Structure of the human CpG-island-containing lung Kruppel-like factor (LKLF) gene and its location in chromosome 19p13.11–13 locus

Sergey V. Kozyrev[a], Lise Lotte Hansen[b], Andrew B. Poltaraus[a], Dmitry A. Domninsky[a], Lev L. Kisselev[a,*]

[a] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 32, Vavilov str., Moscow 117984, Russia*
[b] *Institute of Human Genetics, The Bartholin Building, Aarhus University, DK-8000 Aarhus C, Denmark*

Received 5 February 1999

Abstract We have cloned earlier a short human genomic fragment which showed strong similarity with the mouse cDNA encoding lung Kruppel-like zinc finger transcription factor (LKLF), predominantly expressed in mouse developing lung, spleen, and vascular system, which might play a key role in programming the quiescent state of single positive T cells and blood vessel wall morphogenesis. Here we report the successful cloning of the human LKLF cDNA, its genomic structure and chromosomal localization at the 19p13.11–p13.13 locus. The full-length human LKLF cDNA has longer 5′-UTR with higher GC content than mouse cDNA and encodes a predicted protein of 355 amino acids which has three zinc fingers at the C-terminus and a proline-rich N-terminal domain. Human and mouse proteins share 87.3% identity and 90.2% amino acid similarity. The human LKLF gene consists of three exons. From the proximal promoter to the end of the second exon, we have found a CpG island with an average 76% GC content and two regions of unusually high GC density.

Key words: Lung Kruppel-like factor gene; Kruppel-like protein family; CpG island; *Not*I jumping clone; Human chromosome 19

1. Introduction

Eukaryotic transcription factors containing Cys_2-His_2-type (Kruppel-like) zinc fingers are involved in various processes of cell growth and differentiation and many of these factors exhibit highly restricted patterns of expression in tissues and during development [1–3]. These factors share a common principle of structural organization. While zinc finger domains have a conservative structure, the activation domains differ significantly and often contain regions rich in proline, glutamine, and dicarbonic amino acids [4,5]. Recently, some new proteins of this family have been identified: EKLF [6], BKLF [7], LKLF [8] and GKLF [9]. It was demonstrated that mouse lung Kruppel-like factor (LKLF) is expressed in the lung, the developing vascular system, heart, skeletal muscles, kidney, testis and in lymphoid tissues [1,2]. After examination of different cells, LKLF expression was revealed in mature SP thymocytes and splenocytes, B220+/IgM+ splenic B cells and in bone marrow macrophages [1]. Interestingly, LKLF was expressed only in mature resting T cells down-regulated immediately after their activation [1,10]. LKLF-deficient T cells produced by gene targeting were very susceptible to Fas-mediated apoptosis because of the rapidly increasing level of Fas ligand expression [1]. The Fas ligand belonging to the tumor necrosis factor family is known to be expressed in activated T cells and induces apoptosis when it binds to its receptor Fas. Thus it may be suggested that LKLF might play a key role in establishing the quiescent state of T cells by preventing spontaneous activation and subsequent apoptosis of these immune cells [1]. LKLF also appeared to take place in regulating blood vessel wall morphogenesis as it was found that LKLF is expressed in vascular endothelial cells throughout mouse embryogenesis [2,11].

Earlier, during the total human genome *Not*I mapping project we cloned a short *Not*I jumping clone, which is part of the CpG island and displays similarity with mouse LKLF cDNA [12]. In this study we report the cloning of the human LKLF cDNA, its gene structure and chromosomal mapping.

2. Materials and methods

2.1. Cloning of human LKLF cDNA

5′-RACE PCR and 3′-RACE PCR were applied to isolate cDNA of the human LKLF gene. The plasmid clone jl003 containing a 111-bp *Not*I human genomic fragment obtained as described [12] was used for the design of two sets of the gene-specific PCR primers: RN-1 (5′-GGCCTCCAGCAGCTCCAGCGGGG-3′), and nested RN-2 (5′-GCAGGCGGCGTGAGGAGACC-3′), and DN-1 (5′-CGCGGTCTCCTCACGCCG-3′), and nested primer DN-2 (5′-CCCCGCTGGAGCTGCTGGAGGCC-3′). Ap1 and Ap2 primers were provided in the Marathon-Ready cDNA kit purchased from Clontech Laboratories. The human lung adaptor-ligated Marathon-Ready cDNA was used as a template for PCR: 30 cycles of denaturation at 95°C for 20 s, annealing at 64°C for 20 s, polymerization at 72°C for 1.5 min in the buffer containing 50 mM Tris-HCl, pH 9.2, 16 mM ammonium sulfate, 1.75 mM $MgCl_2$, 0.2 mM of each dNTP, 0.5 U of Taq polymerase, 5% DMSO and 1 M betaine for 5′-RACE and for 3′-RACE in the same buffer with 5% DMSO only. After the first round of PCR nested PCR was carried out in the same conditions.

2.2. Genomic structure analysis

Exon-intron boundaries were determined by using PCR with K-1 (5′-CCCATCCTGCCGTCCTTTTCCACTTTCG-3′) and RN-3 (5′-CCGGGACCTCGCATGCA-3′) primers for the first intron and DN-2 and RN (5′-CTGTGCCCGGCCCGGTTCTCT-3′) primers for the second intron. To clone the 5′-upstream region we carried out inverse PCR as described [13]. Briefly, genomic DNA was completely digested with *Bam*HI restriction endonuclease and self-ligated at 1.5 μg/ml of DNA concentration. Primary PCR was performed with 4.5 ng of self-ligated genomic DNA and RN-3 and DN-2 primers at the conditions described above for cDNA cloning. Then, 1/50 part

*Corresponding author. Fax: (7) (095) 135 1405.
E-mail: kissel@imb.imb.ac.ru

Abbreviations: SP, single positive; UTR, untranslated region; CGI, CpG island; RACE, rapid amplification of cDNA ends; PCR, polymerase chain reaction; LKLF, lung Kruppel-like factor; EKLF, erythroid Kruppel-like factor; BKLF, CACCC-box-binding Kruppel-like factor; GKLF, gut-enriched Kruppel-like factor

PII: S0014-5793(99)00348-8

of the diluted primary PCR product was used as a template for secondary PCR with RN-5 (5′-CGACAGGGCAATGGGCG-3′) and I-2d (5′-CCTTGGGGCGTGGCTCA-3′) primers.

2.3. Sequence analysis

All PCR products were purified from 1% low-melting-point agarose gel, cloned into pUC 19 via blunt ends and sequenced on an automated DNA sequencer (Applied Biosystems) with ABI Prism Dye Terminator Cycle Sequencing Ready Reaction kit from Perkin-Elmer. Computational analysis and nucleotide sequence homology searches were performed using the BLAST program at http://www.ncbi.nlm.nih.gov/BLAST/, TFSEARCH program at http://www.genome.ad.jp/SIT/TFSEARCH.html and TESS program at http://agave.humgen.upenn.edu/utess/tess?request=SBS-FRMREQ-BasicQuery and DNASIS software from Hitachi Software Engineering. Protein alignment was done with ClustalX program. The GenBank accession number for the human LKLF gene is AF123344.

2.4. Chromosomal mapping of the human LKLF gene

A Genebridge 4 radiation hybrid panel obtained from Research Genetics (Huntsville, AL, USA) consisting of 93 radiation hybrid clones, one human cell clone and a hamster recipient cell clone was used. Three independent PCR amplifications were performed on the panel samples and controls, and the products were analysed in a 2% agarose gel. A 313-bp fragment was obtained by PCR amplification with the primers M-1 (5′-TGACAACAGTGGGGAGTGG-3′) and M-2 (5′-GTGTGCTTTCGGTAGTGG-3′). The PCR amplification conditions were: 20 pmol of each primer, 1×PCR buffer (supplied with the enzyme, Boehringer Mannheim), 6% DMSO, 1 M betaine, 25 ng of DNA and H_2O to a final volume of 8.9 μl, and incubated at 99°C for 10 min. The samples were immediately placed on ice and 0.25 mM dNTP and 0.5 U of Taq polymerase (Boehringer Mannheim) were added. 37 cycles were performed consisting of 94°C 2 min, 67°C 30 s and 72°C 1 min, 1 cycle; and 94°C 45 s, 67°C 30 s and 72°C 1 min for 35 cycles followed by extension for 10 min. The PCR results were sent to the Whitehead Institute/MIT Center for Genome Research at the http://www.genome.wi.mit.edu for establishing the localization of the human LKLF gene and the found location data were compared with the Marshmed genetic map (http://www.marshmed.org/), Genetic Location Database (LDB) (http://cedar.genetics.soton.ac.uk/public_html/) and Genome Database (GDB) (http://www.gdb.org).

3. Results and discussion

3.1. Cloning and sequencing of the human LKLF cDNA

To clone human LKLF cDNA, we used information from the previously characterized *Not*I jumping clone jl003, which we obtained during the total *Not*I human genome mapping project [12,14]. The size of the DNA insert in this clone was 111 bp, of which 92 bp were G/C. To overcome the difficulties caused by high GC content during PCR we used 1 M betaine and 5% DMSO known to be efficient in melting GC-rich regions [15,16]. Two rounds of 5′- and 3′-RACE PCR on the human lung Marathon-Ready cDNA template with the gene-specific primers based on the sequence from the jl003 clone and the adaptor primer Ap1 and the nested adaptor primer Ap2 resulted in the isolation of LKLF cDNA fragments. Analysis of the products by 1% agarose gel electrophoresis revealed that the 5′-RACE PCR product, unlike the 3′-RACE PCR product, was non-homogeneous even after nested PCR (Fig. 1A). We have found that about half of the cloned and sequenced 5′-RACE PCR fragments contained the entire sequence of the first intron at the 5′ end instead of the first exon sequence (Fig. 1B), but we could not uncover the presence of the second intron sequence in the cDNA template. Marathon-Ready cDNA appears to contain partially spliced cDNA sequences.

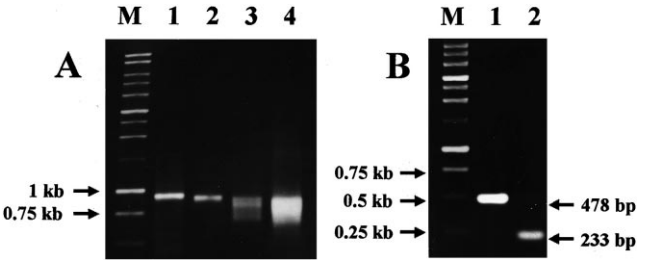

Fig. 1. Human LKLF cDNA PCR amplification. A: 5′- and 3′-RACE PCR fragments of human LKLF cDNA. 3′-RACE PCR products amplified with primer sets DN-1/Ap1 for the primary PCR (lane 1) and Dn-2/Ap2 for the secondary PCR (lane 2). 5′-RACE PCR products using primer pair RN-1/Ap1 (lane 3) and nested primer pair Rn-2/Ap2 (lane 4). B: Amplification products generated from the Marathon-Ready cDNA using exon-specific primers K-1 and Rn-3 (lane 1) and intron-specific primer Rn-5 and adaptor primer Ap2 (lane 2). Lane M (A and B): DNA size marker.

The length of the longest cDNA obtained after 5′-RACE and 3′-RACE PCR was 1671 bp. Many human expressed sequence tags (EST) were found after a search in the GenBank (release of 16 September 1998), which represent different short sequences located between the second *Not*I site and the 3′-UTR. The 5′-UTR is longer and has higher GC content than in mouse cDNA (GenBank U25096).

3.2. The deduced amino acid sequence of human LKLF

The entire human LKLF cDNA encodes a 355-amino acid protein with a calculated molecular mass of 37.4 kDa and an

```
human  MALSEPILPSFSTFASPCRERGLQERWPRAEPESGGTDDDLNSVLDFILSMGLDGLGAEA   60
mouse  ..........A......-.........N...A....E...N.................N

human  APEPPPPPPPPAFYYPEPGAPPPYSAPAGGLVSELLRPELDAPPGPALHGRFLLAPPGRL  120
mouse  P.....Q..................I..DS.GT.....D..P.Q.................

human  VKAEPPEADGGGGYGCAPGLTRGPRGLKREGAPGPAASCMRGPGGRPPPPPDTPPLSPDG  180
mouse  .......V....-........AH......L.....ATGA.....A...............

human  PARLPAPGPRASFPPPFG-GPGFGAPGPGLHYAPPAPPAFGLFDDAAAAAAALGLAPPAA  239
mouse  .L.I..S...NP......P..S..G...A...G....G.....E..............T

human  RGLLTPPASPLELLEAKPKRGRRSWPRKRTATHTCSYAGCGKTYTKSSHLKAHLRTHTGE  299
mouse  .......S.....................A........TN.....................

human  KPYHCNWDGCGWKFARSDELTRHYRKHTGHRPFQCHLCDRAFSRSDHLALHMKRHM*     355
mouse  .......E.............................................*
```

Fig. 2. Alignment of the human and mouse LKLF protein sequences. Amino acids are shown in standard one-letter code. Dots represent identical amino acids, dashes indicate gaps. The three zinc fingers are boxed. Proline residues, alanine stretch and nuclear localization signal are in bold letters.

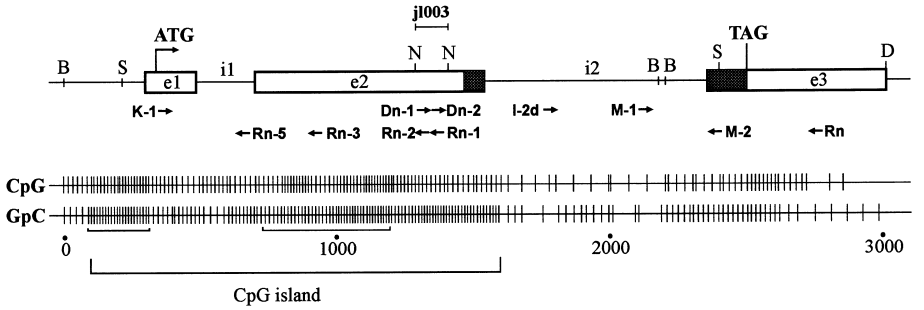

Fig. 3. Genomic organization of the human LKLF gene. Exons (e1, e2, e3) are shown as boxes separated by intronic sequences (i1, i2). Zinc finger region is shaded. The letters B, N, S and D indicate the restriction sites for *Bam*HI, *Not*I, *Sac*I and *Dra*I, respectively. The jl003 *Not*I clone and PCR primers are shown above and below the gene structure. A map plot of CpG and GpC distribution through the gene is represented by the short vertical lines. CpG island and two regions of encreased CpG/GpC density are marked by horizontal brackets.

estimated p*I* of 8.8. The protein is rich in proline residues which constitute 17.5% of the total amino acids and form Pro-rich repeats of three to eight amino acids in a row. An additional feature of the human LKLF is an alanine stretch at the border between the putative activation domain and the zinc finger domain (Fig. 2). We also revealed a 16-amino acid stretch adjacent to the first zinc finger sequence KPKRGRRSWPRKRTAT which should be considered a potential nuclear localization signal [17]. This sequence is identical in human LKLF and mouse gut-enriched Kruppel-like factor (GKLF) [9]. The three zinc fingers are similar to the zinc finger consensus sequence Cys-X_{2-4}-Cys-X_{12}-His-X_{3-4}-His [18] while only the first inter-finger spacer, the so-called H/C link, consisting of seven amino acids, has the conservative sequence TGEKPYH. Thus, the human protein has a canonical Kruppel-like transcription factor structure and is divided into two domains, the Pro-rich transactivating N-terminal domain and the C-terminal zinc finger containing domain. Human and mouse proteins share 87.3% identity and 90.2% amino acid similarity (Fig. 2).

3.3. Gene structure analysis

The genomic organization was determined by applying standard PCR for identifying exon-intron boundaries and by inverse PCR for cloning the 5′-flanking region. The human LKLF gene spans over 3 kb and is composed of three exons having sizes of 174 bp, 817 bp and 680 bp, interrupted by two introns (Fig. 3). Both introns show classical GT/AG sequences flanking the splice junctions [19] and are 217 bp and 819 bp in size. The splice site for the first intron occurs between the 25th and 26th amino acid codons. The second intron disrupts the glycine codon GGT located within the first H/C link.

Although the 5′-flanking region of the gene manifests typical features of GC-rich promoter with multiple transcription factor Sp1 binding sites, it nevertheless contains two putative TATA box consensus motifs as revealed by database analysis using the TFSEARCH and TESS programs. It is unclear yet whether these elements are functionally active.

As one can see from the plot of CpG and GpC distribution (Fig. 3), the human LKLF gene contains a CpG island where 76% of the nucleotides are either C or G. This CGI is about 1.6 kb and spans from the promoter region to the end of the second exon. Interestingly, inside the CGI body regions of the proximal promoter and the central part of the second exon have as much as 80% and 82% GC, respectively. The presence of these two regions separated by the intron with lower CpG/GpC density could be the consequence of a mutational process which does not affect the functionally essential regions. Promoter regions in the mammalian genes are often protected from the action of DNA methyltransferase by DNA binding transcription factors and contain unmethylated CpG dinucleotides. Many of other CpGs scattered throughout the genome are methylated during the cell cycle and are prone to mutate to TpG or CpA [20]. The sequence of the second GC-rich region encodes CG-containing proline codons in the transactivation domain and therefore could also be protected from the mutational changes. As was shown earlier [21], many of the zinc finger proteins encoding genes encompass CpG islands. Moreover, part of these genes involved in cell growth and differentiation have clustered *Not*I sites, which can serve as CGI markers due to their recognition site GCGGCCGC and thus facilitate gene cloning and mapping [12].

3.4. Chromosomal localization of the human LKLF gene

Chromosomal localization of the human LKLF gene was ascertained by PCR amplification using a Genebridge 4 radiation somatic cell hybrid panel. Results obtained during PCR were sent to the WICGR mapping service which placed the human LKLF gene 10.9 cR from the distal marker WI-6344/D19S825 on 19p13.13 with a LOD score of 19 (Fig. 4). Ac-

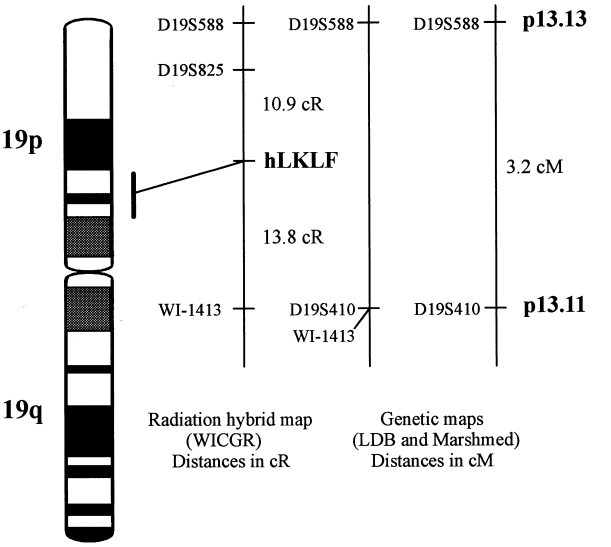

Fig. 4. Chromosomal mapping of the human LKLF gene. The results from three different maps are compared.

cording to the Marshmed genetic map and the Genetic Location Database (LDB) the human LKLF gene was placed between D19S588 and D19S410 markers, which cover a region of 3.2 cM. The distal marker D19S588 has been assigned to 19p13.13 according to LDB and the proximal marker D19S410 to 19p13.11 via Genome Database (GDB) which locates LKLF gene within 19p13.11–13.

Interestingly, there are a number of zinc finger genes mapped on human chromosome 19 [22]. Moreover for some zinc finger genes in cytogenetic band 19p12 a head-to-tail arrangement was demonstrated [23]. The human erythroid Kruppel-like factor (EKLF) gene was reported [24] to map in 19p13.12–p13.13, the same locus where the human LKLF gene is located as shown above. Both genes encode the transcription factors belonging to the Kruppel family and their expression is developmentally regulated in a tissue-specific fashion [6,8]. The co-localization of these genes and similar patterns of expression in mouse blood cells pose the question of whether these genes form not only a structural but also a functional cluster with possible common regulation.

Acknowledgements: This work was supported by the Russian National Human Genome Program and The Danish Cancer Society.

References

[1] Kuo, C.T., Veselits, M.L. and Leiden, J.M. (1997) Science 277, 1986–1990.
[2] Kuo, C.T., Veselits, M.L., Barton, K.P., Lu, M.M., Clendenin, C. and Leiden, J.M. (1997) Genes Dev. 11, 2996–3006.
[3] Bellefroid, E., Lecocq, P.J., Benhida, A., Poncelet, D.A., Belayew, F. and Martial, J.A. (1989) DNA 8, 377–387.
[4] Mitchell, P.J. and Tjian, R. (1989) Science 245, 371–378.
[5] Pabo, C.O. and Sauer, R.T. (1992) Annu. Rev. Biochem. 61, 1053–1095.
[6] Miller, I.J. and Bieker, J.J. (1993) Mol. Cell. Biol. 13, 2776–2786.
[7] Crossley, M., Whitelaw, E., Perkins, A., Williams, G., Fujiwara, Y. and Orkin, S.H. (1996) Mol. Cell. Biol. 16, 1695–1705.
[8] Anderson, K.P., Kern, C.B., Crable, S.C. and Lingrel, J.B. (1995) Mol. Cell. Biol. 15, 5957–5965.
[9] Shields, J.M., Christy, R.J. and Yang, V.W. (1996) J. Biol. Chem. 271, 20009–20017.
[10] Kuo, C.T., Veselits, M.L. and Leiden, J.M. (1997) Science 278, 788–789.
[11] Wani, M.A., Means Jr., R.T. and Lingrel, J.B. (1998) Transgenic Res. 7, 229–238.
[12] Domninskii, D.A., Prokunina, L.V., Kozyrev, S.V., Bannikov, V.M., Baev, A.A., Poltaraus, A.B., Zabarovskii, E.R. and Kiselev, L.L. (1996) Mol. Biol. (Moscow) 30, 84–92.
[13] Silver, J. (1994) in: PCR: A Practical Approach (McPherson, M.J., Quirke, P. and Taylor, G.R., Eds.), Vol. 1, pp. 137–146, Oxford University Press, New York.
[14] Zabarovskii, E.R., Domninskii, D.A. and Kiselev, L.L. (1994) Mol. Biol. (Moscow) 28, 1231–1244.
[15] Baskaran, N., Kandpal, R.P., Bhargava, A.K., Glynn, M.W., Bale, A. and Weissman, S.M. (1996) Genome Res. 6, 633–638.
[16] Henke, W., Herdel, K., Jung, K., Schnorr, D. and Loening, A. (1997) Nucleic Acids Res. 25, 3957–3958.
[17] Boulikas, T. (1993) Crit. Rev. Eukaryot. Gene Express. 3, 193–227.
[18] Miller, J., McLachlan, A.D. and Klug, A. (1985) EMBO J. 4, 1609–1614.
[19] Shapiro, M.B. and Senapathy, P. (1987) Nucleic Acids Res. 15, 7155–7174.
[20] Bird, A.P. (1986) Nature 321, 209–213.
[21] Luoh, S.-W., Jegalian, K., Lee, A., Chen, E.Y., Ridley, A. and Page, D.C. (1995) Genomics 29, 353–363.
[22] Lichter, P., Bray, P., Ried, T., Dawid, I.B. and Ward, D.C. (1992) Genomics 13, 999–1007.
[23] Eichler, E.E., Hoffman, S.M., Adamson, A.A., Gordon, L.A., McCready, P., Lamerdin, J.E. and Mohrenweiser, H.W. (1998) Genome Res. 8, 791–808.
[24] van Ree, J.H., Roskrow, M.A., Becher, A.M., McNall, R., Valentine, V.A., Jane, S.M. and Cunningham, J.M. (1997) Genomics 39, 393–395.