3846–3857 *Nucleic Acids Research, 1996, Vol. 24, No. 19*

Cloning and characterization of the genomic DNA of the human *MSSP* genes

Christian Haigermoser[+], Mitsuaki Fujimoto, Sanae M. M. Iguchi-Ariga[1] and Hiroyoshi Ariga*

Faculty of Pharmaceutical Sciences and [1]College of Medical Technology, Hokkaido University, Kita 12, Nishi 6, Kita-ku, Sapporo 060, Japan

Received April 22, 1996; Revised and Accepted August 12, 1996

DDBJ/EMBL/GenBank accession nos D82351–D82362

ABSTRACT

MSSP proteins have been identified by their binding to an upstream element of *c-myc*. Independently, two different approaches yielded two cDNA clones highly homologous to the MSSP cDNAs, suggesting an involvement of MSSP in the regulation of the cell cycle (scr2) and in the repression of HIV-1 and ILR2 α-promoter transcription (human YC1). Screening human genomic libraries, we have isolated clones belonging to two different gene loci. Whereas the human *MSSP* gene 1 turned out to be intronless, the organization of the coding sequence within gene 2 is more complex. It spans more than 60 kb and contains 16 exons (including two alternative first exons), ranging from 48 to 287 bp, respectively. The intron sizes vary from 0.1 to more than 13 kb. Gene 1 has been completely sequenced. A deletion series of its upstream region was conjugated to the luciferase gene, but the transfection of the constructs did not display any promoter activity. Moreover, compared with gene 2 and the cDNA sequences known so far, about 20 point mutations as well as flanking direct repeats have been detected in the *MSSP* gene 1, showing that it possesses all the characteristics of processed retropseudogenes. Sequence analysis of a 1.7 kb fragment of the 5′ flanking region of the *MSSP* gene 2 revealed that the promoter of gene 2 lacks consensus sequences for TATA and CCAAT boxes, is GC-rich, and contains numerous potential transcription factor binding elements including an Sp1 binding site. DNase I footprinting experiments showed that the putative Sp1 site was bound by proteins. The results of primer extension and S1 mapping analyses suggested the transcription of the gene starts at multiple positions upstream from the initiator methionine codon. Luciferase assays employing progressive deletions of the 1.7 kb promoter region allowed us to define the minimal promoter region of 428 bp (–488/+) and revealed a complex pattern of the transcriptional regulation the human *MSSP* gene 2. Furthermore, it can be concluded that the *MSSP* gene 2 encodes both MSSP-1 and MSSP-2, and moreover scr2 and human YC1.

INTRODUCTION

The involvement of the proto-oncogene *c-myc* in crucial functions such as cell proliferation and differentiation requires a careful control in every cell type. Its delicately regulated expression and the elements involved therein have been the subject of numerous studies (1–3). A sequence of 21 bp about 2 kb upstream of the human *c-myc* gene has been shown to be essential for replication and transcription and to constitute both a putative DNA replication origin and a transcriptional enhancer (4,5). Its stimulation of SV40 DNA replication (6) and the functional substitution of its core sequence for the AT-stretch of the SV40 origin (7), in addition to the binding of a *c-myc* protein complex to it, suggested the role of the 21 bp sequence as a target for DNA–protein interaction (8). Indeed, several proteins showing direct binding to either of its strands could be identified and were named MSSP (*c*-*myc* single-strand binding proteins). Two cDNA clones coding for members of this protein family (MSSP-1 and MSSP-2) were characterized in respect to their DNA binding specificity as well as their promoting activity on DNA replication (9,10). Independently, the phenotypic complementation of *cdc2* and *cdc13* mutants of *Schizosaccharomyces pombe* yielded a third cDNA clone almost identical to MSSP-1 and MSSP-2, although containing a longer open reading frame, as well as a closely related one (scr2 and scr3, respectively) (11), suggesting an involvement of MSSP in the regulation of the cell cycle (12,13), especially at the G_1 to S transition, when its expression level rapidly increases (9). The sequence of a fourth highly homologous clone, termed human YC1, was submitted to GenBank as a potential human repressor of HIV-1 and ILR2 alpha promoter transcription (L11289). Though the predicted sequences of these homologous proteins display some discrepancies, they share as a common feature a novel RNA binding motif, RNP-1 (14,15), which in addition to other RNP consensus sequences is required for DNA binding (10).

To shed full light upon this complex matter, the knowledge of the gene structure and mode of expression of MSSP is a prerequisite. In the present report, human genomic clones representing two gene loci have been isolated and characterized in order to gain insight into the regulation of the *MSSP* gene(s). Whereas *MSSP* gene 1 shares all the characteristics of processed pseudogenes, *MSSP* gene 2 codes for all MSSP cDNAs, as well

* To whom correspondence should be addressed

[+]Present address: Institute of Molecular Biology, Academia Sinica, Nankang, Taipei 115, Taiwan

as scr2 and YC1. Besides the description of the detailed structure of *MSSP* gene 2, we have also investigated its 5′-flanking region to identify *cis* elements important to drive transcription by deletion analysis.

MATERIALS AND METHODS

Screening of human genomic libraries

A human placental genomic library in EMBL3-SP6/T7 and a human leukocyte genomic library in EMBL3 were purchased from Clontech. The phages were propagated in host bacteria NM538 and LE392. The infection and plating procedures were according to the recommendations of the manufacturer. Plaques were screened under stringent conditions with [α-^{32}P]dCTP-labelled gel-purified DNA fragments (10^5–10^6 c.p.m./ml of hybridization solution) as probes. Human MSSP-1 was labelled with [α-^{32}P]dCTP using a random primer labelling kit from Boehringer Mannheim. MSSP-2 cDNA and genomic DNA fragments (Fig. 1) were labelled with [α-^{32}P]dCTP by nick translation (16). Prehybridization and hybridization were performed in 50% formamide at 42°C (16). Membranes were washed twice in 3× SSC/0.1% SDS at 37°C and twice in 0.1× SSC/0.1% SDS at 50–68°C (depending on the background of the respective probe) for 30 min. The filters were then autoradiographed and analyzed with a bioimaging analyzer (BAS 2000, Fuji Film Co.) and/or placed in contact with a Fuji X-ray film and an intensifying screen and exposed at –70°C for 1–3 days. Selected recombinants that hybridized to the screening probes were rescreened and purified. Phage DNA from the purified positive plaques was prepared by the bacteriophage lysate method (16). Inserts were excised and subcloned into pUC19 and pBluescript (Stratagene). Subcloned fragments were analyzed and intron lengths were determined by a combination of restriction endonuclease digestion, Southern blotting, PCR analysis, and nucleotide sequencing. PAC clones were obtained from Genome Systems [St. Louis, MO; clone addresses PAC-85-H1 (*MSSP* gene 2), PAC-280-C14 (*MSSP* gene 1) and PAC-320-C24 (*MSSP* gene 1)] after screening with a probe spanning exons II–IV, which was produced by PCR with exon-specific primers and MSSP-2 cDNA as a template. With these clones PCR and sequence analysis were performed.

Nucleotide sequence analysis

Nucleotide sequencing was performed both manually using the chain termination method of Sanger *et al.* (17) with a Sequenase 2.0 kit (US Biochemical Corp.) and *Bca*Best kit (Takara Shuzo Co., Ltd.) and automatically on a model 373A DNA sequence (Applied Biosystems) using a fluorescent dideoxy terminator kit. In the case of the *MSSP* gene 1, nested deletions were performed on appropriately double-digested subclones with *Exo*III for different time points, mung bean nuclease-treated, ligated and transformed in *Escherichia coli* strains DH5α or C600. The resultant plasmids were sequenced with universal primers. To determine the sequence of each exon and adjacent sequence of the *MSSP* gene 2, synthetic oligonucleotides corresponding to known cDNA and genomic sequences were used besides universal primers (M4, RV, SP6, T3, and T7).

Preparation of total RNA

HeLa cells were washed with phosphate-buffered saline (PBS) and lysed with ISOGEN (Nippon Gene) according to the instructions of the manufacturer. After purification by an additional phenol extraction and ethanol precipitation, RNA was resuspended in sterile water.

Primer extension analysis

Primer extension analysis was performed using a 20mer oligonucleotide complementary to position +39 to +20 of the *MSSP* gene 2 (GCCGTGCAGGGTCGCGGACA). The primer was labelled at the 5′ end with [γ-^{32}P]ATP and T4 polynucleotide kinase. The labelled nucleotide was purified on a Sephadex G-50 spin column. A quantity of 6×10^5 c.p.m. of the primer nucleotide was co-precipitated with 120 µg of HeLa total RNA and resuspended in 30 µl hybridization buffer (40 mM PIPES, pH 6.4, 0.4 M NaCl, 1 mM EDTA, and 0.2% SDS). The mixture was heated to 55°C during 10 min and annealed at 37°C for more than 3 h. The RNA and the annealed oligonucleotide were precipitated with isopropanol and rinsed with 70% ethanol. The pellet was resuspended in reverse transcriptase buffer with 1 mM dNTP, 130 U of RNase inhibitor, and 25 U Moloney Murine Leukemia virus reverse transcriptase. Elongation was carried out for 2 h at 37°C. The reaction products were phenol/chloroform-extracted, ethanol-precipitated, and resuspended in formamide loading buffer. One third was electrophoresed on an 8% polyacrylamide–8 M urea sequencing gel along with a sequencing ladder using the same primer for the dideoxy sequencing of a 5′-genomic clone.

S1 mapping

pMSSP-Luc was digested with *Hin*dIII and treated with bacterial alkaline phosphatase before digestion with *Sma*I. The *Sma*I–*Hin*dIII fragment of 545 bp containing MSSP-1 promoter was end-labelled with [γ-^{32}P]ATP and T4 polynucleotide kinase and was used for a probe. Cytoplasmic RNA (100 µg) and the labelled DNA probe (10^5 c.p.m.) were co-precipitated with ethanol, suspended with 50 µl of hybridization buffer containing 80% formamide, 40 mM PIPES (pH 6.4), 400 mM NaCl and 1 mM EDTA, heated at 100°C for 8 min, and hybridized overnight at 42°C. The RNA mixtures were then mixed with 400 µl of ice-cold S1 nuclease buffer containing 0.25 M NaCl and 300 mM sodium acetate, 3 mM $ZnSO_4$, 100 µg/ml of salmon sperm DNA, and 10 U S1 nuclease (Takara), and incubated 25°C for 30–60 min. The S1-resistant DNA hybrids were precipitated and electrophoresed on a 10% polyacrylamide denaturing gel.

Reporter plasmids construction

The promoterless plasmid pGV-B (PicaGene^{TM}, TOYO INK) served as the vector backbone for all the luciferase expression constructs. pMSSP-Luc: Nucleotide sequences of the PCR primers used were 5′-GCTCGAGGTCTAAACCATAGAAC-3′ for MSSP-N and 5′-GAAGCTTCATGAAGCTGGAAGGG-3′ for MSSP-C. After the PCR reaction with the above primers on the λ clone containing the upstream region of MSSP gene 2 as a template, the product was digested with *Xho*I and *Hin*dIII and was inserted to the *Xho*I–*Hin*dIII sites of pGV-B. pΔX-Luc: The *Xba*I–*Hin*dIII fragment of pMSSP-Luc was first inserted to the *Xba*I–*Hin*dIII site of pBluescript SK(–). The *Sac*I–*Hin*dIII fragment from the construct was then inserted to the *Sac*I–*Hin*dIII sites of pGV-B. pΔB-Luc: The

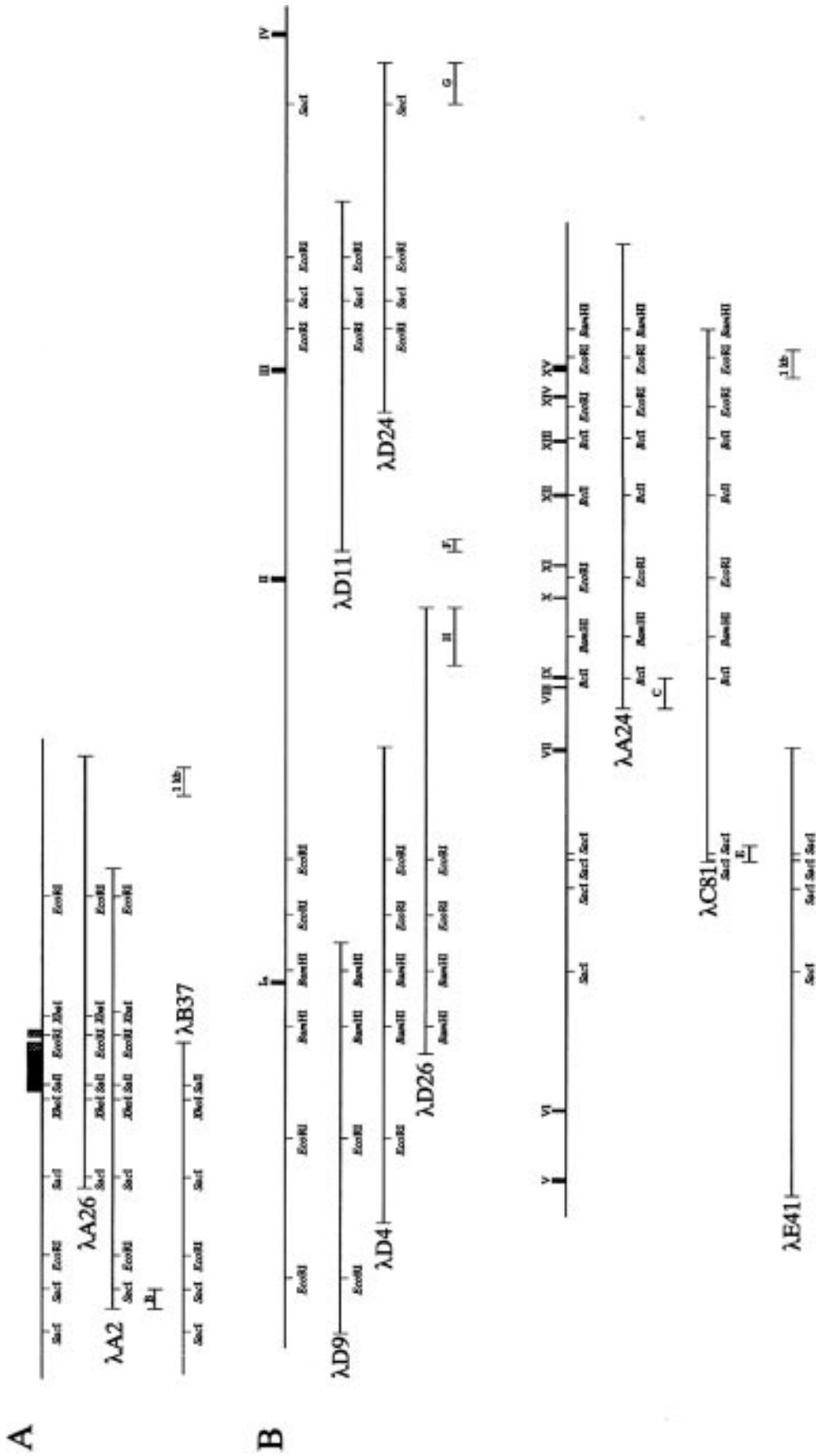
A
λA2
λA26
λB37
B
λD9
λD4
λD26
λD11
λD24
λA24
λC81
λE41

Figure 1. Organization of the human *MSSP* genes. Physical maps of the human *MSSP* gene 1 (**A**) and gene 2 (**B**) are shown. Exons (or homologous sequences, respectively) are represented as black boxes. Regions in gene 1 homologous to sequences downstream of the putative polyadenylation signal in gene 2 are shown as shaded boxes. Numbers above the boxes indicate exon numbers for the *MSSP* gene 2 transcripts. λ clones used in determining gene structure are shown below the genomic map. The fragments B, C, E, F, G and H were used for genomic library screenings. Partial restriction maps of the two loci are presented for restriction enzymes used in subclonings.

*Bam*HI–*Hin*dIII fragment of pMSSP-Luc was inserted to the *Bam*HI–*Hin*dIII site of pBluescript SK(–). The *Sac*I–*Hin*dIII fragment from the construct was then inserted to the *Sac*I–*Hin*dIII sites of pGV-B. pΔS-Luc: pMSSP-Luc was digested with *Sma*I and the larger fragment yielded was self-annealed. pΔP-Luc: pMSSP-Luc was digested with *Xho*I and *Pst*I and the larger fragment yielded was treated with Klenow fragment prior to self-ligation.

Cell culture and transient transfection

Human HeLa cells were cultured in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% calf serum. Five μg of the respective reporter plasmid and 2 μg of the β-galactosidase expression vector (pCMV-β-gal), carrying the cytomegalovirus (CMV) promoter, were co-transfected to the cells (60% confluent) by the calcium phosphate co-precipitation method (18). Four to five hours after transfection, the cells were boosted with 20% glycerol for 2–3 min at room temperature, then incubated for 48 h.

Luciferase and β-galactosidase assays

The transfected cells were washed with PBS and lysed in the plates using 200 μl of a detergent solution (lysis buffer PicaGene^{TM}, TOYO INK). The cell extract was then centrifuged for 5 s in an Eppendorf microcentrifuge and the supernatant was collected. The transfection efficiency was normalized by a β-galactosidase assay, set up in 300 μl according to the standard procedure (16), by incubation at 37°C until a yellow colour developed. The absorbance of the solution was then measured on a double beam spectrophotometer at 420 nm. The luciferase activity of the extract was determined by mixing standardized aliquots in a total of 20 μl lysis buffer with 100 μl of luciferase substrate (PicaGene^{TM}, TOYO INK) in a vial. Immediately after mixing, the light intensities emitted by the samples were measured on a luminometer (lumicounter ATP-300, Advantec Toyo Ltd.). Background luciferase activity was assessed in assays from parallel cultures transfected with the promoterless plasmid pGV-B.

DNase I footprinting

The 259 bp fragment from –547 to –289 in the MSSP gene 2 promoter was used as a probe for DNase I protection assays. To create restriction enzyme sites at both ends, PCR was carried out on pMSSP-Luc as a template using primers of FtC (5′-AGGATC-CTTGGTGCCAGGCGGCA-3′) and FtN (5′-GCTCGAG-CAACCGCGAGCCTGGG-3′). The PCR product was digested with *Xho*I and *Bam*HI, and inserted to the *Xho*I–*Bam*HI sites of pBluescript SK(–). The fragment was excised and labelled at either end with T4 polynucleotide kinase and [γ-^{32}P]ATP. The labelled probe (~3 × 10^4 c.p.m.) was mixed with HeLa nuclear extract (5, 10 or 15 μg) in a buffer containing 10 mM HEPES (pH 7.9), 150 mM KCl, 1 mM DTT, 10 mM $MgCl_2$, 1 mM EDTA and 2 mM poly(dI-dC). After incubation at 25°C for 30 min, DNase I was added to finally 10 ng/μl in the reaction. The reaction

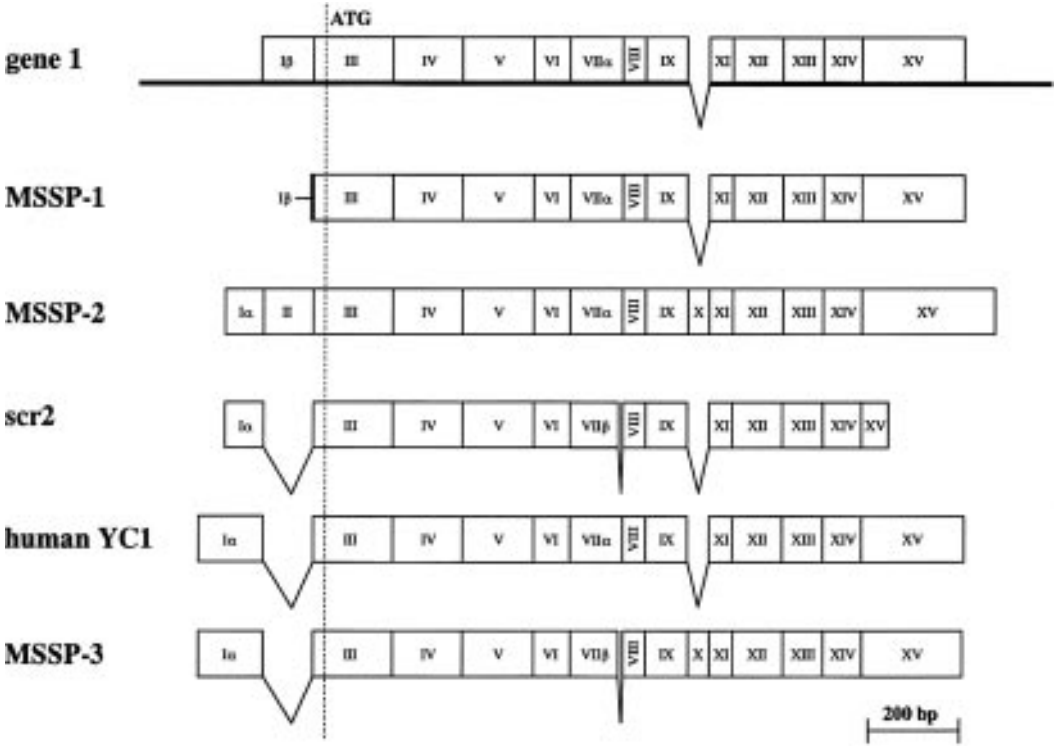

Figure 2. Comparison of the exon compositions of the known MSSP-related cDNAs and the structure of the *MSSP* gene 1. ATG indicates the position of the start codon in the MSSP-1 and MSSP-2 cDNAs.

mixtures were kept at 0°C for 80 s, then denatured and separated in an 8% polyacrylamide denaturing gel.

RESULTS

Isolation of human *MSSP* genomic DNA clones

A human placental genomic library in EMBL3 SP6/T7 was screened with labelled full-length MSSP-1 cDNA (9) using the plaque hybridization method. Three intense clones (A2, A24, and A26) were isolated, subcloned and subjected to further analysis. The restriction analysis and the sequencing of the ends of the subclones revealed that these clones represented two different genomic loci, termed *MSSP* gene 1 (A2 and A26) and *MSSP* gene 2 (A24) (Fig. 1). Partial sequence analysis of the 5′-end of the genomic clone A26 down to the unique *Sal*I site (223 bp downstream from the putative translation start in MSSP-1 cDNA) (Fig. 1A) revealed an intronless region identical to the 5′-end of the MSSP-1 cDNA (except for three mismatches) and indicated that this genomic clone extends about 3.5 kb upstream of the start of the cDNA. Attempts to screen for a sequence homologous to the head sequence of scr2 (11) (exon Iα in Fig. 2) in the far distal 5′-region of gene 1 with the genomic fragment B (Fig. 1A) and by genomic PCR did not yield the expected result. Following genomic Southern experiments showed the uniqueness of this sequence in the human genome as well as the uniqueness of another short region specific for MSSP-2 cDNA (10) (exon X in Fig. 2), missing in both MSSP-1 cDNA and gene 1 (data not shown). The discovery of the co-occurrence of these two stretches in one cDNA clone (MSSP-2) theoretically excluded the possibility of the existence of a sequence homologous to the scr2 head upstream of gene 1 (Fig. 2).

Since A24 encompassed only the last eight exons of the human *MSSP* gene 2 (Fig. 1B), further screenings with genomic fragments (C and E) and a fragment, containing ~500 bp of the 5′-end of the MSSP-2 cDNA down to the unique *Pst*I site (10) (D series), were performed and yielded clones falling into three non-overlapping regions. As neither the screening of the placental genomic library and of a human leukocyte genomic library with other genomic fragments (F, G, and H) nor PCR amplification across the gaps using genomic DNA as template proved a successful strategy for obtaining the missing sections, a PAC library was screened with a probe spanning exons II–IV. PCR and sequence analysis revealed that two of these clones belong to gene 1 (clone addresses PAC-280-C14 and PAC-320-C24) and one to gene 2 (clone address PAC-85-H1). Using exonic, intronic and Iα 5′-flanking primers, it could be demonstrated that this clone encompasses at least the region from the second *Eco*RI restriction site upstream of exon Iα (Fig. 1B) down to exon VII, and the existence of the exons Iβ, II and IV therein was confirmed. These exons had not been covered by the bacteriophage clones. Thus, a contig spanning the whole genomic locus of MSSP gene 2 was finally established with the overlapping bacteriophage and PAC clones.

Characterization of the human *MSSP* genes

The sequences in and around the exons of the *MSSP* gene 2 and those of the junctions of bacteriophage and plasmid subclones were determined. The comparison of the sequence of the human *MSSP* gene 2 (Fig. 3) with those of the cDNAs published so far shows that the human *MSSP* gene 2 is organized into 16 exons (including two alternative first exons, two optional ones and one with an internal splice site) (Fig. 2) and 15 intervening sequences, spanning a total of more than 60 kb. The exons are distributed sparsely at the 5′-end of the gene but rather densely at the 3′-end (Fig. 1B). Their sizes are rather small (Table 1). All the exon–intron junction sequences conform to the GT/AG rule (19) (Table 2). The sequence of the known exons coincides with those of all the known cDNAs, except for some mismatches most probably introduced by cloning procedures or sequence misreadings.

Table 1. The sizes of exons and introns in the human *MSSP* gene 2

Exon	Length (bp)	Amino acid position[a]	Intron	Type[b]	Length (kb)
Iα	132	–33 to –9	1	0	>13
Iβ	≥103	–31 to –9	1/2 β	0	>5
II	110		2	–	>5
III	176	(–8) 1 to 51	3	2	>12
IV	151	51 to 101	4	0	>0.5
V	158	102 to 154	5	2	3
VI	80	154 to 181	6	1	13
VIIα(β)	116 (107)[c]	181 to 219 (216)[c]	7	0 (0)	2.0
VIII	50	220 to 236	8	2	0.125
IX	94	236 to 267	9	0	2.8
X	48	268 to 283	10	0	0.892
XI	51	284 to 300	11	0	2.9
XII	111	301 to 337	12	0	1.4
XIII	81	338 to 364	13	0	1.7
XIV	85	365 to 389	14	–	0.865
XV	287				

[a]Based upon the human *MSSP* gene 2 sequence in Figure 3.
[b]Intervening sequence type is defined where type 0 indicates placement between codons, type 1 interrupts a codon between the first and second nucleotide, and type 2 occurs between the second and third nucleotide of a codon (39).
[c]Alternative splicing sites (Table 2)

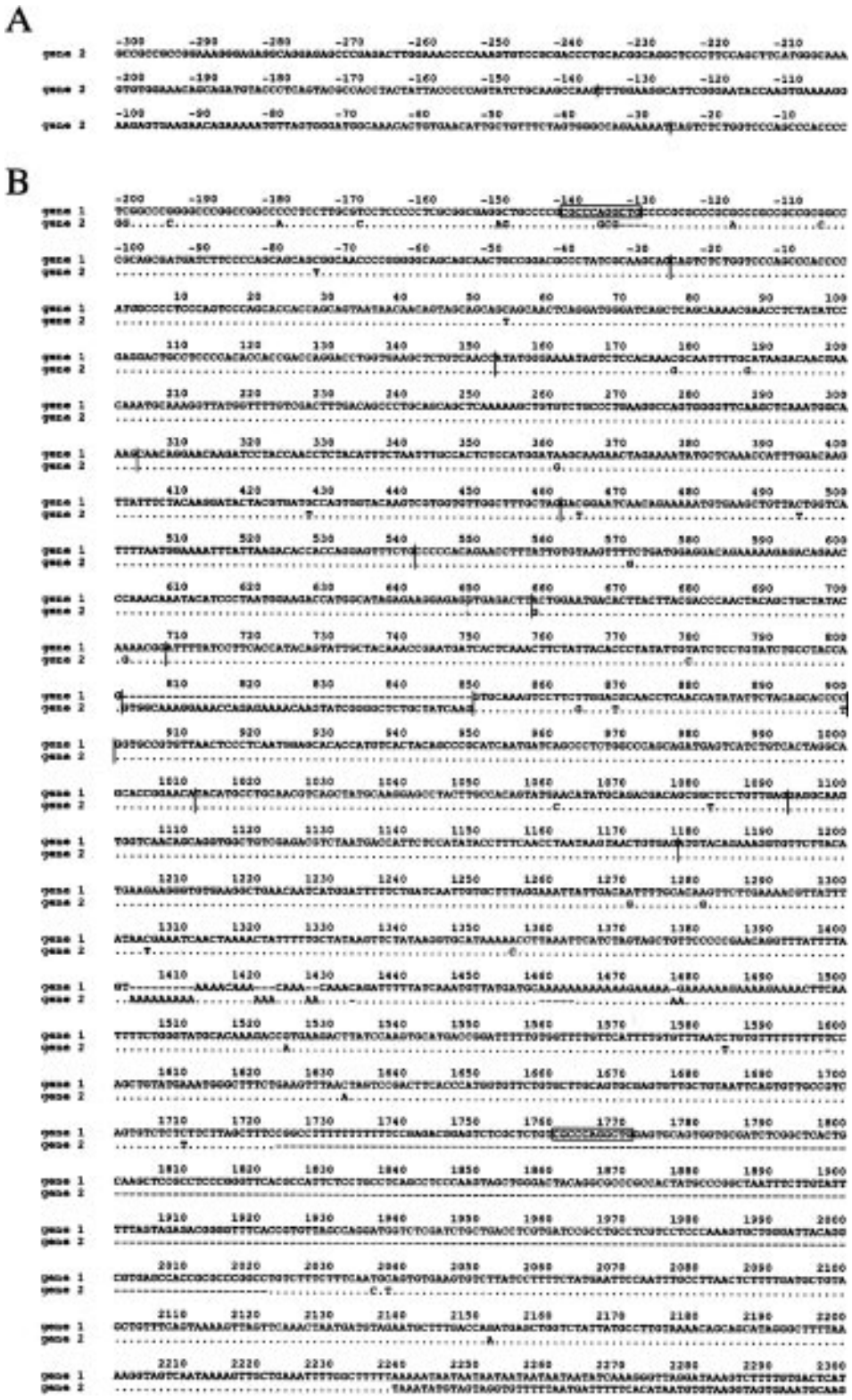

Figure 3. Intronless sequences of the human *MSSP* genes. The A residue which corresponds to the first nucleotide in the start codon of the MSSP-1 and MSSP-2 cDNAs (Fig. 2) has been assigned as residue +1. Exon–exon junctions are marked by vertical lines (a broken line for an internal splice site). (**A**) The 5′-flanking region of exon Iα is linked to the sequences of the exons Iα, II and III down to –1, which corresponds to the splicing pattern of the MSSP-2 cDNA (Fig. 2). (**B**) Alignment of the sequences of the human *MSSP* genes. The sequence of gene 2 is composed by linkage of the 5′-flanking sequence of exon Iβ and the 3′-flanking sequence of exon XV to the exon sequences of gene 2 spliced according to the pattern found in the MSSP-1 cDNA (Fig. 2), except for the insertion of the sequence of exon X. Moreover, the positive part of gene 2 down to the polyadenylation signal corresponds to the MSSP-2 cDNA and can also be regarded as a continuation of A. Dots denote sequence identity, and deletions are marked by dashes. The putative start codons (–93 and 1 for gene 1 (B), –209 and 1 for gene 2 (A or B, respectively), the stop codon (1168) and two potential polyadenylation signals (1414 and 1469) are italicized. The 11 bp direct repeats flanking the *MSSP* gene 1 are boxed.

Table 2. Exon–intron boundaries of the human *MSSP* gene 2

5'								3'
		-9/-135				-/-134		
exon Iα	CAA GCC AAG:	gtaaaggggcgccggggaga	intron 1	gtttcttttatttcctggag:	TTT GGA AGG	exon II
		-9/-25				-8/-24		
exon Iβ	CGC AAG CAG:	gtgaggcggcggcngggaan	intron 1/2β	tcctgatctttgttttccag:	CAG TCT CTG	exon III
		-/-25				-8/-24		
exon II	CAG AAA AAT:	gtatgtataaatcttccctg	intron 2	tcctgatctttgttttccag:	CAG TCT CTG	exon III
		51/152				102/304		
exon III	TGT CAA CC :	gtaagtagtagttgttccta	intron 3/4	aacactctcttctgttttag:	CAA CAG GAA	exon V
		154/461				154/462		
exon V	TTT GCT AG :	gtaagccttctgcccagtct	intron 5	tctatttttattataaatag:	G ATG GAA	exon VI
		181/541				181/542		
exon VI	GTT TCT G :	gtatgttgtttgtctgaaac	intron 6	atttttctctgtttacttag:	CC CCC ACA	exon VIIα/β
		219/657				220/658		
exon VIIα	GTG AGA CTT:	gtaagtccttttcttgaaact	intron 7α	tttctttctctttcccatag:	GCT GGA ATG	exon VIII
		216/648				220/658		
exon VIIβ	(GAA GGA GAG:	gtgagacttgtaagtccttt	intron 7β	tttctttctctttcccatag:	GCT GGA ATG)	exon VIII
		236/707				236/708		
exon VIII	CAG AAC GG :	gtatgtcgttaaaataaatc	intron 8	ggcttttgtgtttgttctag:	A TTT TAT	exon IX
		267/801				268/802		
exon IX	GCC TAC CAG:	gtttgtacctttaggattcg	intron 9	ttaactttggtctgccttag:	GTG GCA AAG	exon X
		283/849				284/850		
exon X	GCT ATC AAG:	gtaaatcgcaatacattctt	intron 10	caaaatcctccctttctag:	GTG CAA AGT	exon XI
		300/900				301/901		
exon XI	CAG CAC CCT:	gtaagtttttatcttaatg	intron 11	tcctgtctccaactccccag:	GGT GCC GTG	exon XII
		337/1011				338/1012		
exon XII	ACC GGA ACA:	gtaggtggcaaataattatc	intron 12	gacctcattcttctttctag:	TAC ATG CCT	exon XIII
		364/1092				365/1093		
exon XIII	CCT GTT GAG:	gttggtagagaccatccagt	intron 13	ttctgctttgtttgttgtag:	GAG GCA AGT	exon XIV
		-/1177				-/1178		
exon XIV	ACT GTG AG :	gtatgagggaaaggtcttca	intron 14	cttttatttttctcttccag:	A TGT ACA	exon XV

Exon sequence is presented in upper case letters and intron sequence in lower case.
The numbers identify codons within the coding sequence or the nucleotides of the *MSSP* gene 2 in Figure 3, respectively.
The less common type intron 7β, which only occurs in scr2 and MSSP-3, is in parentheses.

Table 3. Amino acid exchanges in the human *MSSP* gene 1

Number	Position	Nucleotide exchange	Amino acid number	Amino acid exchange
1	–73	AG**T**→AG**C**	–25	no
2	54	AG**T**→AG**C**	18	no
3	177	AA**G**→AA**C**	59	Lys→Asn
4	187	**G**AT→**C**AT	63	Asp→His
5	361	**G**AG→**A**AG	121	Glu→Lys
6	427	**T**CC→**G**CC	143	Ser→Ala
7	464	AT**G**→A**C**G	155	Met→Thr
8	494	AT**T**→A**C**T	165	Ile→Thr
9	571	**G**CT→**T**CT	191	Ala→Ser
10	658	**G**CT→**A**CT	220	Ala→Thr
11	702	CA**G**→CA**A**	234	no
12	779	G**C**A→G**T**A	260	Ala→Val
13	864 (816)	TC**G**→TC**T**	288 (272)	no
14	869 (821)	AT**G**→A**C**G	290 (274)	Met→Thr
15	900 (852)	CC**T**→CC**C**	300 (284)	no
16	1061 (1013)	G**C**A→G**A**A	354 (338)	Ala→Glu
17	1082 (1034)	GT**T**→G**C**T	361 (345)	Val→Ala

Nucleotides subject to exchanges are indicated in bold type.
The amino acid number and position identify codons within the coding sequence or nucleotides of the *MSSP* gene 2 (Fig. 3B), respectively. Differing values for the *MSSP* gene 1 are shown in parentheses.

The completion of the sequencing of the coding region of the *MSSP* gene 1 and its 3′-flanking region down to the first *Xba*I site (Fig. 1A) revealed that gene 1 is intronless. Figure 3B shows a sequence comparison of the human *MSSP* genes at the nucleotide level. The *MSSP* gene 1 harbours five nucleotide exchanges with no effect on the amino acid composition of a possible gene product, whereas 12 nucleotide exchanges would lead to amino acid exchanges (Table 3), one of which would destroy an RNP consensus (Met^{155}→Thr^{155}). While there are neither insertions nor deletions in the coding sequence, added or missing nucleo-

Figure 4. The 5′-regulatory region of the human *MSSP* gene 2. Nucleotides are numbered in the right margin relative to the transcription start site determined by primer extension, 57 bp upstream from the translation start codon in exon Iα (italicized). Consensus sequences for selected putative transcription factors are underlined, with abbreviations of the corresponding transcription factors shown below the sequences. Restriction enzyme sites are denoted in outlined letters and labelled. The sequence complementary to the oligonucleotide used in the primer extension analysis is represented in bold letters. The start site of transcription is marked by a hooked arrow. TG and GGA repeats are boxed. Intronic sequences are shown in lower case.

tides can be observed at the end of the region homologous to the longest cDNAs (region 1402–1470 in Fig. 3B). The theoretical length of 404 or 373 amino acids corresponds exactly to the size of the coding sequence of the correspondingly spliced product of gene 2 or the MSSP-1 cDNA, respectively. Although consensus sequences for transcriptional factors were discovered in the 5′-flanking region of gene 1, deletion studies with luciferase constructs did not display any promoter activity, neither were specific bands detected by nuclease S1 mapping analysis (data not shown). The discovery of the existence of flanking 11 bp direct repeats definitely confirmed that the human *MSSP* gene 1 shares all the characteristics of processed pseudogenes (20–22).

Sequence of the 5′-flanking region of the *MSSP* gene 2

Nucleotide sequence analysis of the 5′ end of D26, containing ~1.7 kb of the 5′-flanking region, exon Iα and part of intron 1 (down to the first *Eco*RI site) (Fig. 1B) of the *MSSP* gene 2, shows that this region is GC-rich and does not contain TATA and CAAT boxes at their characteristic positions in relation to the transcription initiation site (Fig. 4). However, it contains several consensus sequences for transcription factor binding sites such as Sp1 and AP-2. Two series of GGA repeats with sequence homology to the Epstein-Barr virus IR3 and a stretch of the Herpes simplex virus 2 genome were noted from –173 to –102, surrounding two

potential PEA3 binding sites. Another intriguing element found in the *MSSP* gene 2 promoter is an alternating GT residues stretch of 26 bp located between nucleotides –1001 and –976. A similar sequence has been found in the 5′-flanking region of several other genes (23–27). It has been suggested that such 'TG element' sequences may have a Z-DNA conformation and serve as an enhancer element (28). With the exception of these simple sequence repeats upstream of exon Iα, none of the new DNA sequence showed statistically significant similarity to those already in GenBank.

Mapping of the transcription initiation sites in the 5′-flanking region of the *MSSP* gene 2

To determine the transcription initiation sites in the 5′-flanking region of gene 2, primer extension and S1 nuclease mapping were performed. Primer extension using a 20 bp oligomer complementary to the sequence from position + 20 to +39 gave a major initiation site at A (Fig. 5A) localized 57 nucleotides upstream of the first putative ATG initiation codon in exon Iα. Accordingly, this base was designated as +1 bp and corresponded to the 5′ end of the longest cDNAs human YC1 and MSSP-3 [unpublished clone isolated along with MSSP-2 (10)]. S1 mapping using the probe spanning from –488 to +61 revealed several initiation sites including the +1 position determined by the primer extension analysis above (Fig. 5B), and the major initiation sites were nucleotide A at +11 and nucleotide A at +12. The sequence surrounding the initiation sites matches nucleotides with the loose consensus sequences defining an initiator element (29,30).

Functional analysis of the promoter in the 5′-flanking region of the *MSSP* gene 2

To determine the region bearing an active promoter in the *MSSP* gene 2, various segments upstream from the gene were cloned into a promoterless vector pGV-B containing the firefly luciferase gene and the constructs were examined for expression of the enzyme activity. The chimeric MSSP promoter-luciferase plasmids were introduced into HeLa cells by the calcium-phosphate transfection method, and the luciferase activity of the cell lysate was assayed (Fig. 6). The transfection efficiency was monitored by β-galactosidase activity due to the co-transfected plasmid containing the β-galactosidase gene driven by the CMV promoter. In preliminary experiments, two constructs harboring large MSSP gene segments, starting at the second *Bam*HI site upstream of exon Iα, yielded a reasonably high level of luciferase synthesis. The region from –1709 to +61 were then tested for promoter activity (Fig. 6A). When the sequence was deleted from upstream as far as position –1283 (*Xba*I site), a higher level of promoter activity appeared than that due to the whole region, suggesting that negative regulatory element(s) exist in the region from –1709 to –1024. Deletion as far as the position –1024 (*Bam*HI site) decreased the luciferase activity, which implies positive element(s) therein. The transcription activity was drastically reduced by deleting the 292 bp fragment between –488 and –196. Positive regulatory element(s) of importance were thus suggested to exist between the positions –488 and –196, where a cluster of consensus binding sites for various transcription factors including Sp1 and AP-2 (Fig. 4) is located.

Potential protein binding to the sequence between –545 and –369 in the MSSP gene 2 promoter region was tested by DNase I footprinting analysis using HeLa nuclear extract (Fig. 7). The nucleotides from –473 to –431 in upper strand, and those from –473 to –440 in lower strand, were protected from DNase I digestion by the proteins in HeLa nuclear extract. The protected segment contains an Sp1 recognition sequence, implying the involvement of Sp1 in the transcriptional regulation of the MSSP gene 2.

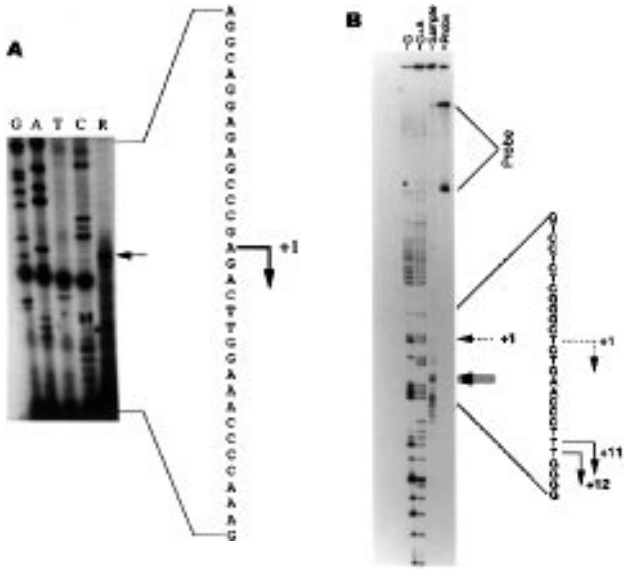

Figure 5. Determination of the transcription start sites in the 5′-flanking region of the human *MSSP* gene 2. (**A**) Primer extension analysis. Total RNA from HeLa cells was hybridized to a 5′ end-labelled 20mer oligonucleotide corresponding to positions +39 to +20 in Figure 4. The primer was extended with reverse transcriptase and analyzed by electrophoresis on a denaturing polyacrylamide gel (lane R). Sequencing of a 5′-genomic clone with this nucleotide was used for calibration (lanes G, A, T and C). The start site of transcription is marked by an arrow, and the corresponding nucleotide is indicated by a bent arrow. (**B**) S1 nuclease mapping. Total RNA was hybridized with a ^{32}P-labelled probe of the sequence spanning from –488 to +61, digested with S1 nuclease, and run on a denaturing polyacrylamide gel (sample). A control reaction without RNA was also performed and similarly analyzed (probe). The G and G+A reactions (Maxam-Gilbert chemical cleavages) of the probe were run on the same gel in parallel as size markers.

DISCUSSION

Screening human genomic libraries with the MSSP-1 and MSSP-2 cDNAs and several genomic fragments, clones from two different genomic loci were obtained. Whereas the human *MSSP* gene 1 turned out to be intronless, the organization of the coding sequence within gene 2 is more complex. Gene 1 has been completely sequenced. The alignment of gene 1 with the upstream region of the alternative first exon Iβ and the downstream region of the last exon of gene 2 shows homology close to identity (Fig. 3B). Compared with the exonic sequences of gene 2 and the cDNA sequences known so far, including those of recently published expressed sequence tags (EST) and that of MSSP-3, the *MSSP* gene 1 contains about 20 point mutations, none of which interrupts the reading frame nor causes frameshifts, which demonstrates that none of the cDNAs results from a transcript of this gene. The region of gene 1 which corresponds to the cDNA sequences is bounded by 11 bp direct repeats, the homology between the two genes, however, extends even beyond them. The 5′ direct repeat is surrounded by sequences related to

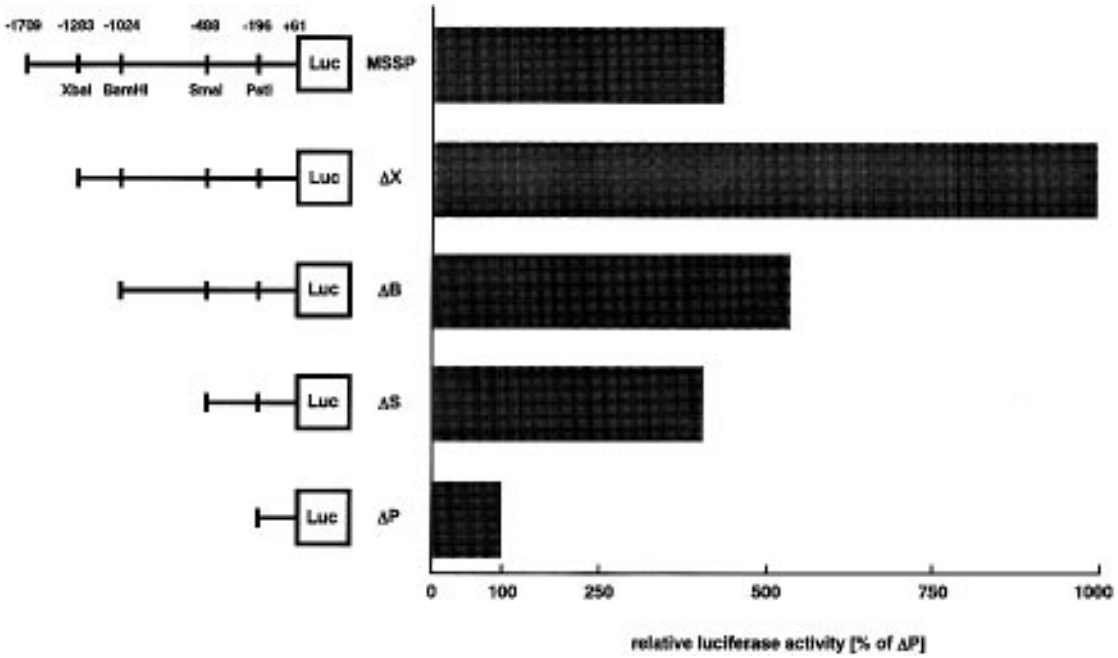

Figure 6. Promoter activities of upstream regions of the *MSSP* gene 2. The schematic diagram represents the human *MSSP* gene 2 promoter from –1709 to +61. The numbers refer to the position of the ends of the construct inserts relative to the transcription initiation site. Restriction sites used for subcloning are also indicated. HeLa cells were co-transfected with the indicated luciferase plasmids and pCMV-β-gal. Luciferase activity is expressed as the average percentage of the chemiluminescence count for each construct standardized by that of pΔP-Luc (set at 100) after correction of transfection efficiency. Data represent the mean from at least three independent transfection experiments.

exon Iβ. The situation of the sequence around the 3′ direct repeat is more complex. The homology between gene 1 and gene 2 in the 3′-flanking region continues beyond the putative polyadenylation signal, spanning about 470 bp interrupted after ~250 bp by a 299 bp insert, which comprises the downstream direct repeat sequence. The lack of this tract in gene 2 could be interpreted by its deletion in gene 2 rather than by its insertion in gene 1 after the formation of gene 1. Interestingly, the bounding sequences of the gap show a high degree of homology. As long as no sequence data of exon Iβ upstream of –200 in Figure 3B are available, further homology in the 5′-flanking region cannot be excluded, either. The lack of a poly-A stretch before the 3′ direct repeat might be due to reverse transcription before its addition to the fully spliced RNA. Negative Northern data, no hybridization of the homologous 3′-flanking sequences to HeLa RNA, would support this hypothesis. The percentage identity to the mRNA at the nucleotide level is 98.6% considering only the coding sequences, a typical value for a processed retrogene, suggesting that the *MSSP* gene 1 has arisen relatively recently in evolution. Commonly, the analysis of retrogenes does not go beyond a sequence comparison. Experiments to demonstrate the lack of promoter activities or transcription sites are hardly performed. In view of the absence of deletions, insertions or mutations to stop codons in the reading frame of gene 1, these approaches were considered necessary to provide further evidence for a final conclusion. Deletion series (from the 5′ and the 3′ end) of its upstream region were cloned upstream of the luciferase gene into vectors with and without the SV40 promoter, but never was any significant activity above background level obtained with constructs of the SV40-promoterless series. An S1 mapping was started at a very early stage, before the knowledge of the structure of gene 2, assuming the uniqueness of the 5′ sequence of gene 1. We primarily obtained numerous non-specific bands due to S1 nuclease-sensitive sites. After the discovery of a second copy of this region in the human genome (exon Iβ) it was realized that an independent analysis of this region is impossible and that an optimal choice of the probe and an accurate interpretation of the results has to await the determination of the precise 5′ end of the homology between the two genes and the functional analysis of the Iβ5′-flanking region. Since only the fortuitous placement of a retrosequence next to an active promoter could theoretically result in transcription, the absence of deletions, insertions or mutations to stop codons within the reading frame is only of secondary importance. In the case of gene 1, even on the assumption that it may be transcribed at a very low level, which cannot be ruled out completely, the mutation $Met^{155} \rightarrow Thr^{155}$ disrupts an RNP consensus and most probably precludes translation of the putative transcript into a functional polypeptide. In conclusion, although the precise reconstruction of its formation may be difficult to accomplish, the *MSSP* gene 1 arguably shares all the hallmarks of retropseudogenes.

A set of recombinant DNA clones that contain the entire *MSSP* gene 2 were isolated, and the gene 2 was shown to span more than 60 kb and to contain a total of 16 exons (including two alternative first exons), ranging from 48 to 287 bp. The intron sizes vary from 0.1 to more than 13 kb. The 5′-flanking region of the *MSSP* gene 2 contains a functional promoter: A relatively high promoter activity was observed in the region from –488 to +61 (relative to the transcription site), which can be divided into a sequence essential for transcription (the region up to –196) and the other sequence necessary for maintenance of a high expression level (–488 to –196). The complex patterns of transcriptional activity of the 5′ deletion mutants suggests that both positive and negative elements are involved in the regulation of *MSSP* gene 2 transcription. The presence of consensus binding sites for various transcription factors implies, but does not prove, their involve-

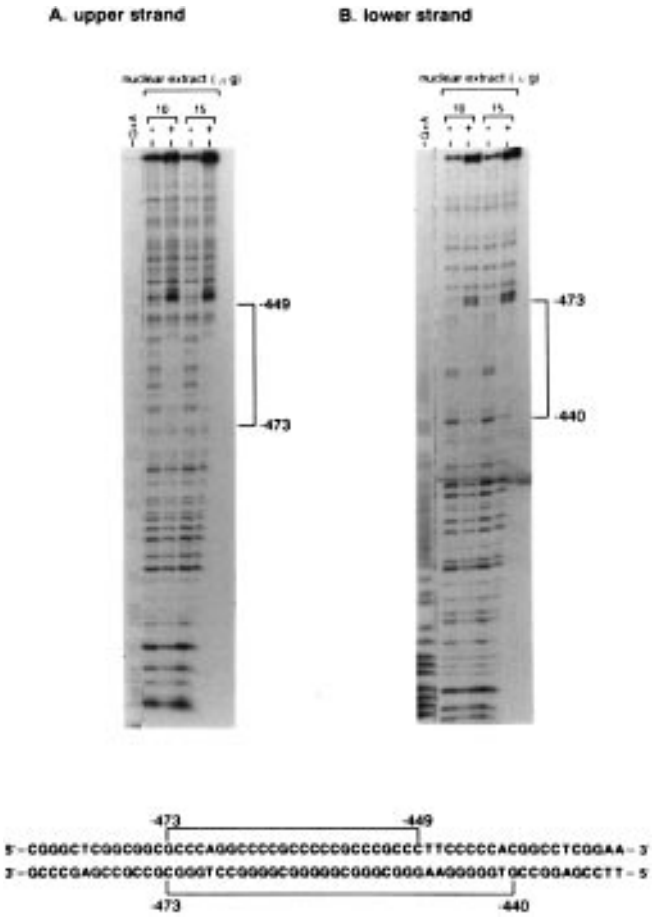

Figure 7. DNase I footprinting analysis of the region from –547 to –289 in MSSP gene 2. The *Xho*I (upper strand) or the *Bam*HI (lower strand) site at either end was labelled. Approximately 3×10^4 c.p.m. of either labelled probe was incubated with the nuclear extract of HeLa cells, partially digested with DNase I and separated on a 10% denaturing polyacrylamide gel. A+G and G indicate the positions of A and G residues of the fragment analyzed as determined by Maxam-Gilbert chemical cleavages. + or – shows the presence or absence of the nuclear extract of HeLa cells in DNase reactions. The sequences bound by proteins are indicated by lines beside the gel and in the lower panel of the figure.

ment in the regulation of the promoter activity. The presence of overlapping binding sites (e. g. Sp1, AP-2, GC-box etc. at –460) suggests a balance activity of these transcription factors in different physiological conditions. Among the putative binding sites, at least the Sp1 binding site was protected from DNase I digestion by the nuclear proteins from HeLa cells, suggesting the involvement of Sp1 in the MSSP gene 2 transcription. Other notable features of the nucleotide sequence in the 5′-flanking region of the *MSSP* gene 2 are the presence of a 26 bp TG repeat (TG-element) at position –1001 to –976 and two series of GGA repeats from –173 to –102. Tandem repeat sequences of 2–5 reiterated nucleotides are frequently found in eukaryote genomes. The most common type is the dinucleotide CA repeat which can appear in up to 10^5 different locations, each of which contains up to 60 bp (31). Their varying length provides a useful system for the generation of genetic markers which can be used for mapping and linkage analysis (32). These elements can induce a conformational change from right-handed B-DNA to left-handed Z-DNA and negative supercoiling (33). It is not clear whether these conformational changes play a role in transcriptional regulation, either as part of an enhancer or as part of a silencer. Additional investigations by deletion analysis and site-directed mutagenesis will be required to define more precisely the *cis*-acting elements and *trans*-acting factors important in the cell cycle-regulated expression of the human *MSSP* gene 2.

Potential polyadenylation sequences (AACAAA and AAGAAA) in the 3′ end of the human *MSSP* genes differ from the putative poly(A) signal AAUAAA. However, polyadenylation sites different from the canonical poly(A) signal have been reported (34,35), suggesting that a perfect hexanucleotide AAUAAA is not an absolutely essential element in the efficient polyadenylation of MSSP transcripts.

The divergence of the cDNA sequences at the point where the 5′ exons splice onto exon III (or also onto exon II in the case of exon Iα) suggested that gene 2 might have an additional 5′ exon located somewhere upstream of exon III. The 12 bp head sequence in the MSSP-1 cDNA (corresponds to the region –25 to –36 in Fig. 3B), which is only slightly different from the 3′ end of exon 1α (–135 to –149 in Fig. 3A), also exists in gene 1. Assuming that the homology between the two genes continues farther towards the 5′ end, an 18 bp primer homologous to the 12 bp and 6 additional bp in gene 1 (–25 to –42 in Fig. 3B) was designed and used for sequencing the PAC clone containing gene 2. Gratifyingly, this approach finally resulted in the discovery of exon 1β and extended the known scope of homology between the two *MSSP* genes up to –200 in Fig. 3B. All cDNAs cloned so far are consequently not transcripts of different genes but alternatively spliced products of the same gene. In other words, gene 2 encodes MSSP-1, MSSP-2 and MSSP-3, as well as scr2 and YC1.

Further studies will be necessary to pinpoint exon Iβ, to perform functional analysis of its 5′-flanking region and to map its transcription start site(s), although we could not detect any transcript in the region upstream of exon III under the conditions used in this experiment. These should also include: a comparison of the expression patterns of Iα and Iβ transcripts; their relative abundance at different developmental stages and/or in different tissues; the identification of the set or combination of transcription factors for each promoter; and a comparison of their secondary structures. Thus, the meaning and importance of the differential promoter usage could be determined. The differential utilization of two alternative promoter sequences may even provide a mechanism for translational regulation of the *MSSP* gene 2. Due to the non-excludable possibly non-coding nature of exon Iα and Iβ (all fusion proteins of MSSP used in *in vitro* studies lacked the scr2 head sequence and started at +1 in Fig. 3B), transcription initiation from both promoters might even result in identical protein products. Thus, the transcription under the control of multiple promoters would result in an alternative splicing of sequences within the 5′ UTR, which might regulate the expression of the *MSSP* gene 2 at the translational level. Lastly, determining the sequence of the alternate candidate promoter in the Iβ 5′-flanking region should reveal the precise range of homology between the two *MSSP* genes and should eventually allow the development of more specific probes to prove or disprove the possibility of a transcription of gene 1.

Along with the knowledge of the gene structure, the dissection of the structure and transcriptional activity of the *MSSP* gene 2 promoter region reported herein will facilitate further study on its regulation, thus contributing to a better understanding of the

physiological and cellular roles of MSSP, e.g. its involvement in the modulation of the biological functions of *c-myc*.

ACKNOWLEDGEMENTS

This work was supported by grants from the Japanese Ministry of Education, Science and Culture. We thank Kiyomi Takaya for technical assistance. C. H. received a Foreign Student Scholarship from the Japanese Ministry of Education, Science and Culture.

REFERENCES

1 Alitalo, K., Koskinen, P., Mäkelä, T. P., Saksela, K., Sistonen, L., and Winqvist, R. (1987) *Biochim. Biophys. Acta* **907**, 1–32.
2 Lücher, B. and Eisenman, R. N. (1990) *Genes Dev.* **4**, 2025–2035.
3 Meichle, A., Philipp, A. and Eilers, M. (1992) *Biochim. Biophys. Acta* **1114**, 129–146.
4 Iguchi-Ariga, S. M. M., Okazaki, T., Itani, T., Ogata, M., Sato, Y., and Ariga, H. (1988) *EMBO J.* **7**, 3135–3142.
5 Ariga, H., Imamura, Y., and Iguchi-Ariga, S. M. M. (1989) *EMBO J.* **8**, 4273–4279.
6 Kumano, M., Nakagawa, T., Imamura, Y., Galli, I., Ariga, H., and Iguchi-Ariga, S. M. M. (1992), *FEBS Lett.* **309**, 146–152.
7 Galli, I., Iguchi-Ariga, S. M. M., and Ariga, H. (1993) *FEBS Lett.* **318**, 335–340.
8 Negishi, Y., Iguchi-Ariga, S. M. M., and Ariga, H. (1992) *Oncogene* **7**, 543–548.
9 Negishi, Y., Nishita, Y., Saegusa, Y., Kakizaki, I., Galli, I., Kihara, F., Tamai, K., Miyajima, N., Iguchi-Ariga, S. M. M., and Ariga, H. (1994) *Oncogene* **9**, 1133–1143.
10 Takai, T., Nishita, Y., Iguchi-Ariga, S. M. M., and Ariga, H. (1994) *Nucleic Acids Res.* **22**, 5576–5581.
11 Kataoka, Y. and Nojima, H. (1994) *Nucleic Acids Res.* **22**, 2687–2693.
12 Nurse, P. (1985) *Trends Genet.* **1**, 51–55.
13 Norbury, R. and Nurse, P. (1992) *Annu. Rev. Biochem.* **61**, 441–470.
14 Bandziulis, R. J., Swanson, M. S., and Dreyfuss, G. (1989) *Genes Dev.* **3**, 431–437.
15 Dreyfuss, G., Swanson, M. S., and Pinol-Roma, S. (1988) *Trends Biochem. Sci.* **13**, 86–91.
16 Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, Second Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.
17 Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl Acad. Sci. USA* **74**, 5463–5467.
18 Graham, F. L. and Van der Eb, A. J. (1973) *Virology* **52**, 456–467.
19 Breathnach, R. and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383.
20 Rogers, J. H. (1985) *Int. Rev. Cytol.* **93**, 187–279.
21 Vanin, E. F. (1985) *Annu. Rev. Genet.* **19**, 253–272.
22 Weiner, A. M., Deininger, P. L., and Efstratiadis, A. (1986) *Annu. Rev. Biochem.* **55**, 631–661.
23 Altmeyer, A., Klampfer, L., Goodman, A. R., and Vilcek, J. (1995) *J. Biol. Chem.* **270**, 25584–25590.
24 Xie, Q.-W., Whisnant, R., and Nathan, C. (1993) *J. Exp. Med.* **177**, 1779–1784.
25 Danoff, T. M., Lalley, P. A., Chang, Y. S., Heeger, P. S., and Neilson, E. G. (1994) *J. Immunol.* **152**, 1182–1189.
26 Lin, Y.-H., Shin, E. J., Campbell, M. J., and Niederhuber, J. E. (1995) *J. Biol. Chem.* **270**, 25968–25975.
27 Mori, Y., Folco, E., and Koren, G. (1995) *J. Biol. Chem.* **270**, 27788–27796.
28 Hamada, H., Seidman, M., Howard, B. H., and Gorman, C. M. (1984) *Mol. Cell. Biol.* **4**, 2622–2630.
29 Kaufmann, J. and Smale, S. T. (1994) *Genes Dev.* **8**, 821–829.
30 Javahery, R., Khachi, A., Lo, K. Zenzie-Gregory, B., and Smale, S. T. (1994) *Mol. Cell. Biol.* **14**, 116–127.
31 Hamada, H. and Kakunaga, T. (1982) *Nature* **298**, 396–398.
32 Weber , J. L. and May, P. E. (1989) *Am. J. Hum. Genet.* **44**, 388–396.
33 Haniford, D. B. and Pulleyblank, D. E. (1983) *Nature* **302**, 632–634.
34 Lee, K.-L. D., Pentecost, B. T., D'Anna, J. A., Tobey, R. A., Gurley, L. R., and Dixon, G. H. (1987) *Nucleic Acids Res.* **15**, 5051–5068.
35 Stros, M. and Dixon, G. H. (1993) *Biochim. Biophys. Acta* **1172**, 231–235.