BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **229,** 494–497 (1996)
ARTICLE NO. 1832

The Exon–Intron Structure of Human *LHX*1 Gene

Fabio Bozzi,* Stefano Bertuzzi,† Dario Strina,* Cettina Giannetto,*
Paolo Vezzoni,* and Anna Villa*

**Istituto Tecnologie Biomediche Avanzate, Consiglio Nazionale delle Ricerche, via Ampère 56, 20131, Milan, Italy; and †Laboratory of Mammalian Genes and Development, National Institutes of Health and Human Development, NIH, Bethesda, Maryland 20892*

Received October 21, 1996

We have determined the genomic structure of the human *LHX*1 gene, a member of the LIM/homeobox (*Lhx*) gene family. The transcript is assembled from five exons, which are separated by introns ranging in size from 93 nt to 2.3 kb. The two LIM domains are entirely contained in the first and second exons, respectively, while the homeodomain is split into exons three and four. This structure closely parallels the organization of other mouse and human Lhx genes whose genomic structure is known. An exception is the mouse and human *isl*1 genes, whose homeodomain does not contain introns. An intron at the same position also occurs in the *Xlim*1 gene as well as in other homeobox genes, such as *evx*1 and *evx*2, suggesting that this intron insertion represents an ancestral event, from which homeobox genes of different families originated. In this context, evolution of the *Lhx* gene family probably involved the shuffling of this intron-containing homeobox in the proximity of a LIM-only gene, while *Islet* genes were formed either by the shuffling of an intronless homeobox to the same LIM domain or, alternatively, by intron loss during their evolution.

LIM genes constitute a large gene family whose gene products carry a unique cysteine-rich zinc-binding domain called the LIM domain. This domain, CxxCx(17-19)HxxCxxCxxCx(16-20)Cxx(D/H/C)x, was first identified in the proteins coded by three different genes: *mec-3* (11) and *lin-11* (3) in C. elegans and *Isl-1* from rat (4). At least 40 members of this family, distributed in four groups according to the number of LIM domains and to the presence of homeodomains, kinase domains and others, have been identified in vertebrates and invertebrates (2, 7, 9).

One of the four subfamilies of LIM proteins is characterized by the presence of a homeodomain located 3′ to the two LIM domains. At least nine members of this group have been identified so far, and a new nomenclature, which we will refer to in the present paper, has been suggested (2). The evolutionary origin of these genes has not yet been thoroughly investigated, but probably involved the shuffling of a homeobox to an ancestral LIM containing gene, followed by several duplication events and subsequent nucleotide divergence. An evolutionary dendrogram based on cDNA sequences has recently been proposed (2). In this regard, comparison of the genomic structure of these genes could provide additional information. For instance it is well known that, although the length of introns is subjected to high variation during evolution, exon-intron boundaries are often conserved thus revealing a relationship between related genes. However, very little data exist on the genomic structure of these genes. We report here the exon-intron structure of human *LHX*1 gene and compare it to that of the few known LIM/homeobox genes.

RESULTS AND DISCUSSION

In order to obtain genomic clones of human Lhx subfamily, we screened a cosmid library with a mouse *Lhx*5 cDNA probe (Bertuzzi et al, unpublished), essentially as previously de-

0006-291X/96 $18.00

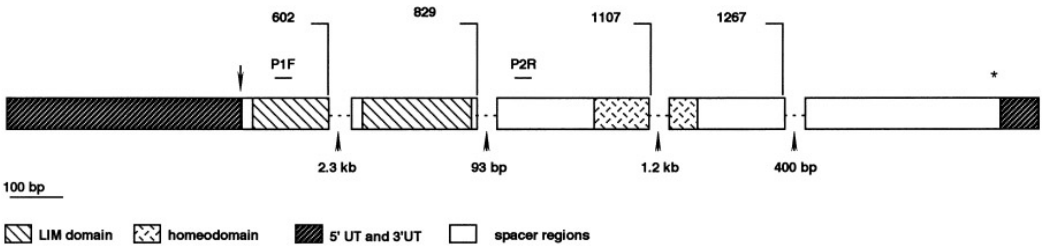

FIG. 1. Exon-intron structure of *LHX1*. All the boxes represent exons, while the dashed line indicates introns. The differently patterned boxes show various segments of the transcript in relation to the functional domains encoded by the exons. Numbers at the end of each box indicate the last nucleotide of the exon. The arrow indicates the ATG codon, the asterisk indicates the stop codon, and the numbers with arrowheads refer to the approximate intron length. The primer pair P1F and P2F are indicated.

scribed (10). A cosmid clone was identified and subcloned. Hybridization-positive bands were partially sequenced, and comparison to data bases showed identity with the human *LHX*1 gene, whose unpublished sequence is deposited in data bases (a.n. U14755).

The exon-intron structure (Fig 1) was determined by partial sequencing of cosmid subclones or PCR products and comparison to the *LHX*1 cDNA sequence with a FastA algorithm to define exon-intron boundaries. The *LHX*1 sequence is interrupted by four introns, at positions 602, 829, 1107 and 1267. All the splice sites conform to the GT/AG rule: they are shown in Table 1, together with 50 nt of flanking sequence. Exon sizes are 602, 227, 278, 160 and 446 respectively. Intron sizes range from the 93 nt of intron 2 to approximately 2.3 kb of the estimated size of intron 1. The whole *LHX*1 gene is contained in a relatively short DNA fragment of approximately 6 kb.

The 5′ and 3′ untranslated regions are completely contained in the first and last exon, respectively. Comparison of the exon-intron structure to the protein domains is shown in Fig 1. Each separate LIM domain is entirely contained in the first and second exon, respectively. The ''linker'' region between the second LIM domain and the homeodomain is contained in the third exon, while the homeodomain is split into exons three and four by an intron located exactly after codon 47 of the 61-long homeobox.

At least five distinct LIM/homeobox genes (*Lhx*1 to *Lhx*5), in addition to *Isl*1 and *Isl*2 genes have been identified in mouse (2, 12 and Bertuzzi et al: unpublished). The genomic structure

TABLE 1
Exon–Intron Junctions of the *LHX 1*

EXON	Acceptor site	Donor site
1		....ACTTCTTCCG gtgagtactttcctcccacgcctctgctgct acctccccgcgggcccttc
2	ccctggctgacccatccccgcccccgccccccaccc ccaccccccgcag GTGTTTCGGT	....CTTCACTCGG gtgaggccccaattcctggctggctaggtgc aagcgggtcctgggg
3	gaggaaggctcgccaaggccccggctcatctgtcc tttccctcttag CCACCACGGG	....CGTCATTCAG gtcaggccccggcgcgcctctccatcccaca gaggcccacactgccactt
4	gggtcgggggtggagtctcggtggcctcaccccgcc gccatgtgctgcag GTCTGGTTCC	....TTCTACGGAG gtgggtgcgcgccgaatggcggggcacggcc aggtcggggcgggcttcgt
5	cggcccccggccgggccctgacgtcctgcgccctcc ccgccgctccgcag ATTACCAGAG	....

Note. Computer analysis of the sequence was performed with the GCG software package running on a Sun workstation. Lowercase letters indicate the partial intron sequences (Accession Nos. X96914 to 96921) while the uppercase letters indicate the partial exon sequences (Accession No. U14755). The sequence of the second intron (93 bp) is complete.

of the five Lhx genes and of the mouse and human Islet-1 genes has been reported: *Lhx*3: (12); *Lhx*4: (8); *Lhx*5: (Bertuzzi et al. submitted); mouse *Isl*1: (5) and human *Isl1* (6) or is available through personal communication (*Lhx*1: Fuji et al. personal communication; *Lhx*2: Xu et al. personal communication). The general structure of *LHX1* is identical to the one reported for all the five mouse Lhx genes, with each LIM domain contained in single exons, and the homeodomain split into two exons at the same position. Interestingly, the homeodomain of mouse and human Islet-1 is not interrupted by introns (5, 6), while the C. elegans *mec-3* gene is interrupted at a different position, between amino acids 12 and 13 of the homeodomain (11). However, an intron at the same position is found in the *Xlim*1 gene (Dawid and Rebberts, personal communication) as well as in other homeobox-containing genes such as mouse *evx*1 and *evx*2 (1).

This suggests that intron insertion in an ancestral homeobox was an ancient event predating the divergence of amphibians from mammals. After the shuffling of this intron-containing homeobox to a LIM-containing protogene, several duplication events occurred, giving rise to all the members of the *Lhx* family. In this context, intron absence in *Isl* genes is somewhat puzzling, but could be explained either by a separate shuffling or recombination event to the same LIM-only gene involving a different intronless homeobox, or could represent an example of intron loss during evolution.

Finally, the *LHX*1 gene has been mapped to chromosome 11p12 (unpublished data deposited in data bases with accession no.: U14755). The definition of the genomic structure and of the exon-flanking sequences will allow the amplification of the whole gene. In this way, the possibility that the gene is involved in diseases characterised by abnormalities in the development of the nervous system mapping to this region can be tested by its direct sequencing in patients.

MATERIALS AND METHODS

Isolation of the genomic clone and subcloning of pertinent fragments for DNA analysis. The genomic clone corresponding to human *LHX1* was obtained from a commercial cosmid library (Stratagene), made from human genomic DNA in pWE15 vector, as previously described (10). The probe was a mouse *Lhx5* cDNA probe covering the two LIM domains and the homeodomain (Bertuzzi et al. unpublished). Cosmid subcloning was performed in pBluescript II ks+ vector (Stratagene) and pertinent fragments were sequenced with the Sanger dideoxy chain termination method using Sequenase (USB). Computer analysis of nucleotide and amino acid sequences was performed with the GCG software package on a VAX 3600 computer. Homology searches were performed on GenBank, Embl and PIR. Exon-intron boundaries were determined by comparison of the genomic and cDNA sequences. The exon flanking sequences have been deposited in Genbank, accession numbers from X96914 to X96921.

DNA extraction, Southern blot and PCR amplification. Human genomic and cosmid DNA was purified with the standard phenol/chloroform procedure and used for Southern blot and PCR amplification. Introns 1 and 2 were obtained by PCR amplification from cosmid DNA with the following primers: P1F (sense: ctggcacgtcaagtgcgtcc) and P2R (antisense: tccgacacgttggcgctct).

PCR amplification was performed with the following reaction components: 5-10 ng of purified cosmid DNA template, 10m*M* Tris-HCL (pH 8.3), 50m*M* KCL, 0.001% gelatin, 2m*M* MGCL2, 200μM dNTP, 20 pmol of primers, 2.5 U of *Taq* polymerase and water to 50μl. A Perkin-Elmer thermal cycler was used for 30 cycles of reactions under conditions of denaturing at 94°C for 1 min annealing at 62°C for 1 min, and extension at 72°C for 2 min. PCR amplification products were cloned in TA vector (Invitrogen). DNA was prepared by Wizard

Miniprep System (Promega) and sequenced by the dideoxynucleotide chain termination method using the Sequenase kit (USB).

For Southern analysis, 10μg of genomic DNA or 50 ng of cosmid DNA was digested with EcoRI enzyme, according to manufacturer's instructions (Promega). The DNAs were separated on a 1.0% agarose gel and transferred to Hybond N+ (Amersham). Probes were labelled to a specific activity of $2 \times 10/9$ cpm/μg by random oligonucleotide priming. Hybridization was carried out at 65°C in $5 \times$ SSPE, $5 \times$ Denhardt's solution, 0.1% SDS and denaturated salmon sperm DNA (100μg/ml). The filters were washed at a final stringency of $0.1 \times$ SSPE, 0.1% SDS at 65°C for 30 minutes followed by autoradiography for 18 hours at −80°C using Kodak XAR film with an intensifying screen.

ACKNOWLEDGMENTS

We are grateful to Professor R. Dulbecco and Dr Heiner Westphal for their suggestions and to Professor L. Rossi Bernardi for his encouragement. The technical assistance of Lucia Susani and Valeria Fasolo is gratefully acknowledged. We also thank Sophie Bevan for her careful typing of the manuscript. This article is paper No. 4 of the Genome 2000/ITBA project funded by CARIPLO.

REFERENCES

1. Bastian, H., and Gruss, P. (1990) *EMBO J.* **9,** 1839–1852.
2. Dawid, I. B., Toyama, R., and Taira, M. (1995) *C.R. Acad Sci. Paris* **318,** 295–306.
3. Freyd, G., Kim, S. K., and Horvitz, H. R. (1990) *Nature* **344,** 876–879.
4. Karlsson, O., Thor, S., Norberg, T., Ohlsson, H., and Edlund, T. (1990) *Nature* **344,** 879–882.
5. Pfaff, S., Mendelsohn, L. M., Stewart, L., Edlund, T., and Jessell, T. M. (1996) *Cell* **84,** 309–320.
6. Riggs, A., Tanizawa, C. Y., Aoki, M., Wasson, J., Ferrer, J., Rabin, D. U., Vaxillaire, M., Froguel, P., and Permutt, M. A. (1995) *Diabetes* **44,** 689–694.
7. Sanchez-Garcia, I., and Rabbitts, T. H. (1994) *Trends Genet.* **10,** 315–320.
8. Singh, G., Kaur, S., Stock, J. L., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., and Potter, S. S. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 10706–10710.
9. Taira, M., Evrard, J. L., Steinmetz, A., and Dawid, I. D. (1995) *Trends Genet.* **11,** 431–432.
10. Villa, A., Patrosso, C., Biunno, I., Frattini, A., Repetto, M. M., Mostardini, M., Evans, G., Susani, L., Strina, D., Redolfi, E., Lazzari, B., Pellegrini, M., and Vezzoni, P. (1992) *Genomics* **13,** 1231–1236.
11. Way, J. C., and Chalfie, M. (1988) *Mec*-3, *Cell* **54,** 5–16.
12. Zhadanov, A. B., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., and Westphal, H. (1995) *Genomics* **27,** 27–32.