USA. Tel. (1-919) 962-4380; Fax (1-919) 962-3155; e-mail: steve—crews@unc.edu

* Present address: Biology Department, University of Massachusetts at Amherst, Amherst, MA 01003, USA. Tel. (1-413) 545-2782.

Abbreviations: aa, amino acid(s); Ahr, aromatic hydrocarbon receptor; AHRC, Ahr complex; Arnt, aryl hydrocarbon receptor nuclear translocator; bHLH, basic helix-loop-helix; *bHLH*, gene encoding bHLH; bp, base pair(s); CNS, central nervous system; *Dm*, *Drosophila melanogaster*; HIF, hypoxia-inducible factor; kb, kilobase(s) or 1000 bp; *Kina*, gene encoding protein kinase Kina; LZ, leucine zipper; nt, nucleotide(s); ORF, open reading frame; PAS, Pas-Single-minded-Arnt; *PAS*, gene encoding PAS; Per, Period; *per*, gene encoding Per; Sim, Single-minded; *sim*, gene coding for Sim; Sima, Similar; *sima*, gene coding for Sima; SSC, 0.15 M NaCl/0.015 M Na_3·citrate pH 7.6; Trh, Trachealess; *trh*, gene coding for Trh; wt, wild type.

PII S0378-1119(96)00060-1

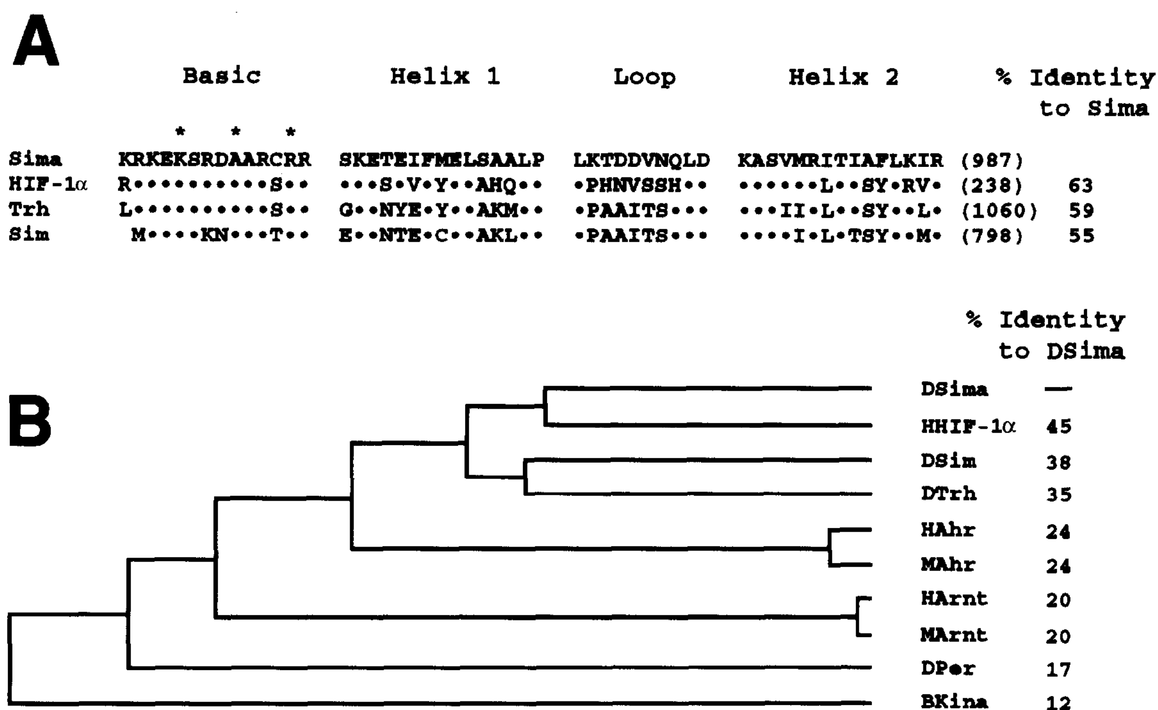

Fig. 1. Sequence relationships between Sima and PAS domain proteins. (A) Sima bHLH region is related in sequence to HIF-1α, Sim, and Trh. The bHLH sequences of the four proteins were aligned and residues identical to Sima are indicated by dots. The nt positions of the sequences shown are indicated in parentheses (GenBank accession No. U43090, Sima; U22431, HIF-1α; U42699, Trh; M19120, Sim), and percent identity within the bHLH region is indicated between Sima and the other three proteins. Asterisks indicate residues of the Max protein sequence that are responsible for DNA contacts. Labels above the sequence indicate the extent of each bHLH sequence domain. (B) PAS domain sequence comparison tree of PAS family proteins. Published PAS domains were aligned and a sequence comparison tree generated. Comparisons used the PAS-A and PAS-B subregions; the poorly conserved spacer region was excluded. Each protein name is preceded by an abbreviation of the organism that the gene was isolated from: *Bacillus subtilis* (B), *D. melanogaster* (D), human (H), and mouse (M). The proteins included are: Ahr, Arnt, HIF-1α, Kina, Per, Sim, Sima, and Trh. The% aa identity between the Sima PAS domain and other PAS domains is shown to the right. **Methods:** (A) The λ8-1 genomic clone representing the *sima* gene was identified by low-stringency hybridization of a *Dm* λEMBL3 genomic DNA library with a ^{32}P-labeled 256-bp *Nde*I-*Pst*I *bHLH* fragment of the *Dm sim* gene that encodes the entire bHLH region, and an additional 20 aa of spacer between the bHLH and PAS domain. Hybridization was performed at 37°C in 5 × SSC/10 mM Tris pH 7.5/1 mM EDTA/5 × Denhardt's solution/0.1%SDS/100 μg per ml sheared salmon sperm DNA/10% formamide. Filters were washed in 1 × SSC/0.5%SDS at 60°C. (B) Sequence alignments were generated using the multiple alignment program in GeneWorks (Intelligenetics), the Genetics Computer Group, Madison, WI, USA Bestfit program, and manual alignment. The sequence comparison tree was generated using the GeneWorks Unweighted Pair Group Method with Arithmetic Mean Tree program.

PAS domain is found not only in eukaryotes, but in prokaryotes. The *Bacillus subtilis Kina* gene encodes a PAS domain-bearing protein kinase (Wang et al., 1995; Perego et al., 1989). In this paper we describe another member of the bHLH-PAS protein family, the *Dm* Similar protein.

EXPERIMENTAL AND DISCUSSION

(a) Isolation of a *sim*-related *bHLH* gene

Genes related in sequence to *sim* were sought by screening a *Dm* genomic DNA library at reduced stringency with a *sim bHLH* DNA probe (bHLH proteins generally

Fig. 2. Sequence of *sima* cDNA clone. The complete nt sequence (5763 bp) of the NB-1 *sima* cDNA clone is shown (GenBank accession No. U43090). A poly(A) tract of 12 bp is present at the 3′ end of the sequence and there is a polyadenylation specification sequence (AATAAA) (underlined) that begins 13 bp before the poly(A) stretch. The predicted ORF of 1505 aa begins at Met^{613}, and is shown using the one letter aa code. The Met start codon has flanking sequences [CACA] which match well (3 out of 4) with the preferred *Dm* translation start sequence (MAAM) (where M = A/C) (Brown et al., 1994). Both the bHLH domain (aa 72 through 125) and PAS domain (aa 171 through 433) are boxed. The PAS repeats lie between aa 182 to 223 and 321 to 362. Homopolymeric and acidic aa stretches are underlined. They include 13 stretches of poly[Gln] that follow the bHLH and PAS domains, two stretches of poly[Ser], one His-rich region, and one Pro-rich stretch. The Pro-rich stretch (aa 577–587) is part of a longer Pro-rich region that is 16% Pro over aa 506–635. There are six acidic regions also underlined. **Methods:** Isolation of *sima* cDNA clones utilized a ^{32}P-labeled 600-bp *Pst*I genomic DNA fragment that contains the *sima bHLH* region to screen 500000 pNB40 plasmid cDNA clones derived from *Dm* 4–8 h embryonic mRNA (Brown and Kafatos, 1988). Seven positive clones were identified, purified, analyzed by restriction enzyme mapping and Southern blot analysis, and sequenced.

CGAATCAAAGAAAACGGAAAAATAAAAGCGTATTAAGATTTCGTGTGGAAAGGAAAAATGCTGTAGGCGCAGTGTAAAATGCATATATTGCGGATAAATCGCGTTTGTGAGATGCTTAAACGCGAATTCAACATCAAGTCGGTGCAGAAC 150
TTCAAGGTGCAATAACAAAAATATCCGAATTATCGAAAAAAGCAAAGTGCAATTGGCCAGTTGAGACTACAACCACAATAACAACGACAACAACAGCAACAGCAGTGAAATAATAATATAACCTTGTGTTTTGTTTTTTGTGGCTCAAAA 300
ATCAGATATACACAAAGTTAATCGAGGTGAACGGTGTAAAAGTGATTAAATAATAACGGTTTAACAGGAAAACAGAGTTATAGTGTGTGCGAGTGTGTGTGTGCGTAGTGACAACAATACAAACGCAGAAACCAATACAAGCGCACAATA 450
GTTTTCATACATAGGAAAACAACGAAAGTAACAGCGATAAATAATAGAAAACAACAACGAAGAGTCTCAAAAAGTGCTGGAAAAACTCGCGATAAGGAAAACAGGAACCTCCACATCGGCAGATCTTTTGAATGCTGCCGCACTTTGAAA 600

TCCGGAGACACAATGGTTTCATTAATTGACACAATCGAAGCGGCGGCCGAGAAACAGAAGCAGTCCCAAGCGGTGGTCACCAACACCTCCGCCTCCTCCTCCTCCTGTTCCTCTTCGTTCTCGGCCTCCCCACCATCTTCGTCCGTTGGC 750
M V S L I D T I E A A A E K Q K Q S Q A V V T N T S A S S S S C S S S F S A S P P S S S V G 46

TCCCCATCTCCTGGCGCCCCAAAAACCAATCTCACGGCCAGCGGCAAACCAAAGGAGAAAAGAAGGAACAATGAAAAACGCAAGGAAAAGTCTCGCGATGCGGCTAGATGTCGTCGCTCCAAGGAGACAGAGATATTCATGGAACTCTCG 900
S P S P G A P K T N L T A S G K P K E K R R N N E K R K E K S R D A A R C R R S K E T E I F M E L S 96

GCTGCTTTGCCATTGAAAACCGACGATGTCAACCAACTGGACAAGGCCTCCGTGATGAGGATCACCATCGCTTTCCTCAAGATTCGCGAAATGCTGCAGTTCGTGCCCAGCTTGCGCGACTGCAATGATGACATCAAGCAGGACATTGAA 1050
A A L P L K T D D V N Q L D K A S V M R I T I A F L K I R E M L Q F V P S L R D C N D D I K Q D I E 146

ACTGCAGAGGATCAGCAGGAAGTCAAGCCCAAGCTGGAAGTGGGCACTGAGGATTGGCTCAACGGAGCTGAGGCCAGAGAGCTCCTCAAACAGACCATGGACGGATTCCTGTTGGTCTTGTCCCACGAGGGAGACATCACCTACGTGTCC 1200
T A E D Q Q E V K P K L E V G T E D W L N G A E A R E L L K Q T M D G F L L V L S H E G D I T Y V S 196

GAAAACGTAGTCGAGTATCTGGGCATCACCAAGATTGACACACTTGGCCAACAGATCTGGGAGTACTCGCACCAGTGCGATCACGCCGAGATCAAGGAGGCTCTCAGCCTAAAGCGTGAACTTGCCCAGAAGGTCAAGGATGAGCCGCAG 1350
E N V V E Y L G I T K I D T L G Q Q I W E Y S H Q C D H A E I K E A L S L K R E L A Q K V K D E P Q 246

CAGAACTCTGGAGTGAGCACCCATCACCGGGATCTGTTTGTGCGCTTGAAGTGCACCTTGACCAGTCGTGGACGAAGCATCAACATCAAGAGCGCTTCGTATAAGGTCATTCACATCACGGGACACTTGGTGGTCAATGCAAAGGGCGAA 1500
Q N S G V S T H H R D L F V R L K C T L T S R G R S I N I K S A S Y K V I H I T G H L V V N A K G E 296

CGTCTGCTGATGGCCATTGGCCGACCAATTCCGCATCCTTCGAACATTGAGATTCCGTTGGGAACTAGCACCTTCCTTACCAAACACTCCCTGGATATGCGTTTCACCTATGTGGATGACAAGATGCATGACTTACTGGGATATTTGCCC 1650
R L L M A I G R P I P H P S N I E I P L G T S T F L T K H S L D M R F T Y V D D K M H D L L G Y L P 346

AAGGATCTACTTGACACCTCTTTGTTTTCGTGCCAGCATGGAGCAGATTCCGAACGCCTGATGGCCACCTTCAAGAGCGTGCTGAGCAAGGGACAGGGCGAGACCTCCCGCTACAGATTTCTGGGAAAATACGGTGGCTACTGCTGGATT 1800
K D L L D T S L F S C Q H G A D S E R L M A T F K S V L S K G Q G E T S R Y R F L G K Y G G Y C W I 396

CTTAGCCAGGCCACGATCGTCTACGATAAGTTAAAGCCGCAGAGCGTCGTTTGCGTTAACTATGTCATAAGTAATCTTGAGAATAAACATGAGATATACTCACTTGCTCAACAAACGGCGGCAAGTGAGCAAAAGGAACAACACCACCAA 1950
L S Q A T I V Y D K L K P Q S V V C V N Y V I S N L E N K H E I Y S L A Q Q T A A S E Q K E Q H H Q 446

GCAGCTGAAACCGAGAAGGAGCCCGAAAAAGCAGCCGATCCCGAAATAATCGCTCAGGAAACCAAGGAAACAGTAAACACACCTATTCACACCTCAGAGCTGCAAGCTAAGCCATTGCAACTTGAGAGCGAGAAAGTAGAGAAAACCATC 2100
A A E T E K E P E K A A D P E I I A Q E T K E T V N T P I H T S E L Q A K P L Q L E S E K V E K T I 496

GAGGAAACCAAAACAATTGCAACAATCCCACCAGTAACTGCAACCTCAACGGCTGACCAGATAAAGCAGCTGCCCGAGAGCAATCCGTACAAGCAGATCCTCCAGGCGGAGTTGCTCATCAAGCGGGAGAACCACTCGCCAGGACCGCGC 2250
E E T K T I A T I P P V T A T S T A D Q I K Q L P E S N P Y K Q I L Q A E L L I K R E N H S P G P R 546

ACCATCACTGCCCAGCTGCTCAGCGGCAGTAGCAGCGGACTGCGTCCCGAGGAGAAGCGTCCCAAGTCGGTGACCGCTTCGGTCCTCCGCCCGAGTCCAGCTCCACCCCTAACACCACCGCCCATAGCGGTCCTCTGCAAGAAGACTCCG 2400
T I T A Q L L S G S S S G L R P E E K R P K S V T A S V L R P S P A P P L T P P P I A V L C K K T P 596

CTCGGCGTGGAGCCCAATCTGCCGCCAACCACCACCGCTACGGCAGCCATAATCTCCTCGAGCAATCAGCAGCTGCAGATCGCACAGCAAACCCAACTGCAGAATCCCCAGCAACCAGCTCAAGATATGAGTAAAGGATTCTGTTCGCTA 2550
L G V E P N L P P T T T A T A A I I S S S N Q Q L Q I A Q Q T Q L Q N P Q Q P A Q D M S K G F C S L 646

TTTGCCGACGATGGACGAGGTCTGACAATGCTTAAGGAAGAACCTGACGATTTGTCCCACCACCTGGCCTCCACCAATTGCATTCAGCTGGACGAGATGACTCCCTTCAGTGACATGTTGGTTGGCCTCATGGGCACCTGCCTGCTGCCC 2700
F A D D G R G L T M L K E E P D D L S H H L A S T N C I Q L D E M T P F S D M L V G L M G T C L L P 696

GAGGATATCAATTCCCTGGACTCGACCACCTGCTCCAAGACGGCCAGCGGCCAGCACTACCAGAGCCCCAGCAGCAGCAGCACCTCTGCACCCAGCAATACCAGTAGCAGCAATAATAGCTATGCCAACAGTCCGCTCAGCCCGCTCACA 2850
E D I N S L D S T T C S K T A S G Q H Y Q S P S S S S T S A P S N T S S S N N S Y A N S P L S P L T 746

CCCAATTCCACGGCAACCGCTAGCAATCCCAGTCATCAGCAGCAACAGCAACATCACAACCAGCAGCAGCAGCAGCAGCAGCAGCAACAACAGCAGCAGCATCATCCCCAGCACCACGACAACAGCAACAGCAGCAGCAACATTGATCCG 3000
P N S T A T A S N P S H Q Q Q Q Q H H N Q Q Q Q Q Q Q Q Q Q Q Q H H P Q H H D N S N S S S N I D P 796

CTGTTCAACTATCGCGAGGAGTCGAATGACACAAGCTGCTCGCAGCATCTGCATTCGCCCAGCATCACATCAAAGAGTCCCGAGGACAGCAGCCTCCCATCGTTGTGCAGTCCCAATTCACTAACACAGGAGGATGACTTCTCCTTCGAG 3150
L F N Y R E E S N D T S C S Q H L H S P S I T S K S P E D S S L P S L C S P N S L T Q E D D F S F E 846

GCCTTTGCCATGCGTGCGCCCTACATCCCGATCGATGATGATATGCCCCTGCTAACGGAGACGGATCTCATGTGGTGCCCACCAGAGGATCTGCAGACCATGGTGCCCAAGGAGATCGACGCCATACAGCAGCAGTTGCAACAGTTGCAA 3300
A F A M R A P Y I P I D D D M P L L T E T D L M W C P P E D L Q T M V P K E I D A I Q Q Q L Q Q L Q 896

CAGCAGCAACATCACCAGCAATACGCCAGCAACACTGGCTACCAGCAACAGCAGCAACAGCCGCAACTCCAGCAGCAGCATTTCTCCAACTCCCTGTGCTCCTCGCCAGCTTCAACGGTCTCATCATTGTCGCCATCGCCCGTGCAACAG 3450
Q Q Q H H Q Q Y A S N T G Y Q Q Q Q Q Q P Q L Q Q Q H F S N S L C S S P A S T V S S L S P S P V Q Q 946

CATCATCAGCAGCAACAGGCAGCGGTCTTCACCAGCGATTCCAGCGAGCTGGCGGCATTGCTGTGCGGATCCGGAAATGGAACACTTTCGATCCTGGCGGGTAGTGGTGTTACGGTTACGGAGGAGTGCAACGAGCGGCTGCAGCAACAT 3600
H H Q Q Q Q A A V F T S D S S E L A A L L C G S G N G T L S I L A G S G V T V T E E C N E R L Q Q H 996

CAGCAGCAGCAACAGCAGACGAGTGGCAACGAGTTCAGGACTTTCCAGCAGCTGCAGCAGGAGTTGCAGTTGCAGGAGGAACAGCAGCAGCGTCAGCAGCAGCAACAACAGCAGCAACAGCAGCAGCAGCAACAGCAGTTACTCAGCCTA 3750
Q Q Q Q Q Q T S G N E F R T F Q Q L Q Q E L Q L Q E E Q Q Q R Q Q Q Q Q Q Q Q Q Q Q Q Q Q L L S L 1046

AATATCGAGTGCAAAAAAGAGAAATACGATGTGCAAATGGGAGGCAGTTTATGTCATCCCATGGAGGATGCATTTGAGAATGACTACAGCAAGGATTCCGCCAATCTGGACTGCTGGGACCTGATTCAAATGCAGGTGGTGGACACGGAA 3900
N I E C K K E K Y D V Q M G G S L C H P M E D A F E N D Y S K D S A N L D C W D L I Q M Q V V D T E 1096

CCCGTGAGTCCCAATGCCGCATCTCCCACTCCCTGCAAGGTGTCTGCCATCCAGTTGCTCCAGCAACAGCAGCAACTGCAGCAGCAGCAGCAGCAGAACATTATCTTGAATGCAGTGCCATTGATTACCATACAGAATAACAAGGAATTG 4050
P V S P N A A S P T P C K V S A I Q L L Q Q Q Q Q L Q Q Q Q Q Q N I I L N A V P L I T I Q N N K E L 1146

ATGCAACAGCAGCAGCAGCAACAGCAGGAGCAGCTGCAGCAGCCGGCTATAAAACTGCTGAATGGAGCCAGCATTGCGCCGGTGAACACCAAGGCCACCATCCGCCTGGTGGAGAGCAAGCCGCCCACCACGACGCAGTCTCGAATGGCG 4200
M Q Q Q Q Q Q Q Q E Q L Q Q P A I K L L N G A S I A P V N T K A T I R L V E S K P P T T T Q S R M A 1196

AAGGTGAACCTCGTGCCGCAGCAGCAGCAGCATGGAAACAAGCGCCACCTGAACAGCGCGACAGGAGCGGGCAATCCCGTGGAGTCCAAGCGTCTGAAGAGCGGAACCCTCTGCCTGGACGTGCAGTCCCCTCAGCTCCTGCAGCAGCTA 4350
K V N L V P Q Q Q Q H G N K R H L N S A T G A G N P V E S K R L K S G T L C L D V Q S P Q L L Q Q L 1246

ATTGGCAAGGATCCTGCCCAGCAGCAAACCCAAGCAGCCAAGCGAGCAGGCAGTGAGCGCTGGCAACTGTCGGCGGAGAGCAAACAGCAGAAGCAGCAGCAACAGCAGTCGAACTCGGTACTAAAGAACCTGCTGGTCAGCGGACGCGAT 4500
I G K D P A Q Q Q T Q A A K R A G S E R W Q L S A E S K Q Q K Q Q Q Q S N S V L K N L L V S G R D 1296

GATGACGATTCCGAAGCAATGATAATCGATGAGGACAACTCCCTGGTGCAGCCAATTCCATTGGGCAAATATGGCTTGCCCCTGCACTGTCACACCTCCACATCTTCGGTGCTGCGGGACTATCATAATAACCCGCTGATCAGTGGCACT 4650
D D D S E A M I I D E D N S L V Q P I P L G K Y G L P L H C H T S T S S V L R D Y H N N P L I S G T 1346

AACTTTCAGCTGTCTCCAGTCTTCGGAGGCTCCGATTCCAGTGGAGGTGACGGCGAGACTGGGTCAGTGGTTTCCCTAGATGACTCGGTGCCACCCGGCTTGACAGCCTGCGACACGGATGCCTCCAGCGACAGTGGCATCGATGAGAAC 4800
N F Q L S P V F G G S D S S G G D G E T G S V V S L D D S V P P G L T A C D T D A S S D S G I D E N 1396

AGCCTGATGGATGGGGCATCGGGGTCTCCGCGTAAGAGGTTGTCCTCAACTTCCAATAGCACTAATCAGGCTGAGTCGGCACCACCGGCTTTGGACGTGGAGACTCCAGTGACCCAGAAGTCGGTAGAAGAGGAGCTCGAAGGATGCGGA 4950
S L M D G A S G S P R K R L S S T S N S T N Q A E S A P P A L D V E T P V T Q K S V E E E L E G C G 1446

AGCGGCAATAACGCCCCCAGCAGGAAGACTTCCATTTCCTTCCTGGACAGCAGCAATCCTCTGCTGCACACGCCGGCGATGATGGACCTGGTCAACGACGACTACATCATGGGCGAGGGTGGCTTCGAGTTTAGCGGCAACCAGTTGGAG 5100
S G N N A P S R K T S I S F L D S S N P L L H T P A M M D L V N D D Y I M G E G G F E F S G N Q L E 1496

CAAGTCCTGGGTTGGCCAGAAATTGCCTAGATAACCAGTACCTTAGCATGCACTGCATTTGATATGAGGCATTTGTTGGCGGAGGACGAGAGAAAACATTGAAATTAAACTGAGCATCGAGAAGGAAGCTAACCAAAAGATTGATTTTGA 5250
Q V L G W P E I A * 1505

GCCTTGCGTGAAAGTAGGACCCCGCACAACCAACTTGACAACCTGACAGGACCCATCCCCACCAACCATTTCCCCCTGACGATGATCCAACACTTGACCCCTGACCCCGACCCTGACCCTGGCCCCGACCCCATGCTAACCGACACATTT 5400
ACAAACAGAAAGGGTTAATCACACTTGAATAGTTTTGTAGCCATGTTAACTGCCTAGAACAGAAGAATATACACGTATGTAATTTAGTCTTAATGTTTTTCTATATGCAAACCGTTTACACAACAGATGGATGAGGAGCAAAATAGCAGA 5550
GGAGCGTGAAAAACCAGTCGAAGAATAATGATAAACCTATATATTACAATGAGAAAATACCGATGAAAAATGTTCATAGTTTTTTGCTGGAAGGCTCAAACAAGTTTTTTGGTGATGGTTGTATTCTGGTTTATAACTTTCAAGCAAAAA 5700
TGCTGTGACGAGAAAAAAAGTTTTTGAACATGTTTTACAATAAATATTAACAAAAAAAAAAAA 5763

show highest sequence homology within the bHLH region). One clone, λ8-1, possessed an authentic *bHLH* region. This gene is referred to as *sima*. The Sima bHLH domain has 55% aa identity to Sim (Fig. 1A). It has even higher identity (63%) to HIF-1α, and is closely related to Trh (59%). Within the bHLH region, the highest homology is within the basic region and the loop-helix 2 junction. The presence of identical basic region residues at position numbers shown by the crystal structure of the bHLH-LZ Max protein to be DNA binding residues suggests that all four bHLH-PAS proteins could bind the same core half-site sequence (Fig. 1A; Ferre-D'Amare et al., 1993). The loop is the least conserved region in sequence although its length is the same in all four proteins. The Sima bHLH sequence clearly places Sima within the Sim-related gene family since it shows significantly lower identity to other bHLH proteins: Ahr (28%), Arnt (28%), MyoD (20%) and Myc (15%).

(b) The Sima protein is a member of the bHLH-PAS protein family

The complete coding sequence of *sima* was determined by analyzing corresponding cDNA clones. The inserts of seven embryonic cDNA clones ranged in size from 3 to 5.8 kb, and sequence and restriction map analysis showed they completely overlap. These results do not provide evidence for multiple embryonic mRNA species. This is consistent with Northern blot analysis of embryonic RNA with a ^{32}P-labeled *sima* cDNA probe that detected a single 6.1 kb mRNA species (data not shown). The longest clone, NB-1, was completely sequenced (Fig. 2). It is 5763 nt in length, contains a predicted ORF of 1505 aa (Fig. 2), considerably longer than Sim (673 aa), HIF-1α (826 aa), and Trh (929 aa). In vitro transcription and translation of a full length *sima* cDNA clone resulted in synthesis of a polypeptide of approximately 150 kDa, close to the expected size of Sima (M. Ward and S.T.C., unpublished results).

The Sima coding sequence shares structural features with bHLH-PAS proteins including a bHLH domain, a PAS domain with the 41–44 aa PAS repeats (Nambu et al., 1991), and C-terminal Glu-rich regions. Arrangement and spacing of the domains is also conserved between Sima and other members of the family. Both of the PAS repeats have the invariant PAS repeat residues (S.T.C., unpublished results, Wang et al., 1995) in the 44-aa consensus repeat sequence: Phe^{1}, His^{41}, and Asp^{44}. Several PAS domain residues have been shown by mutational analysis to affect Per PAS domain dimerization (Huang et al., 1993): (*1*) Val^{205} of Per (the *per^{L}* mutation), (*2*) Gly^{375}-Tyr^{376}, and (*3*) Pro^{378}-Leu^{381}. These residues are also conserved in the Sima PAS domain (aa 185, 343–344, 346–349, respectively). The C-terminal half of Sima contains numerous long stretches of poly[Glu], His-rich, Pro-rich, Ser-rich, and acidic regions (Fig. 2) that may function as transcriptional activation domains (Mitchell and Tjian, 1989).

(c) Sequence relationships among bHLH-PAS domain proteins

Fig. 1B shows a sequence identity tree of the published PAS proteins. The proteins can be partitioned into five groups, all with a single homologous member except the Sim-related group, which is comprised of Sim, Sima, HIF-1α, and Trh. Statistical considerations indicate that Sima and HIF-1α are significantly more related to each other than to Sim or Trh. However, analysis of the incomplete human Sim (Chen et al., 1995) and *Dm* Sim sequences indicate they have considerably higher conservation than human HIF-1α and *Dm* Sima. Thus, sequence analysis is unclear regarding whether Sima and HIF-1α are homologous genes.

(d) Ubiquitous embryonic expression of *sima*

Whole-mount embryonic in situ hybridization experiments using different *sima* cDNA probes were performed to determine the *sima* expression pattern. Hybridization was observed in most or all cells from the blastoderm stage through the end of embryogenesis (Fig. 3). The significant amounts of *sima* transcripts in syncitial blastoderm embryos suggest a maternal contribution of *sima*. The pattern of *sima* transcripts completely overlaps that of *sim* and *trh*.

(e) Cytological mapping of the *sima* gene

The cytological position of the *sima* gene was mapped by in situ hybridization to larval polytene chromosomes using biotinylated *sima* cDNA and genomic DNA fragments as probes. Specific hybridization was detected with both probes at position 99D on the right arm of chromosome 3. Thus, unlike the related *bHLH* members of the *achaete-scute* complex and the *Enhancer of split* locus, the *sima* gene is not closely linked to *sim* (87DE) or *trh* (61BC). Further localization of *sima* was derived by in situ hybridization mapping with respect to deficiency and duplication chromosomes (Fig. 4). These experiments localized *sima* to the chromosomal interval 99D6–9.

(f) Conclusions

(*1*) The *Dm sima* gene encodes a novel member of the bHLH-PAS protein family. Sequence data suggests that Sima is a DNA-binding transcriptional activator.

(*2*) Sima is most closely related in sequence to the human HIF-1α protein that controls the cellular response to hypoxia. However, further genetic and cellular analysis

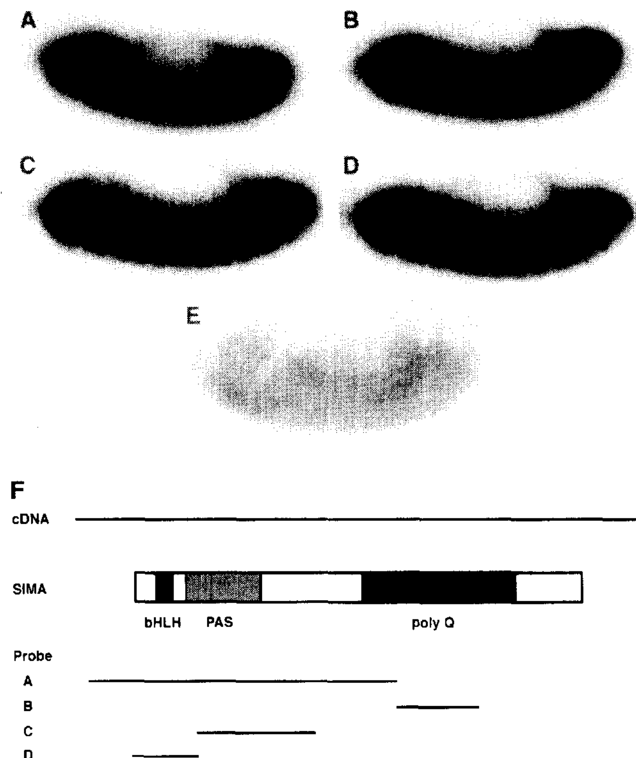

Fig. 3. The *sima* gene is expressed ubiquitously during embryogenesis. *Drosophila* embryos were hybridized in situ to four different *sima* NB-1 cDNA clone fragment probes (to confirm specificity), and all four showed identical uniform expression. The fragments are labeled **A–D** in panel **F**, and their hybridization is shown in panels **A–D**. All embryos are wt stage 13 sagittal views with anterior to the left and ventral up. Probes are: **(A)** 3.4-kb *Eco*RI-*Bam*HI fragment, **(B)** 0.8-kb *Bam*HI fragment, **(C)** 1.2-kb *Bgl*II-*Xho*I fragment, and **(D)** 0.7-kb *Bgl*II fragment. Panel **E** shows a negative control embryo of the same stage hybridized to a non-hybridizing probe. **(F)** Map of the *sima* NB-1 cDNA clone showing the location of the bHLH, PAS, and poly[Glu] (poly Q) domains, and the location of the four probes used for embryo in situ hybridization. **Methods:** Fragments were labeled with digoxygenin using the Genius labeling kit (Boehringer Mannheim). The probe was hybridized to fixed wild-type embryos, and hybridization was detected by alkaline phosphatase histochemistry (Tautz and Pfeifle, 1989).

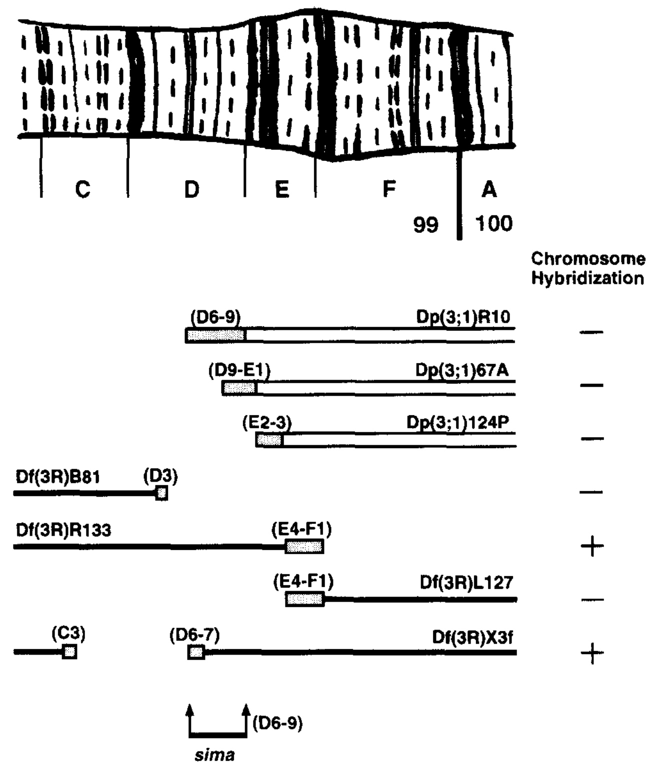

Fig. 4. Cytological mapping of *sima*. In situ hybridization of a *sima* gene probe to polytene chromosomes of wt and deficiency/duplication strains maps the *sima* gene to the third chromosome interval 99D6–9 (bottom of figure). Depiction of polytene chromosome interval from 99C to 100A is shown at the top. Below are shown the names and positions of chromosomal duplications and deficiencies (Kongsuwan et al., 1986; Warmke et al., 1989; Warmke, 1990). Open blocks indicate duplicated (Dp) regions while solid lines represent undeleted material in deficiency (Df) chromosomes. The locations of the breakpoints are indicated by hatched boxes and listed in parentheses. Chromosome hybridization to the *sima* probe is depicted by: (+) detectable hybridization, (−) absence of hybridization. The 99D6–9 interval in which *sima* resides is defined by absence of expression in *Dp(3;1)R10* and the presence of hybridization to *Df(3R)X3f*. **Methods:** Salivary glands from wt and synthetic deficiency *Dm* strains were dissected from third instar larvae in 1 × PBS, fixed in 45% acetic acid, transferred to siliconized cover slips and spread on glass slides. The fixed tissues were hybridized to biotinylated genomic fragments and hybridization detected using horseradish peroxidase-linked avidin (ENZO Biochemicals) followed by a diaminobenzidine/hydrogen peroxide reaction. PBS is 0.125 M NaCl/15.5 mM Na_2HPO_4/8.5 mM NaH_2PO_4, pH 7.0.

is required to determine if hypoxia regulation is one function of *sima*.

(*3*) *sima* transcripts are ubiquitously expressed in the embryo.

(*4*) *sima* maps to *Dm* cytological position 99D, a location unlinked to the other known *bHLH-PAS* genes.

ACKNOWLEDGEMENTS

We thank Jim Adams for contributions in the early stages of this work, Jeff Warmke for helpful advice and fly strains, Gregg Semenza for informative discussions, Benny Shilo for communicating the Trh sequence, Nick Brown for the cDNA library, Jay Nystrom for help with sequence analysis, Mary Ward for communicating Sima protein results, and Margaret Sonnenfeld for critical reading of the manuscript. J.R.N. was supported by an NIH postdoctoral fellowship. This work was supported by grants from the Lucille P. Markey Charitable Trust, Jonsson Cancer Center Foundation (J911209), and the NIH (HD25251) to S.T.C.

REFERENCES

Brown, C.M., Stockwell, P.A., Dalphin, M.E. and Tate, W.P.: The translational termination signal database (TransTerm) now also includes initiation contexts. Nucleic Acids Res. 22 (1994) 3620–3624.

Brown, N.M. and Kafatos, F.C.: Functional cDNA libraries from *Drosophila* embryos. J. Mol. Biol. 203 (1988) 425–437.

Chen, H., Chrast, R., Rossier, C., Gos, A., Antonarakis, S.E., Kudoh, J., Yamaki, A., Shindoh, N., Maeda, H., Minoshima, S. and Shimizu, N.: *Single-minded* and Down syndrome. Nature Genetics 10 (1995) 9–10.

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

Huang, Z.J., Edery, I. and Rosbash, M.: PAS is a dimerization domain common to *Drosophila period* and several transcription factors. Nature 364 (1993) 259–262.

Isaac, D.D. and Andrew, D.J.: Tubulogenesis in *Drosophila*: a requirement for the *trachealess* gene product. Genes Devel. 10 (1996) 103–117.

Kongsuwan, K., Dellavalle, R.P. and Merriam, J.R.: Deficiency analysis of the tip of chromosome *3R* in *Drosophila melanogaster*. Genetics 112 (1986) 539–550.

Mitchell, P.M. and Tjian, R.: Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science 245 (1989) 371–378.

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

Perego, M., Cole, S.P., Burbulys, D., Trach, K. and Hoch, J.A.: Characterization of the gene for a protein kinase which phosphorylates the sporulation-regulatory proteins SpoOA and SpoOF of *Bacillus subtilis*. J. Bacteriol. 171 (1989) 6187–6196.

Rosbash, M.: Molecular control of circadian rhythms. Curr. Opin. Genet. Dev. 5 (1995) 662–668.

Swanson, H.I. and Bradfield, C.A.: The AH-receptor: genetics, structure and function. Pharmacogenetics 3 (1993) 213–230.

Tautz, D. and Pfeifle, C.: A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene *hunchback*. Chromosoma 98 (1989) 81–85.

Wang, G.L., Jaing, B.-H., Rue, E. and Semenza, G.L.: Hypoxia-inducible factor 1 is a basic helix-loop-helix-PAS heterodimer regulated by cellular O_2 tension. Proc. Natl. Acad. Sci. USA 92 (1995) 5510–5514.

Warmke, J.W.: Genetic Analysis of Myosin Light Chain-2 Function in *Drosophila melanogaster*. Ph.D. thesis, The Ohio State University, Columbus, OH, 1990.

Warmke, J.W., Kreuz, A.J. and Falkenthal, S.: Co-localization to chromosome bands 99E1–3 of the *Drosophila melanogaster* myosin light chain-2 gene and a haplo-insufficient locus that affects flight behavior. Genetics 122 (1989) 139–151.

Wilk, R., Weizman, I., Glazer, L. and Shilo, B.-Z.: *trachealess* encodes a bHLH-PAS protein and is a master regulator gene in the *Drosophila* tracheal system. Genes Devel. 10 (1996) 93–102.