Original Contributions

Mammalian Genome
Incorporating *Mouse Genome*

Mammalian Genome 11, 182–190 (2000).

Comparative sequence analysis of the *MECP2*-locus in human and mouse reveals new transcribed regions

Kathrin Reichwald,[1] Jens Thiesen,[2] Thomas Wiehe,[1] Joachim Weitzel,[2] Wolf H. Strätling,[2] Petra Kioschis,[3] Annemarie Poustka,[3] André Rosenthal,[1,4] Matthias Platzer[1]

[1]Institut für Molekulare Biotechnologie, Abt. Genomanalyse, Beutenbergstr. 11, 07745 Jena, Germany
[2]Universitäts-Krankenhaus Eppendorf, Institute für Medizinische Biochemie und Molekularbiologie, Martinistr. 52, 20246 Hamburg, Germany
[3]DKFZ, Abt. für Molekulare Genomanalyse, Im Neuenheimer Feld 506, 69120 Heidelberg, Germany
[4]Friedrich-Schiller-Universität, Schloßgasse 1, 07743 Jena, Germany

Received: 23 July 1999 / Accepted: 12 November 1999

Abstract. Comparative sequence analysis facilitates the identification of evolutionarily conserved regions, that is, gene-regulatory elements, which can not be detected by analyzing one species only. Sequencing of a 152-kb region on human Chromosome (Chr) Xq28 and of the synthenic 123 kb on mouse Chr XC identified the MECP2/*Mecp2* locus, which is flanked by the gene coding for Interleukin-1 receptor associated kinase (IRAK/*Il1rak*) and the red opsin gene (RCP/*Rsvp*). By comparative sequence analysis, we identified a previously unknown, non-coding 5′ exon embedded in a CpG island associated with MECP2/*Mecp2*. Thus, the MECP2/*Mecp2* gene is comprised of four exons instead of three. Furthermore, sequence comparison 3′ to the previously reported polyadenylation signal revealed a highly conserved region of 8.5 kb terminating in an alternative polyadenylation signal. Northern blot analysis verified the existence of two main transcripts of 1.9 kb and ~10 kb, respectively. Both transcripts exhibit tissue-specific expression patterns and have almost identical short half-lifes. The ~10-kb transcript corresponds to a giant 3′ UTR contained in the fourth exon of MECP2. The long 3′ UTR and the newly identified first intron of MECP2/*Mecp2* are highly conserved in human and mouse. Furthermore, the human MECP2 locus is heterogeneous with respect to its DNA composition. We postulate that it represents a boundary between two H3 isochores that has not been observed previously.

Introduction

The nuclear methyl-CpG binding protein, MeCP2, was identified through its ability to recognize DNA that is methylated at CpGs (Lewis et al. 1992). In vertebrates, the majority of the CpGs are methylated, and appr. 90% of methylated CpGs occur in non-expressed or silenced sequences (Yoder et al. 1997). A minority of CpGs is clustered in 1- to 2-kb regions, named CpG islands (Bird 1992).

In vitro, MeCP2 is able to bind to a single methylated CpG. Cytological studies have shown that MeCP2 is an abundant component in the pericentromeric heterochromatin of mouse chromosomes. It is also found on the arms of mouse and rat chromosomes, yet in lower amounts (Lewis et al. 1992). Further studies indicated that MeCP2 can recognize methylated sequences in vivo (Nan et al. 1996). Interestingly, the chicken homolog of MeCP2 was identified through its ability to bind to matrix-attached sequences (MARs) and has thus been named attachment-region-binding protein, ARBP (von Kries et al. 1991; Weitzel et al. 1997). ARBP recognizes chicken repetitive sequences and mouse satellite DNA with the same high affinity as MARs. Besides a role in gene expression, ARBP is thought to function as a structuring protein at the level of higher chromatin order (Buhrmester et al. 1995; Weitzel et al. 1997).

Importantly, MeCP2 also possesses a transcription-repressing activity in the central portion of the protein (aa 207–310) (Nan et al. 1997). This activity was first detected in in vitro experiments and later shown to exist also in vivo. Protein–protein interaction studies indicate that MeCP2 interacts through the transrepressing domain with the corepressor mSin3A that is contained in a large corepressor complex including histone deacetylases HDAC1 and HDAC2. The transrepressing activity can be partially relieved by trichostatin A, a specific inhibitor of histone deacetylases. These results suggest a direct connection between DNA methylation and gene silencing through histone deacetylation (Nan et al. 1998; Jones et al. 1998).

cDNA sequences of MeCP2 have been reported from human, rat, chicken, and Xenopus (acc. no.: L37298, M94064, Y14166, AF051768). The expression pattern of the MeCP2 has been investigated using Northern analysis and Western blotting. In mouse, MeCP2 is present in somatic tissues, and at lower levels in embryonic tissues (Nan et al. 1996). In human, three transcripts of 1.8 kb, >9.5 kb, and >7.5 kb were observed in several adult somatic tissues (D'Esposito et al. 1996; Coy et al. 1999). The murine *Mecp2* gene was mapped between the *L1cam* and *Rsvp* genes in the central span of the mouse X-Chr by Quaderi et al. (1994), who localized the human MECP2 gene to the syntenically equivalent region in human Xq28, as confirmed by FISH (Vilain et al. 1996). Furthermore, it was shown that both the human and mouse MECP2 genes are subject to inactivation (Adler et al. 1995; D'Esposito et al. 1996), and that MeCP2 is essential for mouse development (Tate et al. 1996).

We have been engaged in a large-scale sequencing project on the human X Chr region q28, and its homologous region XA7-C on the mouse X Chr. The alignment of complete sequences of homologous human and murine genomic regions will not only reveal almost all of the exons, but is also effective at discovering novel regulatory elements (Hardison et al. 1997). By comparative sequence analysis of the entire MECP2 locus in human and mouse, we identified novel transcribed regions. The genomic data contribute to the knowledge on MeCP2 and provide a basis for studying the regulation of the MECP2 gene.

The nucleotide sequence data reported in this paper have been submitted to GenBank and have been assigned the accession numbers: AF030876, AF121351, AF158180, AF158181.

Correspondence to: M. Platzer, e-mail: mplatzer@imb-jena.de

Materials and methods

Genomic clones. Cosmid Qc8D3 is from QIZ-derived Xq27.3- Xqter library (Warren et al. 1990); cosmid I219D is from GM07297-F-derived LLOXNCOU1′U′ library (no. 110), and cosmid LC1837 is from GM1416B-derived ICRF library (no. 104) (Nizetic et al. 1991). PAC P671D9 was isolated from a human BAC/PAC library distributed by Research Genetics, Inc. (Huntsville, Ala.; Ioannou et al. 1994). BAC B228O4 is from a mouse BAC library generated at the Genome Research Laboratory at the California Institute of Technology and distributed by Research Genetics, Inc.

Sequencing and assembly. Cosmid, BAC, and PAC DNA preparation and shotgun sequencing were performed as described by Platzer et al. (1997). Sequence reactions were electrophoresed on ABI 377 sequencers. cDNA clones of identified ESTs were ordered at the Resource Center Berlin (http://www.rzpd.de) and sequenced with universal and/or custom-made primers. Sequence assembly and manual editing were performed with GAP4 (Dear and Staden 1991).

Sequence analysis. Homology searches in public domain databases were performed with BLAST, v.1.4 (Altschul et al. 1990) and FASTA, v.2.0 (Pearson, 1990). G+C content, and distribution of G+C were calculated with GCG, v.9.0 (Genetics Computer Group, Inc.); for graphical presentation, a sliding window of 6 kb with a step size of 1 kb was chosen. Criteria for identification of CpG islands were G+C% >50%, ratio of CpG observed/expected (o/e) >0.6, length >250 bp, CpG island finder (Larsen et al. 1992). Exon boundaries were identified by aligning mRNA sequence to genomic sequence with GAP (Huang, 1994). Genome-wide repeats were identified by RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/Repeat Masker). Potential transcription factor binding sites were identified by MatInspector (http://genomatix.gsf.de/cgi-bin/matinspector/matinspector.pl). 3′ Untranslated regions were searched for 3′ UTR specific motifs using UTRscan (http://bio-www.ba.cnr.it:8000/cgi-bin/BioWWW/UTRscanHTML.pl).

The entire human and mouse sequences were aligned by use of SIM96 with default parameters (http://globin.cse.psu.edu). Alignment results were plotted with APLOT (http://www1.imim.es/~jabril/gff2aplot.tgz). The numerical values of homology *h* in exons and introns were determined from the local alignments produced by SIM96 according to:

$$h = \frac{\Sigma\, m(i)}{(n_1 + n_2)/2}$$

where $m(i)$ is the number of exact matches in an alignment i, n_1, and n_2 are the lengths of the considered region in the human and mouse sequence, respectively, and the sum is over all local alignments within the region of interest. The sequence identities of the long 3′ UTR of MECP2/*Mecp2* and of another 55 3′ UTRs extracted from a set of human/mouse and human/rat homologous genes compiled at http://www.ncbi.nlm.nih.gov/Makalowski/PNAS (Makalowski et al. 1996) were calculated as described above. For comparison of the obtained values, match profiles were calculated. A match profile is obtained when the length of all aligned fragments that fall into a certain identity category (e.g., identities between 95% and 100%) are summed, and the obtained sum is divided by the total sequence length. We plotted (i) the fraction of sequence length (taken with respect to the total 3′ UTR length) of the conserved fragments against their sequence identities and (ii) the average length of matching fragments within the 3′ UTRs against their sequence identities.

Cell culture. Human Raji cells and Jurkat cells were grown in RPMI 1640 medium supplemented with 10% fetal calf serum (Boehringer Mannheim), 100 units/ml penicillin, and 100 μg/ml streptomycin.

Cloning of a human MECP2 cDNA. Poly(A)$^+$ RNA from human placenta (Clontech) was reverse transcribed with an oligo(dT) primer and Expand Reverse Transcriptase (Boehringer Mannheim). Double-stranded cDNA was synthesized, and a cDNA adaptor was ligated by use of the Marathon cDNA Amplification Kit (Clontech). Then, a 1181-bp fragment was amplified by PCR with adaptor primer 1 and an internal MECP2-specific primer (5′-TGGAGTTGATTGCGTACTTCG-3′, acc. no.: L37298). The 5′ end of the cDNA was obtained by PCR using a primer from the 5′ region of the 1181-bp fragment (5′-TTGATGTGACATGT-GACTCCC-3′) and a primer (5′-CCTCTCCCAGTTACCGTGAA-3′) derived from the previously reported 5′ end of the cDNA (acc. no.: L37298). Both PCR products were cloned into the vector pCR 2.1 (Invitrogen) and fused at a common *Bst*BI site. The nucleotide sequence of the MECP2 cDNA was deposited in GenBank/EMBL under acc. no. Y12623.

mRNA expression analyses. Multiple human tissue Northern blots were obtained from Clontech. A human MECP2 cDNA probe was prepared with [α-^{32}P]dATP by random prime labeling (Boehringer Mannheim). Prehybridization, hybridization, and washing were performed under high stringency conditions according to the manufacturer. To control for the relative amount of RNA in each lane, after hybridization with MECP2 cDNA, the blots were stripped by incubation in 0.5% (wt/vol) SDS at 95°C for 10 min and reprobed with a human GADPH cDNA.

To analyze the splicing and polyadenylation pattern of the human MECP2 mRNAs, total RNA was extracted from Jurkat cells by the method of Chirgwin et al. (1979). In several lanes, 30 μg of RNA was electrophoresed on formaldehyde gels (Sambrook et al. 1989). Lanes were blotted and then individually hybridized with probes specific for the following exon and intron sequences of human MECP2: exon 2, 5′ end of intron 2, 3′ end of intron 2, complete exon 3, complete intron 3, 5′ end of exon 4, long 3′ UTR probe (derived from EST R125739).

Determination of the MECP2 mRNA half-life in Raji cells. Raji cells were treated with 5 μg/ml of actinomycin D. Poly(A)$^+$ RNA was extracted from appr. 2×10^7 cells as described by Rahmsdorf et al. (1987) at the time of actinomycin D addition and 1.5, 2.5, 4.5, and 6.5 h later. RNA samples were run on formaldehyde gels and blotted (Sambrook et al. 1989). Blots were first hybridized with the MECP2 cDNA probe and then rehybridized with a human GAPDH probe. Autoradiograms were scanned with a Bio-Rad model 620 video densitometer. The signals obtained with a MECP2 probe were normalized for those obtained with the GAPDH probe, taking into account a previous determination of 8 h for the half-life of GAPDH mRNA (Dani et al. 1984).

Results

Sequencing between IRAK/Il1rak and RCP/Rsvp in human and mouse. We have previously sequenced a 400-kb cosmid contig spanning the region between the creatine transporter gene, CRTR, and the gene encoding the interleukin-1 receptor associated kinase, IRAK, in human Xq28 (acc. nos.: U52111 and U52112, from centromere to telomere). Telomeric to U52112, there is a 50-kb cosmid contig (GenBank entries Z47046, Z47066, and Z68193, sequenced at the Sanger Centre, Hinxton, UK), which contains the red opsin gene, RCP (Nathans et al., 1986; Vollrath et al. 1988). Between U52112 and this cosmid contig remained a gap of unknown length, likely to contain the gene for methyl-CpG binding protein 2, MECP2 (Quaderi et al. 1994).

To close this gap, we generated sequence-specific probes from the distal end of U52112 and screened genomic libraries. Initially, we identified one cosmid, Qc8D3. It overlaps 13 kb with U52112 and extends 31 kb into the gap. In a second screening, we used probes corresponding to the distal end of Qc8D3. We identified cosmids I219D, LC1837, and PAC P671D9. P671D9 turned out to overlap both with Qc8D3 (31 kb) and Z47046 (11 kb). It thus bridges the gap between U52112 and Z47046; the size of the gap is 89 kb (Fig. 1). To rule out a deletion in P671D9, we performed Fiber FISH analysis, which confirmed the integrity of the PAC (data not shown). Together, clones Qc8D3, I219D, LC1837, and P671D9 assemble into 112.756 kb of contiguous sequence (acc. no.: AF030876). Including the 50-kb cosmid contig mentioned above, a 151.676-kb contig is formed.

To compare the human region with its murine counterpart, mouse BAC B228O4 was isolated with a murine EST homologous to rat MeCP2 (acc. no.: W10855). The BAC was completely sequenced and assembled to a 123.192-kb contig (acc. no.: AF121351, Fig. 1).

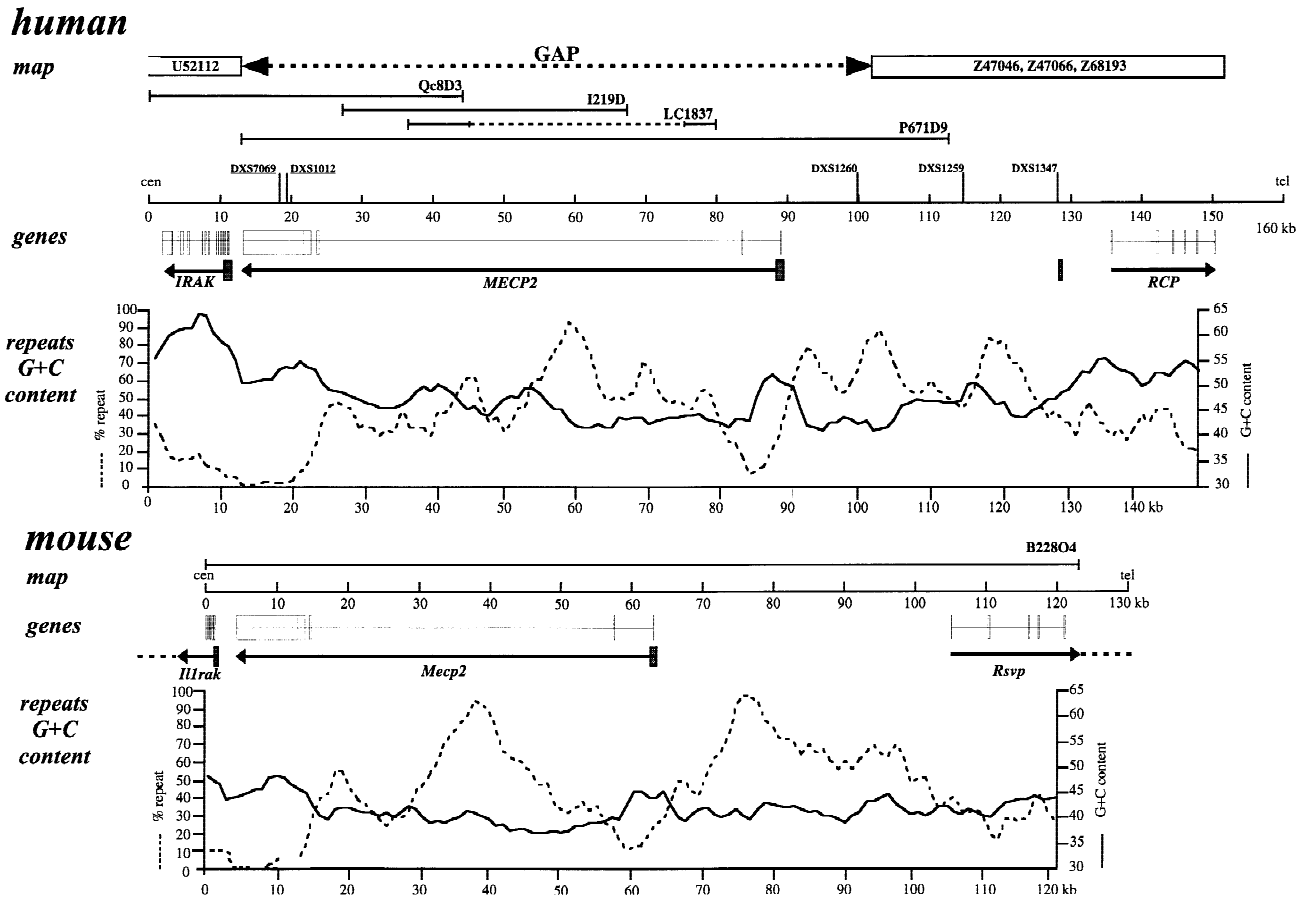


Fig. 1. Comparison of the human and murine genomic regions. **Upper panel:** The 152-kb region on human Chr Xq28. **Map:** The top line depicts previously sequenced contigs U52112 and Z47046, Z47066, Z68193, and the gap at the starting point of our work. Below, sequenced clones are shown. Cosmid Qc8D3 extends U52112 by 31 kb. Cosmid I219D extends Qc8D3 by 23 kb. Cosmid LC1837 bears a 31-kb deletion at its center (dotted line). PAC P671D9 overlaps with Qc8D3 (31 kb) and 47046 (11 kb) and bridges the gap of 89 kb. STS markers are indicated by vertical lines. **Genes:** Identified genes are marked by arrows indicating the direction of transcription, exons are drawn as white boxes, and predicted CpG islands as drawn as black boxes. **Repeats and G+C content:** Repeats are depicted by dotted lines and G+C content by solid lines. Drawings are to scale. **Bottom panel:** The homologous region of 123 kb on mouse Chr XC is encompassed by BAC 228O4.

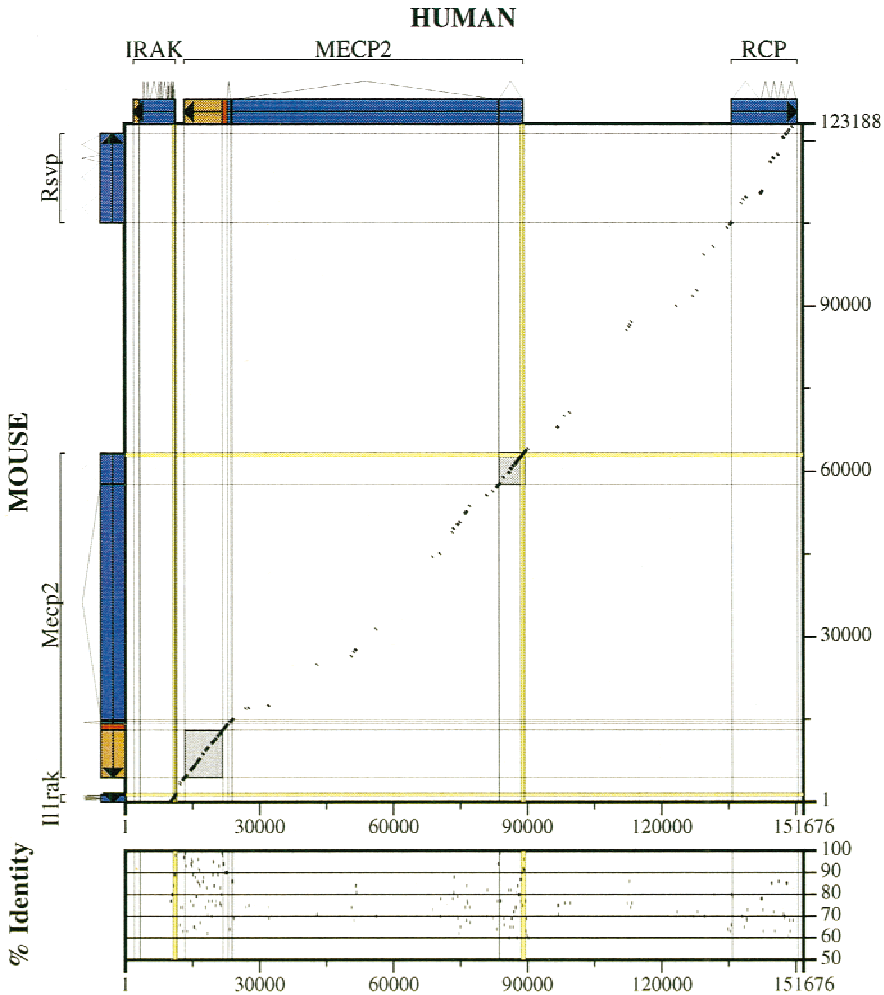
Comparative analysis of the human and mouse loci. Subsequently, we compared the 152 kb of human genomic sequence with 123 kb of homologous murine genomic sequence. We identified three genes in both species: IRAK, MECP2, RCP and their murine homologs *Il1rak*, *Mecp2*, and *Rsvp*, from centromere to telomere (Fig. 1).

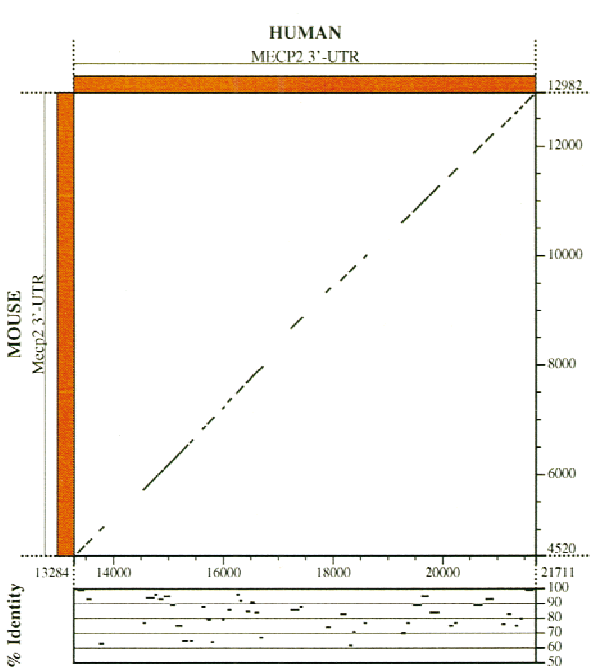
An overall comparison of both genomic sequences represented by similarity and percentage identity plots is given in Fig. 2. In both species, IRAK/*Il1rak* and MECP2/*Mecp2* are transcribed from telomere to centromere, whereas RCP/*Rsvp* is transcribed from centromere to telomere. Exon/intron organization, exon sizes, and coding regions are highly conserved (>80%). Intron sizes are conserved to a lesser extent. In the mouse, the sum of all intron lengths is smaller than in the human (64,282 bp versus 77,442 bp, respectively). In addition, part of the non-coding sequence, that is, regions flanking individual exons are well conserved (http://genome.imb-jena.de/~kathrinr/tbl3.html). Exon/intron boundaries follow the GT/AG rule and are identical in both species.

The average G+C content of the human 152 kb is 49%, while it is 43% in the murine 123 kb. Local analysis indicates three compositionally distinct regions in the human sequence (Fig. 1). The proximal and distal portions exhibit a G+C content of 56% (24 kb) and 53% (16 kb), but the G+C content drops to 46% in the central region (111 kb). Similarly, we observe differences in the G+C content in the mouse sequence. It is 48% in the proximal 15 kb and drops to 41% in the central portion extending over 90 kb. In contrast to the human region, the G+C content remains at 42% in the distal portion (18 kb).

A considerable fraction of both the human and the murine sequences is composed of repeats, representing 47% in the human and 43% in the mouse. In the human, most of the repetitive elements are SINEs and LINEs (22% and 18%, respectively). A total of 123 Alu repeats is present in the human sequence, that is, 1 Alu

Fig. 2. Comparison of the human and murine genomic regions by similarity (SIM) and percentage identity (PIP) plot analysis. **(top)** Overview of the region from IRAK/*Il1rak* to RCP/*Rsvp.* The human and murine genomic sequences are represented on horizontal and vertical lines, respectively. The conservation between both sequences is represented by diagonal lines in the SIM plot. Conserved blocks (CEs) of ≥50 bp, being ≥60% identical and not contained in repetitive elements, are shown. The respective percentage identities of the CEs are drawn as black bars in the PIP plot below each SIM plot. Coding regions of genes are depicted as blue boxes; arrows indicate the direction of transcription. The long 3′ UTR of MECP2/*Mecp2* and the 3′ UTR of IRAK are drawn as red boxes. CpG islands are represented by yellow ribbons. **(bottom)** Enlargement of the 3′ UTR and the 5′ region including intron 1 of MECP2/*Mecp2.*


HUMAN
IRAK
MECP2
RCP
123188
90000
60000
30000
1
MOUSE
Rsvp
Mecp2
Il1rak
1
30000
60000
90000
120000
151676
% Identity
100
90
80
70
60
50


HUMAN
MECP2 3'-UTR
12982
12000
10000
8000
6000
4520
MOUSE
Mecp2 3'-UTR
13284
14000
16000
18000
20000
21711
% Identity
100
90
80
70
60
50

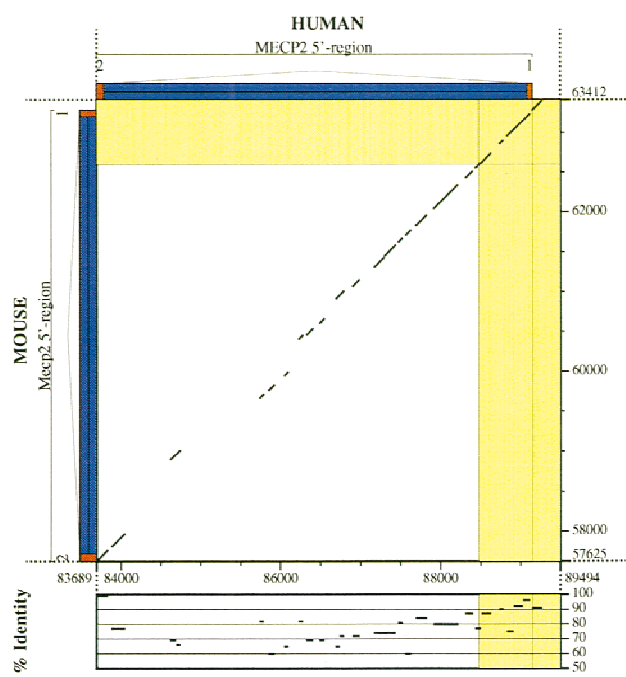

HUMAN
MECP2 5'-region
2
1
63412
62000
60000
58000
57625
MOUSE
Mecp2 5'-region
1
83689
84000
86000
88000
89494
% Identity
100
90
80
70
60
50

Table 1. Comparison of human and mouse MECP2 genes.

Exon				Intron				
no.	human (bp)	mouse (bp)	match (%)	no.	human (bp)	mouse (bp)	match (%)	phase
1 noncoding	69	84	90.2					
				1	5259	5476	64.6	—
2 noncoding	98	98	100.0					
2 coding	26	26	100.0					
				2	59633	4273	12.3	1
3 coding	351	351	88.0					
				3	756	487	34.5	1
4 coding	1084	1078	91.1					
4 noncoding short	127	127	98.4					
4 noncoding long	8555	8590	73.8					
short transcript	1755	1764	94.6					
long transcript	10310	10354	91.6					

per 1.24 kb, which is fourfold the estimated genome average (Deininger 1989). In the mouse, SINEs, LINEs, and LTR elements are almost equally represented (14%, 12%, and 13%, respectively). Specifically, SINEs are present at a density of 1 repeat per kb, which is 20-fold the estimated genome average (Deininger 1989). Furthermore, the repeat content shows drastic local differences. It is high (58%) in the second large intron of MECP2/*Mecp2* and in the intragenic region between MECP2/*Mecp2* and RCP/*Rsvp.* In the remaining portions of MECP2/*Mecp2* and in IRAK/*Il1rak* the repeat content is very low (14%).

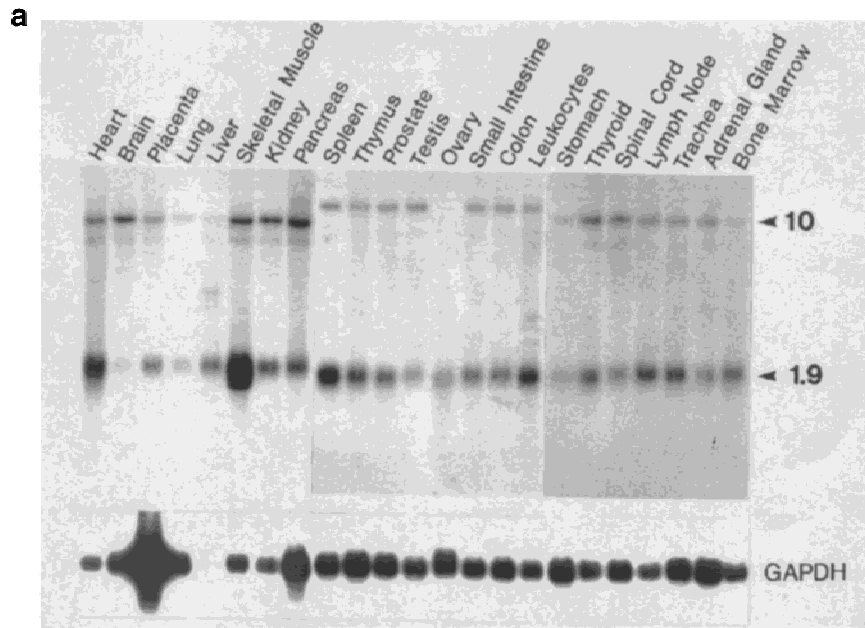
CpG islands are associated with the 5′ ends of both the IRAK/*Il1rak* and the MECP2/*Mecp2* gene. The island (77% GC-rich) found at IRAK is 1034 bp long and covers 314 bp of 5′ flanking DNA, exons 1 to 3, and 26 bp of intron 3. In the mouse, the corresponding island (71% GC-rich) is 598 bp. It covers 209 bp of 5′ flanking DNA, exons 1 and 2, 7 bp of intron 2, and it is 89% identical with the corresponding section of the human island. The islands associated with MECP2 and *Mecp2* are 1020 bp and 881 bp in length, respectively. Both islands are 88% conserved over the entire length of the murine island.

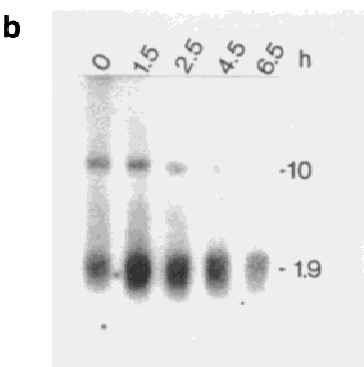
Comparison of the MECP2 gene in human and mouse reveals a new 5′ exon and an alternative polyadenylation site. Presently, GenBank contains four mRNA entries of human MECP2 (acc. nos. L37298, X89430, X99686, Y12623). These mRNAs harbor three exons spanning 62 kb and 45 kb of genomic DNA in human and mouse, respectively. By comparative analysis of human and murine genomic sequences, we identified upstream of the first of these exons a region of 6 kb, which contains blocks of well-conserved sequences, including a highly conserved CpG island. We assumed that this island is associated with the 5′ end of MECP2/*Mecp2.* This hypothesis was confirmed by a murine EST (acc. no.: AI181668), which maps in the murine CpG island and defines a new, non-coding 5′ exon of 84 bp. In this EST the sequence of the new exon (exon 1 in Fig. 1) immediately abuts that of the previously known most upstream exon (exon 2), thus defining an intron 5.5 kb in length. Furthermore, AI181668 overlaps by 336 nt with the human mRNA (acc. no. L37298) and extends it by 97 nt at its 5′ UTR, which defines a new, non-coding 5′ exon of 69 bp, and an intron of 5.3 kb. Thus, the four exons of the human MECP2 mRNA have a length of 1755 nt (acc. no. AF158180), while the murine homolog is 1764 nt (acc. no. AF158181, Table 1). Accordingly, the CpG islands in human and mouse cover 344 bp and 130 bp, respectively, of 5′ flanking DNA, exon 1, and 607 bp and 537 bp, respectively, of intron 1.

In the fourth exon of MECP2/*Mecp2,* there is a conserved, previously known polyadenylation signal 102 nt downstream of the stop codon. In addition to the mRNA database entries mentioned above, there are two ESTs in GenBank, in which that polyadenylation signal is used (acc. nos.: AA279313, AA543089). The corresponding 3′ UTR has a length of 127 bp and exhibits 98% identity between human and mouse. Interestingly, this polyadenylation signal is followed by a region of 8.5 kb, which exhibits 74% identity between human and mouse (Fig. 2). This sequence terminates in two additional, non-canonical polyadenylation signals (underlined) (UAUAAAGAGUUUGCCUUAUAAAUUU**U**$^{\Downarrow}$A*CA*). A database search identified 44 human, 16 mouse, and 4 rat ESTs that match within the 8.5-kb region. These ESTs are mostly expressed in brain, lung, and breast. In four human ESTs (acc. nos. R40020, AA670387, R43025, N25284) the sequence up to the U (bold face) is followed by a poly(A) tail. This indicates that cleavage of the primary transcript occurs at the site marked by the arrow, but not at the canonical poly(A) site CA (italic). Collectively, these ESTs suggest that the 8.5-kb region represents an alternative, unusually long 3′ UTR of MECP2/*Mecp2.*

Expression and half-life of the alternative MECP2 transcripts. To investigate whether the putative, unusually long 3′ UTR is related to MECP2/*Mecp2,* we performed Northern blot analysis with probes from various regions of the human MECP2 gene. A cDNA probe (acc. no. Y12643) as well as probes specific for exon 2, exon 3, or the translated portion of exon 4 simultaneously detected two transcripts of 1.9 kb and ~10 kb (Fig. 3a and data not shown). Both transcripts are ubiquitously expressed but show different tissue expression patterns; for example, the shorter transcript is barely detectable in brain and lung, but very abundant in heart, skeletal muscle, and spleen. On the other hand, the ~10-kb transcript is barely detectable in lung and liver, and absent in ovary, but relatively abundant in skeletal muscle, kidney, pancreas, and brain. Furthermore, there is a third, weak transcript of >7.5 kb. It is mainly expressed in heart, brain, skeletal muscle, and pancreas, but also present in other tissues. A probe from the putative long 3′ UTR exclusively detected the ~10-kb transcript (data not shown). These results are in agreement with those of D'Esposito et al. (1996), but contrast with those of Coy et al. (1999; see Discussion section). The intron probes (see methods) hybridized with neither transcript (data not shown). This indicates that the 1.9-kb as well as the ~10-kb transcript is generated by alternative usage of two poly(A) addition sites in the MECP2/*Mecp2* mRNA. Therefore, the ~10 kb transcript harbors a long 3′ UTR encompassing 8555 bp (acc. no. AF158180) in the human and 8590 bp in the mouse (acc. no. AF158181). Thus, the four exons of MECP2/*Mecp2* span, in fact, 76 kb in human and 59 kb in mouse.

We then determined the half-life of the human MECP2 mRNA by Northern blot analysis after inhibition of RNA synthesis with actinomycin D (Fig. 3b). Both the long and the short transcript are





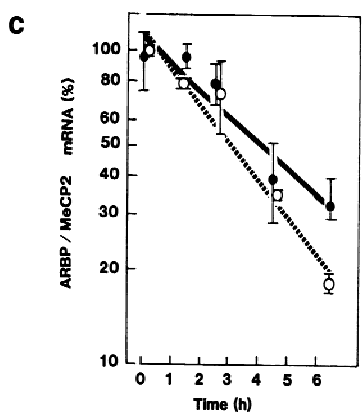


Fig. 3. Expression and half-life of the human MECP2 mRNA. **(a)** Human tissue Northern blots were hybridized using a labeled human MECP2 probe (acc. no. Y12643) and then rehybridized with a labeled GAPDH probe (bottom). Sizes of MECP2 mRNAs are indicated in kb. **(b, c)** Half-life of the MECP2 mRNA in Raji cells. Closed circles and the continuous line show the data for the 1.9-kb transcript, open circles and the dotted line the data for the ~10-kb transcript. Vertical bars indicate standard deviations.

degraded rapidly in Raji cells. For the 1.9 kb transcript we observed a half-life of ~4 h, while the ~10-kb transcript has a half-life of ~3 h (Fig. 3c).

Not only in human, but also in rat and mouse, a short (1.9-kb) and a long (~10-kb) MECP2 transcript are present (data not shown). Both show tissue-specific expression patterns; for example, the long transcript is expressed at relatively high levels in brain, while the short transcript is barely visible in this tissue. We also performed Northern blot analysis, using total RNA preparations from chicken embryos, various chicken tissues, and cell lines (data not shown). We detected two short transcripts of 1.8 kb and ~2.3 kb, both exhibiting tissue-specific expression patterns. Their expression level is low in brain, but high in spleen, a feature also observed for the short transcript in mammals. This indicates that a tissue-specific expression is a cross-species feature of the MECP2 gene.

The long 3′ UTR of MECP2/Mecp2 is well conserved. The unusual length (8.5 kb) and high degree of sequence conservation (74%) of the MECP2/*Mecp2* 3′ UTR prompted us to compare it with other orthologous 3′ UTR pairs. From a table of GenBank accession numbers (Makalowski and Boguski 1998), we chose 24 human/mouse and 31 human/rat homologs (i) that share 92–100% similarity at amino acid level and (ii) whose 3′ UTRs are longer than 1 kb. We calculated the sequence identity of these 55 homologous 3′ UTRs and compared the values with the long 3′ UTR of MECP2/*Mecp2*. When the fraction of sequence length is plotted against identities, the long MECP2/*Mecp2* 3′ UTR fits well the human/mouse reference set (Fig. 4a). In a *t*-test for paired differences, the null-hypothesis („no difference“) is not rejected ($p = 0.9963$). However, it turns out that the long 3′ UTR of MECP2/*Mecp2* is unusual in at least two respects. First, it is much longer than 3′ UTRs in any other of the human/rodent homologous genes known so far (Makalowski et al. 1998); the longest 3′ UTR (4.160 kb) occurs in the gene encoding the mouse zinc finger protein Zfx (Mardon et al. 1990). Second, matching fragments within the long MECP2/*Mecp2* 3′ UTR are on average 14 bp longer than in the reference set (Fig. 4b, *t*-test for paired differences yields $p^* = 0.0429$). The unusual conservation of the long MECP2/*Mecp2* 3′ UTR is also reflected in the values for the category „unaligned“ (Fig. 4a), For the average human/mouse 3′ UTR pairs, 14% of sequence is not aligned at all, however only 6% fall into this category for the long 3′ UTR of MECP2/*Mecp2*.

To identify functional sequences that may act as regulatory elements, we searched both the long and the short 3′ UTR of MECP2/*Mecp2*, using UTRscan. Both 3′ UTRs do not contain

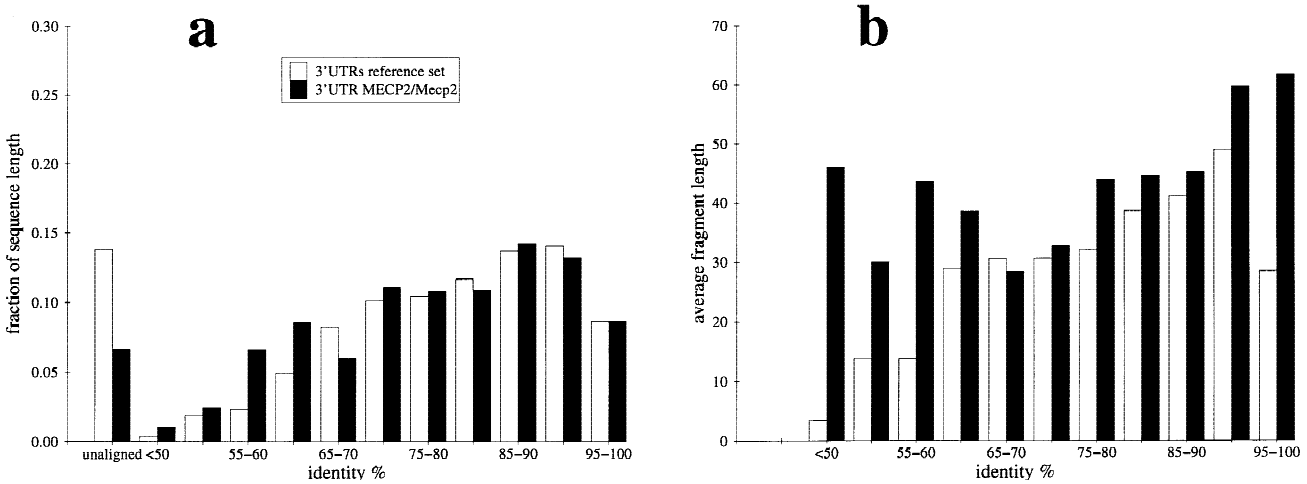


Fig. 4. Match profile of the long MECP2/*Mecp2* 3′ UTR compared with a reference set of homologous human/mouse 3′ UTR pairs. Identity categories are plotted against **(a)** the fraction of sequence length and **(b)** the average fragment length.

iron-responsive elements, 3′ UTR stem-loop structures, or any other motif compiled in UTRdb. With respect to the short half-life of the MECP2 transcripts (see above), we note that the short and the long 3′ UTR do contain AU-rich elements, the most common determinant of RNA stability in mammalian cells (Chen and Shyu 1995).

Conserved elements in the MECP2/Mecp2 gene. There are 170 conserved elements (CEs) within the 152 kb of human and 123 kb of murine genomic sequence. These CEs were obtained by extraction those locally aligned fragments which are at least 50 bp long, have an identity of at least 60%, and are not contained within repetitive elements. Of the 170 conserved fragments, 114 CEs are located in the region occupied by MECP2/*Mecp2,* which corresponds to 16% of the locus in both species (http://genome.imb-jena.de/kathrinr/tbl3.html). Five of these 114 CEs represent exons 1–3 and the coding portion of exon 4, while 109 CEs are found in untranslated regions and intronic sequences.

Intron 1 of MECP2/*Mecp2* is 5.3 kb in human versus 5.5 kb in mouse (Table 1). The intron exhibits an overall identity of 65% between both species. There are 24 CEs within the intron, which are on average 93 bp long and exhibit 87% identity. Together, they comprise 42% of intron 1. Intron 2 spans 60 kb in the human and 43 kb in the mouse. In this intron, the overall sequence conservation is 12%. 37 CEs are observed. These are on average 73 bp long, that is, they are shorter than the CEs in intron 1, but exhibit with 86% identity a similar degree of sequence conservation. Together, these CEs represent 5% of intron 2. Six CEs exhibit a sequence identity above 84%, which was reported as the average identity of human/mouse coding sequences (Makalowski et al. 1996). Intron 3 is 756 bp in human and 487 bp in mouse, and exhibits an overall identity of 35%. There is one conserved element in this intron. It is 66 bp long, exhibits 64% identity between both species, and represents 10% of intron 3. The potential regulatory role of these CEs remains to be investigated.

Discussion

Chromosomal band Xq28 is a terminal reverse band and corresponds to approximately 9 Mb of DNA. Compositionally, a proximal, middle, and distal region are distinguished, which consist of L/H1, H2/H3, and L isochores, respectively (De Sario et al. 1996). The MECP2 locus is located in the middle portion of Xq28, which extends from GABRA to G6PD and has been described to be extremely heterogeneous in base composition. Particularly, this applies to the 152-kb region described here. While its proximal and distal portions exhibit average G+C contents close to neighboring L1CAM and G6PD contigs (56% and 55%, respectively), the G+C content drops to 46% in the middle portion of 111 kb. These compositional differences suggest that the described region may comprise a boundary between two H3 isochores located around the L1CAM and G6PD loci (De Sario et al. 1996). Accordingly, the four exons of MECP2 span 73 kb, which contrasts sharply with the average gene density in the L1CAM and G6PD contigs of one gene per 22 kb and 12 kb, respectively (acc. no. U52112; Chen et al. 1996).

The DNA composition in mouse has been described as less heterogeneous than in human (Cuny et al. 1981), and the genome of rodents underwent secondary changes leading to narrower compositional distributions than in human (Bernardi 1995). Correspondingly, we find that the G+C content of the 123 kb of murine genomic DNA varies only moderately. Furthermore, the G+C content is 48% in the proximal portion in mouse compared with 56% in human, which represents a transition from an H3 to an H2 isochore between both species. Even more, in the distal portion the G+C content is 42% in mouse versus 56% in human, representing a change from an H3 to an H1 isochore. Also, it was shown that CpG islands associated with mouse genes have lower CpG and/or GC levels than their human homologs (Aïssani and Bernardi 1991) and that typical CpG island features are severely depleted in the mouse (Matsuo et al. 1993). Corresponding to these findings, the CpG islands associated with human MECP2 and IRAK are larger and more GC-rich than their murine counterparts.

According to the mRNA sequence information available from GenBank (acc. no. L37298), the human MECP2 gene seems to consist of three exons. The present comparative analysis of human and murine genomic DNA combined with a Northern blot analysis revealed that neither the 5′ end nor the 3′ end of the gene was included in the reported exon sequences. We identified a new untranslated exon appr. 5.4 kb upstream of the previously reported most upstream exon. Since this exon is embedded in a typical CpG island and since ubiquitously expressed (housekeeping) genes routinely have CpG islands associated with their 5′ ends (Antequera and Bird 1993), we assume that the novel exon is the first one. The simultaneously discovered novel first intron is distinguished by a surprisingly high degree of sequence conservation between human and mouse. Sequence comparisons 3′ to the previously reported polyadenylation site with the SIM algorithm revealed an extremely well-conserved (74%) region of 8.5 kb. Thus, the last (fourth) exon of MECP2 harbors two alternatively used polyadenylation sites, located 127 bp and 8555 bp downstream of the stop codon, re-

spectively. We have to infer that this exon contains an unusually long 3′ UTR, which seems to be longest among the 3′ UTRs deposited in GenBank. This alternative 3′ UTR was recently found independently by Coy et al. (1999). Our Northern blot analysis showed that in human (and other mammals) MECP2 gives rise to two prominent transcripts, of 1.9 kb and appr. 10 kb, and a weak transcript of >7.5 kb, which is in good accordance with D'Esposito et al. (1996). In contrast, Coy et al. (1999) reported that in adult tissues the most abundant transcripts detectable by a coding region probe were 5 kb and less than 1.35 kb in length. We did not detect these transcripts. Moreover, we are confident that the transcripts detectable in our Northern blots and in those of D'Esposito et al. (1996) are indeed the most abundant ones, since we obtained consistent results with different fragments of the cDNA as probe under high-stringency conditions.

The extreme length of the 3′ UTR raises the question of conservation of long untranslated regions between different species. Clearly, the alternative long MECP2/*Mecp2* 3′ UTR exhibits a similar degree of sequence conservation as we calculated for a reference set of human/rodent 3′ UTRs. This is in disagreement with Coy et al. (1999), who report a sequence conservation (52% with BESTFIT) below the average of 69.1% for 1196 human/mouse 3′ UTRs estimated by Makalowski et al. (1996). However, if one calculates the conservation of the long 3′ UTR of MECP2/*Mecp2* with the GAP algorithm as Makalowski et al. (1996) did, the identity is 68% between human and mouse, which fits the average.

The expression of the MECP2 transcripts shows conserved tissue-specific differences; for example, the ~10-kb transcript is prominent in brain, while the 1.9-kb transcript prevails in heart and skeletal muscle. The transcripts derive from alternatively used polyadenylation signals. It is not clear why the frequency at which these signals are used varies in a tissue-specific manner. Furthermore, it is unclear whether the two transcripts play different or similar physiological roles. We could show that both transcripts have almost identical short half-lifes. Yet it is possible that they differ in their translatability; there are cases in which either the shorter or the longer transcript is translated more efficiently (Edwalds-Gilbert et al. 1999). Also, the high sequence conservation of the long 3′ UTR between human and mouse suggests an important function of the ~10-kb transcript. Expression of a significant number of other genes is known to be regulated at the level of polyadenylation. However, also in these cases little is known about the molecular mechanisms (Edwalds-Gilbert et al. 1997). It has been proposed that usage of an immunoglobulin alternative poly(A) site is developmentally regulated through differential binding of a basal polyadenylation factor, CstF 64-KDa protein (Edwalds-Gilbert and Milcarek 1995; Takagaki et al. 1996). Furthermore, interaction of two splicing factors, U1 snRNP A and 70K protein, with the basal polyadenylation machinery has been implicated in alternative polyadenylation (Gunderson et al. 1994; Lutz and Alwine 1994). Recently, a basal polyadenylation factor, cleavage factor I, has been shown to possess a domain containing SR (serine-arginine) dipeptides (Ruegsegger et al. 1998). Thus, classical and cell-type-specific SR proteins bound to RNA might interact with cleavage factor I and thus might be implicated in the tissue-specific recognition events of the two polyadenylation signals of the MECP2/*Mecp2* mRNA.

The CpG island at the 5′ end of MECP2 most likely includes the promoter, which contains a number of Sp1 binding sites. Interestingly, homozygous Sp1 knockout mice who die at day 11 show greatly reduced expression of MeCP2 (Marin et al. 1997). MeCP2 performs structural roles in the nucleus and furthermore acts as a transcriptional repressor (Weitzel et al. 1997; Nan et al. 1998; Jones et al. 1998). Therefore, it has been hypothesized that the deficiency of MeCP2 in Sp1 knockout mice contributes significantly to the phenotype of the embryos (Marin et al. 1997).

MeCP2 is the first member of a family of proteins whose common feature is a highly conserved methyl-CpG binding domain (MBD-family; Hendrich and Bird 1998). Interesting novel members of this family are a DNA demethylase (Bhattacharya et al. 1999) and a methyl-CpG binding endonuclease probably involved in DNA mismatch repair (Bellacosa et al. 1999). For all members of the family, the genomic sequence of the methyl-CpG-binding domain is interrupted by an intron at an equivalent site (Hendrich and Bird 1998). Thus, the family members also share similarities in their genomic organization. The complete elucidation of the organization of one family member, MECP2/*Mecp2*, therefore provides a valuable basis for further studies at the genomic level.

Note added in proof. Recently, mutations in the MECP2 gene have been identified as the cause of RRH syndrome (Anir et al., Nat Genet 23, 185–188, 1999)

Acknowledgments. Part of this article is based on the doctoral thesis of K. Reichwald in the Faculty of Biology at the Friedrich Schiller University, Jena, and on the doctoral thesis of J. Thiesen in the Faculty of Biology, University of Hamburg. We thank Diana Wiedemann and Hella Ludewig for expert technical assistance. The sequence data having accession numbers Z47046, Z47066, Z68193 were generated by David Buck of the Human Chromosome-X Sequencing Group at the Sanger Centre. This work was supported by grants to W.H. Strätling from the Deutsche Forschungsgemeinschaft; to A. Rosenthal from the German BMBF (BEO 0311108/0) and from the European Commission (BMH4-CT96-0338); and to M. Platzer from the Deutsche Forchungsgemeinschaft (Pl 173/2-1).

References

Adler DA, Quaderi NA, Brown SD, Chapman VM, Moore J et al. (1995). The X-linked methylated DNA binding protein, Mecp2, is subject to X inactivation in the mouse. Mamm Genome 6, 491–492

Aissani B, Bernardi G (1991) CpG islands, genes and isochores in the genomes of vertebrates. Gene 106, 185–195

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215, 403–410

Antequera F, Bird A (1993) CpG islands. Exper Suppl (Basel) 64, 169–85

Bellacosa A, Cicchillitti L, Schepis F, Riccio A, Yeung AT et al. (1999) MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1. Proc Natl Acad USA 96, 3969–3974

Bernardi G (1995) The human genome: organization and evolutionary history. Annu Rev Genet 29, 445–476

Bhattacharya SK, Ramchandani S, Cervoni N, Szyf M (1999) A mammalian protein with specific demethylase activity for mCpG DNA. Nature 397, 579–583

Bird A (1992) The essentials of DNA methylation. Cell 70, 5–8

Buhrmester H, von Kries JP, Strätling WH (1995) Nuclear matrix protein ARBP recognizes a novel DNA sequence motif with high affinity. Biochemistry 34, 4108–4117

Chen CY, Shyu AB (1995) AU-rich elements: Characterization and importance in mRNA degradation. Trends Biochem Sci 20, 465–470

Chen EY, Zollo M, Mazzarella R, Ciccodicola A, Chen CN et al. (1996) Long-range sequence analysis in Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/CGP and G6PD loci. Hum Mol Genet 5, 659–668

Chirgwin JM, Przybyla AE, MacDonald RJ, Rutter WJ (1979) Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry 18, 5294–5299

Coy JF, Sedlacek Z, Bachner D, Delius H, Poustka A (1999) A complex pattern of evolutionary conservation and alternative polyadenylation within the long 3″-untranslated region of the methyl-CPG-binding protein 2 gene (McCP2) suggests a regulatory role in gene expression. Hum Mol Genet 8, 1253–1261

Cuny G, Soriano P, Macaya G, Bernardi G (1981) The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. Eur J Biochem 115, 227–233

Dani C, Piechaczyk M, Audigier Y, El Sabouty S, Cathala G et al. (1984) Characterization of the transcription products of glyceraldehyde 3-phosphate-dehydrogenase gene in HeLa cells. Eur J Biochem 145, 299–304

Dear S, Staden R (1991) A sequence assembly and editing program for

efficient management of large projects. Nucleic Acids Res 19, 3907–3911

Deininger PL (1989) SINEs: Short interspersed repeated DNA elements in higher eukaryotes. In *Mobile DNA*, DE Berg, MM Howe (eds) (Washington, D.C.: American Society for Microbiology), pp 619–636

De Sario A, Geigl EM, Palmieri G, D'Urso M, Bernardi G (1996) A compositional map of human chromosome band Xq28. Proc Natl Acad Sci USA 93, 1298–1302

D'Esposito M, Quaderi NA, Ciccodicola A, Bruni P, Esposito T, et al. (1996) Isolation, physical mapping, and Northern analysis of the X-linked human gene encoding methyl CpG-binding protein, MECP2. Mamm Genome 7, 533–535

Edwalds-Gilbert G, Milcarek C (1995) Regulation of poly(A) site use during mouse B-cell development involves a change in the binding of a general polyadenylation factor in a B-cell stage-specific manner. Mol Cell Biol 15, 6420–6429

Edwalds-Gilbert G, Veraldi KL, Milcarek C (1997) Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res 25, 2547–2561

Gunderson SI, Beyer K, Martin G, Keller W, Boelens WC et al. (1994) The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. Cell 76, 531–541

Hardison RC, Oeltjen J, Miller W (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res 7, 959–966

Hendrich B, Bird A (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol Cell Biol 18, 6538–6547

Huang X (1994) On global sequence alignment. Comput Appl Biosci 10, 227–235

Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H et al. (1994) A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nat Genet 6, 84–89

Jones PL, Veenstra GJ, Wade PA, Vermaak D, Kass SU, et al. (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet 19, 187–191

Larsen F, Gundersen G, Lopez R, Prydz H (1992) CpG islands as gene markers in the human genome. Genomics 13, 1095–1107

Lewis JD, Meehan RR, Henzel WJ, Maurer-Fogy I, Jeppesen P et al. (1992) Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell 69, 905–914

Lutz CS, Alwine JC (1994) Direct interaction of the U1 snRNP-A protein with the upstream efficiency element of the SV40 late polyadenylation signal. Genes Dev 8, 576–586

Makalowski W, Boguski MS (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci USA 95, 9407–9412

Makalowski W, Zhang J, Boguski MS (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res 6, 846–857

Mardon G, Luoh SW, Simpson EM, Gill G, Brown LG et al. (1990) Mouse Zfx protein is similar to Zfy-2: each contains an acidic activating domain and 13 zinc fingers. Mol Cell Biol 10, 681–688

Marin M, Karis A, Visser P, Groxveld F, Philipsen S (1997) Transcription factor Sp1 is essential for each embryonic development but dispensable for cell growth and differentiation. Cell 89, 619–628

Matsuo K, Clay O, Takahashi T, Silke J, Schaffner W (1993) Evidence for erosion of mouse CpG islands during mammalian evolution. Somatic Cell Mol Genet 19, 543–555

Nan X, Tate P, Li E, Bird A (1996) DNA methylation specifies chromosomal localization of MeCP2. Mol Cell Biol 16, 414–421

Nan X, Campoy FJ, Bird A (1997) MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell 88, 471–481

Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM et al. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 393, 386–389

Nathans J, Thomas D, Hogness DS (1986) Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. Science 232, 193–202

Nizetic D, Zehetner G, Monaco AP, Gellen L, Young BD et al. (1991) Construction, arraying, and high-density screening of large insert libraries of human chromosomes X and 21: their potential use as reference libraries. Proc Natl Acad Sci USA 88, 3233–3237

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183, 63–98

Platzer M, Rotman G, Bauer D, Uziel T, Savitsky K et al. (1997) Ataxia-telangiectasia locus: sequence analysis of 184 kb of human genomic DNA containing the entire ATM gene. Genome Res 7, 592–605

Quaderi NA, Meehan RR, Tate PH, Cross SH, Bird AP et al. (1994) Genetic and physical mapping of a gene encoding a methyl CpG binding protein, Mecp2, to the mouse X chromosome. Genomics 22, 648–651

Rahmsdorf HJ, Schonthal A, Angel P, Litfin M, Ruther U et al. (1987). Posttranscriptional regulation of c-fos mRNA expression. Nucleic Acids Res 16, 1643–1659

Ruegsegger U, Blank D, Keller W (1998) Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. Mol Cell 1, 243–253

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press)

Takagaki Y, Seipelt RL, Peterson ML, Manley JL (1996) The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell 87, 941–952

Tate P, Skarnes W, Bird A (1996) The methyl-CpG binding protein MeCP2 is essential for embryonic development in the mouse. Nat Genet 12, 205–208

Vilain A, Apiou F, Vogt N, Dutrillaux B, Malfoy B (1996) Assignment of the gene for methyl-CpG-binding protein 2 (MECP2) to human chromosome band Xq28 by in situ hybridization. Cytogenet Cell Genet 74, 293–294

von Kries JP, Buhrmester H, Strätling WH (1991) A matrix/scaffold attachment region binding protein: identification, purification, and mode of binding. Cell 64, 123–135

Warren ST, Knight SJ, Peters JF, Stayton CL, Consalez GG et al. (1990) Isolation of the human chromosomal band Xq28 within somatic cell hybrids by fragile X site breakage. Proc Natl Acad Sci USA 87, 3856–3860

Weitzel JM, Buhrmester H, Strätling WH (1997) Chicken MAR-binding protein ARBP is homologous to rat methyl-CpG-binding protein MeCP2. Mol Cell Biol 17, 5656–5666

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13, 335–340