Nucleic Acids Research, 1993, Vol. 21, No. 3 419–426

Structure and expression of the excision repair gene *ERCC6*, involved in the human disorder Cockayne's syndrome group B

Christine Troelstra, Wouter Hesen, Dirk Bootsma and Jan H.J.Hoeijmakers*
MGC, Department of Cell Biology and Genetics, Erasmus University Rotterdam, PO Box 1738, 3000DR Rotterdam, The Netherlands

Received November 16, 1992; Revised and Accepted January 4, 1993

ABSTRACT

The human repair gene *ERCC6*—a presumed DNA (or RNA) helicase—has recently been found to function specifically in preferential nucleotide excision repair (NER). This NER subpathway is primarily directed towards repair of (the transcribed strand of) active genes. Mutations in the *ERCC6* gene are responsible for the human hereditary repair disorder Cockayne's syndrome complementation group B, the most common form of the disease. In this report, the genomic organization and expression of this gene are described. It consists of at least 21 exons, together with the promoter covering a region of 82 – 90 kb on the genome. Postulated functional domains deduced from the predicted amino acid sequence, including 7 distinct helicase signatures, are—with one exception—encoded on separate exons. Consensus splice donor and acceptor sequences are present at all exon borders with the exception of the unusual splice donor at the end of exon VII. The 'invariable' GT dinucleotide in the consensus (C,A)AG/GTPuAGT is replaced by the exceptional GC. Based on 42 GC splice donor sequences identified by an extensive literature search we found a statistically highly significant better 'overall' match of the surrounding nucleotides to the consensus sequence compared to normal GT-sites. This confirms and extends the observation made recently by Jackson (*Nucl. Acids Res.*, 19, 3795 – 3798 (1991)) derived from analysis of 26 cases. Analysis of *ERCC6* cDNA clones revealed the occurrence of alternative polyadenylation, resulting in the (differential) expression of two mRNA molecules (which are barely detectable on Northern blots) of 5 and 7 kb in length.

INTRODUCTION

Nucleotide excision repair (NER) is one of the major DNA repair systems functioning in mammalian cells (1). It is able to remove a broad variety of DNA lesions, such as UV-induced pyrimidine dimers and (6–4) photoproducts, as well as bulky chemical lesions and DNA cross links. After induction of DNA damage, the NER machinery appears to be directed first primarily towards the transcribed regions in the genome. This enables the cells to rapidly resume the vital process of transcription, that otherwise would remain blocked by lesions in the DNA template (2). This preferential repair of active genes is even specific for the transcribed strand (3). Removal of lesions from the non-transcribed strand, as well as from inactive chromatin proceeds more slowly and appears to be incomplete for several lesions (4 for a recent review). In human cells, repair of the non-transcribed strand is more rapid than that of inactive genes (5).

In the rare hereditary disorders xeroderma pigmentosum (XP) and Cockayne's syndrome (CS), deficiencies in NER are thought to underlie the clinical symptoms. XP patients are clinically characterized by severe sun(UV)sensitivity of the skin, pigmentation abnormalities, a highly elevated risk for developing skin tumors, and often neurological degeneration (6). Cell fusion experiments with cells from these patients have distinguished 7 NER-deficient complementation groups (cg; XP-A to XP-G) (7, 8). The biochemical basis for the XP NER-defect probably resides in undefined early steps of the pathway preceding incision and repair (9). For XP-C cells the defect appears to be specific for repair of inactive chromatin; these cells have retained the ability to remove damage from the transcribed strand of active genes (10, 11, 12). CS patients have, in sharp contrast to XP, no pigmentation abnormalities and no elevated risk for skin tumor formation. They are clinically characterised by skeletal and retinal abnormalities, growth retardation, progressive neurological degeneration, and a sunsensitive skin (13, 14 for reviews). Like XP, CS is heterogeneous: 2 cg have been identified (CS-A, -B) (15, 16). The molecular defect in CS-A and -B was shown to affect preferential repair of active genes only, whereas repair of the inactive chromatin regions proceeds normally (17). A third complementation group was defined (CS-C), containing one patient that exhibited a combination of characteristics of XP and CS. This patient is also classified as XP-B (16, 18). Several patients exhibiting this combined XP/CS phenotype are reported; one of these has been assigned to XP group D (8).

* To whom correspondence should be addressed

Another class of mammalian NER-deficient mutants consists of laboratory-induced, UV-sensitive, rodent (mainly chinese hamster) mutant cell lines. Among these, at least 10 cg have been defined (19, 20, 21, 22). A number of rodent mutants have been succesfully used for isolation of human genes capable of correcting the rodent repair deficiency (23 for a recent review). Subsequent introduction of these genes into XP and CS cells has demonstrated an overlap between the hamster and human cg: rodent cg 2 and 3, corrected by *ERCC2* and *-3*, are the rodent equivalents of XP-D and XP-B, respectively (24, 25). Recently, we have isolated the human repair gene *ERCC6* by virtue of its ability to correct the UV-sensitivity of the Chinese hamster ovary (CHO) mutant cell line UV61, cg 6 (26). The gene encodes a protein of 1493 amino acids; the predicted amino acid sequence suggests that the *ERCC6* gene product is a nuclear DNA helicase. Mutations in the *ERCC6* gene appear to be responsible for the repair deficiency of CS-B cells, implying that the *ERCC6* gene product is specifically involved in the process of preferential repair (27). This paper describes the genomic architecture and expression of the *ERCC6* gene.

MATERIALS AND METHODS

General procedures

Purification of nucleic acids, restriction enzyme digests, gel electrophoresis, DNA ligation, synthesis of radiolabeled probes using random oligonucleotide primers, the polymerase chain reaction (pcr), sequence analysis (dideoxy-mediated chain-termination), and filter hybridization were performed according to established procedures (28).

Construction of the cDNA plasmid

Construction of the (almost) full length *ERCC6* cDNA was as described (27). The cDNA insert was subcloned into both the vector pTZ19R (Pharmacia) yielding pTZE6total, and the mammalian expression vector pSLM, a derivative of pSVL (Pharmacia; Van Duin, unpublished results). In the mammalian expression vector the *ERCC6* cDNA is under the control of the SV40 late promoter, whereas in pTZE6total the cDNA is not preceded by a eukaryotic promoter.

Cell culture, transfection, and selection

UV-sensitive CHO cell line UV61 and wt CHO cell line AA8 were grown in 1:1 F10-Dulbecco minimal essential medium supplemented with antibiotics and 8% fetal calf serum. The *ERCC6* cDNA construct pTZE6total (3 μg) was cotransfected with 2 μg pSV2neo and 20 μg lambda phage 6B (see Fig.1) to UV61 cells. In order to release the lambda arms and plasmid vector from the inserts prior to transfection, both pTZE6total and lambda phage 6B were digested with *Sal* I; an enzyme which does not cut within the insert. Transfection and selection were performed as described before (27).

Identification of intron-exon borders

All genomic fragments hybridizing to the *ERCC6* cDNA were subcloned in pTZ19R (Pharmacia) or pBluescript II KS (Stratagene), and sequenced with *ERCC6*-specific primers. All sequence reactions were performed on double-stranded templates by the dideoxy chain termination method, using T7 DNA polymerase (Pharmacia). Intron length was determined either by restriction enzyme digestions and subsequent Southern blot analysis or by pcr with exon-specific primers on subcloned genomic fragments.

Northern blot analysis

RNA samples were separated on an agarose gel and transfered to a nylon membrane (Zeta probe from Bio-Rad) as described by Fourney et al. (29). The filters were hybridized at 65°C, in 3×SSC, 10×Denhardt's reagent, 0.1% SDS, 90 μg/ml dextran sulfate, and 50 μg/ml denatured salm sperm DNA. Filters were washed 2 times (10 min.) in 3×SSC, 0.1% SDS and once (10 min.) in 1×SSC, 0.1% SDS, at 65°C.

Synthesis of strand-specific probes

Strand-specific probes were synthesized according to the method described by Espelund et al. (30), with several modifications. The template was generated by pcr, with one normal and one biotinylated primer. The biotinylated product was then bound to streptavidin-coated magnetic beads (Dynabeads M-280, Dynal) through a 30 min. incubation at 37°C in 5×SSPE (20×SSPE: 3.6M NaCl, 200mM NaH_2PO_4 pH7.4, 20mM EDTA pH7.4). The beads were washed 4× with a solution containing 0.17% (w/v) Triton X-100, 100 mM NaCl, 10 mM Tris.HCl pH7.5, and 1 mM EDTA; subsequently the non-biotinylated strand was removed by 2 cycles (7 min. each) of denaturation in 125 mM NaOH, 100 mM NaCl. After washing twice with 100 mM Tris.HCl pH7.6, 150 mM NaCl, the biotinylated strand is radioactively labeled via primer extension. The beads were washed 3×, and the DNA was denatured by two incubations (7 min. each) in 125 mM NaOH, 100 mM NaCl. The supernatant was neutralized with an equal volume of 1M Tris.HCl (pH7.5), and used in hybridizations.

RESULTS

Architecture of the *ERCC6* gene

ERCC6 promoter region. The *ERCC6* gene was isolated from a lambda EMBL3 library originating from a repair-proficient secondary UV61 transformant. The genomic region coinherited by independent, repair-proficient primary and secondary UV61 transformants, as judged by Southern blot analysis, was approximately 100 kb, and consequently separated over several lambda clones (26). A physical map of this region is shown in Fig. 1. The size of the *ERCC6* locus was determined by hybridization of an (almost) full length, functional cDNA of 4.7 kb, encoding at least the total open reading frame and the 3′ end (of the shortest mRNA, see below), to the different lambda clones. The cDNA, previously localized on chromosome 10q11-21 (31), was shown to encompass 82 kb at the genome level (Fig. 1) and to reside on a 430 kb *Not* I fragment (data not shown). One of the *Not* I sites is part of a CpG-island present in the 5′ region of the *ERCC6* gene (Fig. 1). To pinpoint the region containing the promoter and the 5′ untranslated end of the mRNA (part of which might be lacking in the cDNA), a promoterless cDNA construct (pTZE6total, see Materials and Methods) was cotransfected with lambda clone 6B, containing the first 4 exons present in the cDNA, and approximately 8 kb more upstream sequences (Fig. 1). To be expressed, the promoterless cDNA should recombine within the cell with the cotransfected lambda phage, thus producing a 'mini gene'. After transfection of the two DNAs together UV-resistant clones were obtained, with an efficiency that was ≈10× lower than after

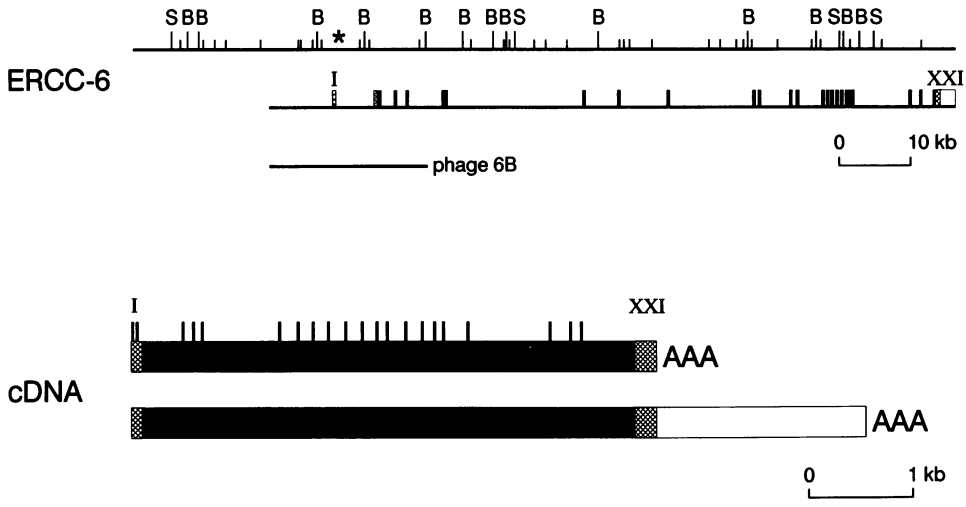

Figure 1. Genomic organization of *ERCC6*. A physical map of the *ERCC6* locus is shown. Exons (I to XXI) are indicated by boxes on the bar underneath the map. cDNAs from the two encoded, differentially poly-adenylated mRNAs are shown; white and hatched areas represent untranslated regions, the black area represents the ORF. Exon borders in the cDNA are indicated by vertical lines on top of the shortest cDNA. Phage 6B: a λ-phage containing the promoter region of the *ERCC6* gene. The * indicates a CpG-island encompassing a *Bss* HI, *Not* I, and a *Sac* II restriction enzyme site located just 3′ of exon 1. Other symbols: S, *Sal* I; B, *Bam* HI; short bars in the map represent *Eco* RI restriction sites.

transfection of a eukaryotic expression vector carrying the *ERCC6* cDNA behind the SV40 late promoter. The UV-sensitivity was corrected to a level within wt range (Fig. 2), as determined by survival experiments with pooled clones. To rule out the possibility that the UV-resistant transformants obtained in the cotransfection experiments were caused by fortuitous integration behind an endogenous promoter, instead of reconstitution of a functional gene by recombination between the 5′ *ERCC6* gene part with the rest of the cDNA, we also transfected the cDNA without a functional promoter at its 5′ end. In this experiment, in which the cDNA was placed in the inverted orientation in vector pSLM, no UV-resistant clones were obtained. These findings strongly suggest that a functional gene has indeed been generated via homologous recombination (between the lambda and cDNA clones) within the cell. The promoter should then be situated within the ±8 kb genomic region 5′ of the first exon present in the cDNA clone. The *ERCC6* locus thus covers a region of 82–90 kb.

Intron exon structure. Appropriate fragments were subcloned into plasmid vectors in order to determine intron-exon borders. Sequence analysis demonstrated the *ERCC6* cDNA to be dispersed over 21 exons (Fig. 1). The CpG-island present in the 5′ region of the *ERCC6* gene is situated within the first intron of *ERCC6* (Fig. 1). The first exon present in the cDNA does not contain any coding information, analogous to the first exon in another repair gene, *ERCC1* (32). Since we have not determined the transcriptional start site, we cannot exclude the presence of an additional exon more 5′ on the genome, which is absent in the cDNA. The presence of a CpG-island in the first intron, however, argues against the existence of an additional 5′ exon located far upstream. The tentatively identified functional domains are—with the exception of helicase signature VI—encoded by separate exons (Fig. 3 and 4).

As shown in Fig. 3, all sequences around intron-exon and exon-intron borders are consistent with the consensus splice

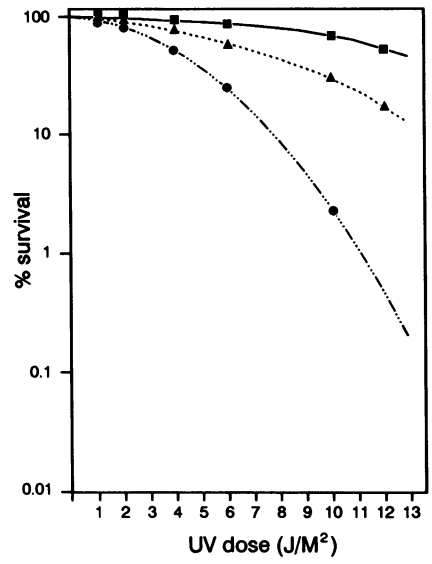

Figure 2. Correction of the UV-sensitivity of mutant UV61 by the *ERCC6* gene. UV survival curves of wild type CHO cell line AA8 (■——■), mutant UV61 (●..–..●), and UV61 cells cotransformed with λ-phage 6B (see Fig.1) and a promoterless *ERCC6* cDNA (▲...▲).

acceptor and donor signals (33), with the exception of the splice donor of the intron VII. The canonical GT dinucleotide at the beginning of the intron is replaced by GC (see below). In all introns, the weakly defined branchpoint sequence (PyNPyTPuAPy) could be tentatively identified at the appropriate distance (10–50 nt) from the splice acceptor site (underlined in Fig. 3) (33, 34). One alternatively spliced cDNA molecule has been isolated, in which exon VIII was found to be skipped. The sizes of internal exons range from 79 to 745 nt. Three of the internal exons (exon II, V, and XVIII) have an exceptional size:

INTRON SPLICE ACCEPTOR		EXON				INTRON SPLICE DONOR	
		I	(≥65)		CCCTGG	?	(6 kb)
TACAATGAATCCTATATAACGAAAA	66				501		
GGCTTATAAATTTTTTTCCTTTTAG	GTAGTC	II	(436)		CCTCAC	GTGAGGTCCAGCC	(1.8 kb)
TATAAAAATGTTATCAGAGTGCAAA	502				622		
ATAGACAATTTATCTTATTTTTCAG	GTCATG	III	(121)		AATAAG	GTGATTCAGAATA	(2.1 kb)
CAATAAATCACTTCATTGCTGTTTT	623				731		
CTGTCTTCTATATTTACCATTCAAG	GAACAA	IV	(109)		ATGCAG	GTGAGGATGCTAG	(4 kb)
CCCTCGGAAAGTTTCATGCTAGTGG	732				1476		
CAAAGCATTGCTATTGTTCTTTCAG	AGCCGG	V	(745)	acidic	GTTAAG	GTCGGTCTGTGGG	(19 kb)
TCTGTCTTGTGATCAAAATAATGGA	1477				1605		
AATGTGATTTTTATTTTCATGGTAG	GAGATG	VI	(129)	NLS/CKII	TTTTAA	GTATGTACCATAT	(5 kb)
TATCCACCATTTGCCATTTTCTCTT	1606				1764		
TTCTTGTTGGTGTTTGTTGTCATAG	GTACCA	VII	(159)	heli I	TTACAG	GCAAGTGCTCCTC	(7.5 kb)
TCTTTCATGGTT	1765				1900		
GTTTTCTTCTTGCGTTTGGATGCAG	GTTTGA	VIII	(136)	heli Ia	AAAAAG	GTAACACAATATT	(9 kb)
GGTGTCACTTCTTATTTAAAACTGA	1901				2071		
CTTTACCATTTTATTGTGGTCTCAG	GAGAAA	IX	(171)	heli II	AAACAG	GTATGACCTCTTT	(0.5 kb)
GATTGCTAAAAGGTAAATTTAATAA	2072				2248		
AAAGGTGCTTCCTTTCCTCAAATAG	TTTCGC	X	(177)	heli III	GTACAG	GTAAAATATTAGG	(3.5 kb)
TTCCTTAAACATATATAGTAAT	2249				2365		
GTTCCCTTCTCTGCTCTTATTAAAG	GTCAAA	XI	(117)		GAACAG	GTCTGTAAATCCA	(2 kb)
GTTAATTTTTTTTTTGAAATTATAG	2366				2461		
TTTCCTTGTTTTTCCCGGTTCGTAG	GTCTTA	XII	(96)		ATGCAG	GTCAGCTAAAAAA	(2 kb)
TTACTTTACATGGGGTCATCTGAGT	2462				2677		
GTACATGTACTCTTTCTTACGACAG	ATTTTC	XIII	(216)	heli IV	AGGCAG	GTGAGTGCACAGA	(0.5 kb)
CTGGGAATGTGTATTTGCTTTGCAA	2678				2788		
ACTCCTATCCCCCACCTCCAAACAG	ATGCTG	XIV	(111)		AATGAG	GTAACATGTGAAT	(0.4 kb)
TGTAACTGGTCTTAAGTGTGTGTGC	2789				2908		
TCAGTGTTGTGTGTCTTACCTCTAG	GACACA	XV	(120)	heli V, VI	ACGCAG	GTTTGTTTTTATT	(0.4 kb)
GTCATTGGGAAGGATTCTCGTTG	2909				3003		
AGAGGTCTCTCTCTCTCTGTTGCAG	GCCCGG	XVI	(95)	heli VI	CCACCG	GTCAGTGCACACA	(1.1 kb)
TAGGTAGAGCTACACATTGTTTTAT	3004				3149		
ACCAGCTTATCTTTTATTTTTTTAG	ACAAAT	XVII	(146)		TTGCAG	GTATTACATAAAA	(85 bp)
TTGCACAAGATGATACAATATAGTA	3150				3857		
TTAGTGGTGTTTTTCCTCTTTACAG	GAACTG	XVIII	(708)	NLS/CKII, NTB	AATCAG	GTAACCATTTGAT	(8.5 kb)
TTTCTTGCATACAGAGTGAAATATC	3858				4062		
ACTTTGCTATTTCTTTTCTTGCTAG	TTGGCG	XIX	(205)		AAAAAA	GTAAGAGATTGCT	(0.8 kb)
AAGTCTCAAAAGCAAACATTTAATC	4063				4141		
TACTGTGATGCTTTTCCTTTTTTAG	GAGTAG	XX	(79)		TGCCAG	GTAATATAGATAA	(1.2 kb)
GGCATAAACTAGAAATTAAATATAT	4142						
CAGTATAGTGGTCTTCTTTTTATAG	GATGGC	XXI	(573 / 2.5 kb)		3'poly(A)		

Figure 3. Structural organization of the *ERCC6* gene. The nucleotide sequence of each intron-exon junction is shown, with the exception of the splice donor of exon I. The vertical lines represent intron-exon borders. The splice donor of exon VII has a GC dinucleotide (indicated in bold) at the position of the canonical GT. All other acceptor and donor sites are in reasonable accordance with the consensus sequence $(Py)_nNCAG/G$ and (C,A)AG/GTPuAGT (33). The nucleotides at the borders of each exon are numbered as reported previously (27). The size of introns and exons are given between parenthesis (in bp, if not indicated otherwise). The postulated functional domains (27) encoded by the different exons are indicated: acidic, stretch of acidic aa; NLS/CKII, putative nuclear location signal followed by a postulated casein kinase II phosphorylation site; heli I to VI, helicase domains I to VI; NTB, putative nucleotide binding fold.

436, 745 and 708 nt, respectively; the average exon length for vertebrate internal exons being 137 nt (35). Intron lengths range from 85 to 19,000 nt; the average size of vertebrate introns being 1,127 nt (35).

ERCC6 gene expression and alternative polyadenylation

Northern blot analysis of human poly(A)$^+$ RNA demonstrated the presence of two *ERCC6* transcripts of approximately 5 and 7 kb in length, both expressed at a very low level (Fig. 5). To examine the structure of the long *ERCC6* mRNA, various cDNA libraries were screened for the presence of clones with extra sequences not found in the 4.7 kb cDNA. One of the isolated (partial) cDNAs had a 3′ untranslated region extending 2 kb more 3′ as the mentioned 4.7 kb construct. It contained a poly(A)tail, preceded by the common polyadenylation signal AATAAA (36, 37). The total length of the isolated cDNA then becomes ±6.7 kb, probably reflecting the longest mRNA. Sequence analysis and restriction enzyme digests showed the 3′ untranslated region to be colinear with the genome. This suggests the 2 mRNA products to be the result of alternative polyadenylation (Fig. 1).

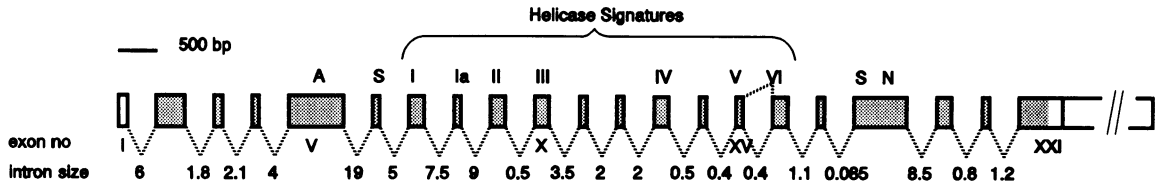

Figure 4. Schematic depiction of the postulated functional domains as encoded by different exons. The shaded area represents the ORF; encoded domains are indicated above the exons. Intron length is given in kb. Symbols: A, acidic amino acid stretch; S, putative nuclear location signal followed by a postulated casein kinase II phosphorylation site; N, putative nucleotide binding fold. The interrupted box added to exon XXI indicates the 3′ end of the longest *ERCC6* mRNA, produced by alternative polyadenylation.

Analysis of mouse tissues revealed the presence of elevated (though still very low) levels of *ERCC6* mRNA in brain and testis; in most other tissues transcription was below detection level (Fig. 5). Also in mouse two transcripts are identified, suggesting that the alternative polyadenylation is conserved. Remarkably, the larger transcript is the most abundant in mouse brain, whereas in testis the smaller mRNA is the more frequently occurring one (Fig. 5).

In human RNA, a third transcript of 3.5 kb is detected, with probes in the 5′ 1.7 kb of the gene. A probe from nt 1662 to 3746 did not detect this short transcript. Both a probe covering the first 350 bp of the cDNA (exon I and part of exon II) and one spanning nt 1020 to 1667 (part of exon V, exon VI and 60 nt of exon VII) recognize this short transcript, suggesting that at least sequences from exon I or II as well as exon V or VI (and maybe the 60 nt of exon VII) should be present. The transcript is detected in both total and poly A^+ RNA. Hybridizations with strand-specific probes, as shown in Fig. 5, have excluded the possibility of an overlapping antisense gene: the transcript appears to be produced from the same strand as the two longer *ERCC6* mRNAs. Since we have been unable to identify *ERCC6* transcripts in mouse tissues with a 5′ probe, it is unclear whether the 3.5 kb transcript is conserved through evolution.

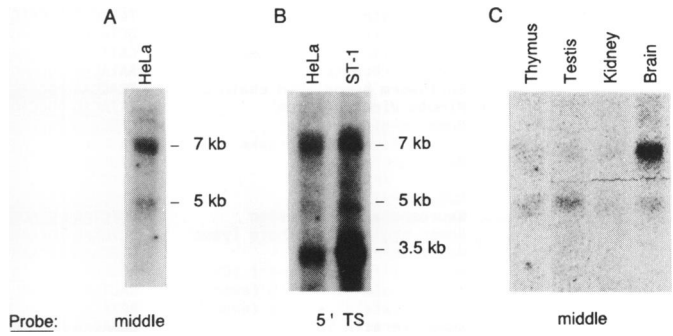

Figure 5. Expression of the *ERCC6* gene. **A**: Northern blot analysis of 5 μg poly(A)$^+$ mRNA from HeLa cells, hybridized with a ^{32}P-labeled *ERCC6* cDNA fragment extending from nt 1662–3746 (middle probe). **B**: Northern blot analysis of 5 μg poly(A)$^+$ mRNA from HeLa cells and ST1, a secondary repair proficient transformant of UV61, containing the human *ERCC6* gene (26). The blot is hybridized with a strand-specific probe (^{32}P-labeled transcribed strand, TS) spanning nt 16 to 1667 from the *ERCC6* cDNA. A radioactively labeled non-transcribed strand did not recognize any transcript (result not shown). **C**: Northern blot analysis of 20 μg of total RNA from different mouse tissues. The probe used is similar to that in *A*.

DISCUSSION

The *ERCC6* gene product, which is predicted to be a nuclear DNA helicase, has recently been shown to be involved in the human DNA repair disorder Cockayne's syndrome (27). In this paper, the expression and structural organization of the *ERCC6* gene are reported.

Two lowly expressed *ERCC6* mRNA molecules of 5 and 7 kb have been identified by Northern blot analysis of human poly(A)$^+$ RNA (Fig. 5). Two types of cDNAs have been isolated, varying in their 3′ end by differential polyadenylation. The second polyadenylation site, accompanied by the common polyadenylation signal AATAAA, is present 2 kb downstream of the first one that contains the less preferred signal ATTAAA (36, 37). Alternative polyadenylation occurs in many genes, including another human NER gene *ERCC1* (32). It is unclear though, whether this has any regulatory function. In this respect, it is interesting to note that often the alternative polyadenylation is evolutionarily conserved, as it is—presumably—for *ERCC6*. We noted differential expression of the two transcripts in mouse brain and testis (Fig. 5). It is unknown whether the elevated expression of the longest *ERCC6* mRNA in brain is related to the (severe) progressive neurological degeneration in CS patients.

Also the function of the (human) 3.5 kb transcript encompassing only the 5′ region of the *ERCC6* cDNA is unresolved. Hybridization with strand-specific probes ruled out the possibility of an overlapping antisense gene. An overlapping gene in sense direction seems unlikely, since it should then have (parts) of at least 2 exons (exons I or II and V, VI or VII) in common with the *ERCC6* gene. A third possibility is that the 3.5 kb transcript is the result of a (strong) transcriptional stop due to the presence of an attenuation site. The attenuation site should then be within or 3′ of exon V, which is approximately 15 kb (or more) from the transcriptional start site. Normally, attenuation takes place within a few hundred bp from the transcription initiation site, although there is one example of a termination site ≈2 kb from the start (c-*myb*) (38). If, for the *ERCC6* gene, the stop site is 3′ of exon V, it could be within the 19 kb long intron V. Cross-hybridization of unique human sequences from this intron to rodent DNA suggested the presence of 2 conserved regions within intron XIX (unpublished observations). It is possible that a second gene is located within this large intron, with transcription termination sequences that can be recognized by RNA polymerase complexes transcribing the *ERCC6* gene. Splicing of this prematurely terminated transcript might result in the observed 3.5 kb mRNA. A more conclusive statement concerning the role of this short (truncated) *ERCC6* transcript should, however, await further research, particularly by the isolation of corresponding

Table I. Summary of GC-splice donor sequences

GENE	SEQUENCE		REFERENCE
	EXON	INTRON	
Human cytochrome P-450	CACTAAG	GCAAGCCCACA	(43 and references therein)
Hamster αA-crystallin	CATCAAG	GCAAGTTACAT	(43 and references therein)
Mouse αA-crystallin	CGTCAAG	GCAAGTTACAT	(43 and references therein)
Mole rat αA-crystallin	CATCAAG	GCAAGTTTCGT	(43 and references therein)
Chicken αD-globin	TTTCAAG	GCAAGCAAAGG	(43 and references therein)
Duck αD-globin	CTTCAAG	GCAAGCGGGGA	(43 and references therein)
Chicken myosin heavy chain	GCTGCAG	GCAAGTGTCTG	(43 and references therein)
Rat heme oxygenase	TCGACAG	GCAAGCGACTA	(43 and references therein)
Soybean nodulin-24	AAAGAGG	GCAAGTTAATT	(43 and references therein)
Human factor XII	AGGACCG	GCGAGTACCCG	(43 and references therein)
Pig growth hormone	GCTGCAG	GCAAGTGCCCC	(43 and references therein)
Human acetylcholine receptor	CCGCAAG	GCAAGGACCCT	(43 and references therein)
Rat pyruvate kinase	CACCCAG	GCATGTGCTAT	(45)
Human superoxide dismutase-1	GCAGAAG	GCAAGGGCTGG	(43 and references therein)
Mouse superoxide dismutase-1	GCAGAAG	GCAAGGCCCGG	(43 and references therein)
Human prothrombin	TGTGCTG	GCAAGTCTGTG	(43 and references therein)
Human factor VII	GCTGCAG	GCGGGGTCGCT	(43 and references therein)
Human erythrocyte α-spectrin	CATTCAG	GCAAGTTCAA	(46)
Rat D2a receptor	AATACAG	GCAAGTCTGGC	(47)
Earthworm hemoglobin chain c	TCACCAA	GCAAGTCTCCC	(43 and references therein)
Minute virus of mice	TTTACAG	GCCTGAAATC	(48)
Human α-glucosidase	CACCAAG	GCAAGA	(49)
Bovine aspartyl protease	TGGCGAG	GCAAGTCCAG	(43 and references therein)
Mouse APRT	ATCGCAG	GCGAGTGGCCT	(43 and references therein)
Hamster APRT	ATCGCAG	GCGAGTGGCCA	(43 and references therein)
Human APRT	ATCGCAG	GCGAGTGCCAG	(43 and references therein)
Neurospora qa repressor	CCTCTAG	GCACGTCTGTA	(43 and references therein)
Human argininosuccinate lyase	GCTGCAG	GCAAGACATCA	(50)
Mouse TRP-1	GTGGAAG	GCAAGTAA	(43 and references therein)
Murine fifth complement (C5)	ACGGGCT	GCAAGTGGT	(51)
Human platelet GP IIb (exon V)	CACTCAG	GCGAGTAGGGA	(52)
Human platelet GP IIb (exon VIII)	ACTACAG	GCAAGAAATCC	(52)
Human keratin 18	CGAAGAG	GCAAGCAGGGG	(43 and references therein)
Mouse RNA polymerase (exon VII)	CTATAAG	GCATGTAAATA	(43 and references therein)
Mouse RNA polymerase (exon XIII)	CTCCCAG	GCAAGATGCTT	(43 and references therein)
E.typhina TUB-8	CAACGAG	GCAAGTCTTCA	(53)
Rat ESP-1	CTATCAG	GCAAGAATGCT	(54)
HSV-1 LAT	CAAGAAG	GCATGTGTCCC	(55)
Human C3	TACCCAG	GCAAGT	(56)
Human DNA ligase I	ACGCAAG	GCAAGT	(57)
Human ERCC3	TATTAAG	GCAAGTGACAG	(40)
Human ERCC6	ATTACAG	GCAAGTGCTCC	(This paper)
Sequence complementary to U1 snRNA	CAG	GTAAGT	(58)

cDNAs. So far, no such cDNAs have been isolated, despite the screening of several cDNA libraries.

The *ERCC6* locus is 82–90 kb in length, and harbors at least 21 exons. The presence of a CpG-island within intron I (Fig. 1) suggests that there is no unidentified exon more 5′ (containing extra 5′ untranslated sequences not present in the cDNA), although we cannot completely exclude this possibility. It is interesting to note that also *ERCC2* and *ERCC3* have CpG-islands within the first intron (39, 40). The putatively identified functional domains in the *ERCC6* cDNA sequence are—with one exception—dispersed over separate exons. It has been proposed that modern eukaryotic genes may have been assembled from a limited number of ancestral exons encoding separate functional domains (41). Within *ERCC6*, only helicase domain VI is split by intron XV (see Fig. 3 and 4). Interruption of helicase domains by introns has been noted before: yeast RNA helicase p68 contains one intron in domain V (42), which is positionally conserved (between human, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*); within the *ERCC3* gene (encoding a putative DNA helicase) both domain I and V are interrupted (40).

In one of the isolated *ERCC6* cDNAs exon VIII was found to be skipped. Although we cannot completely exclude the possibility that a mRNA molecule without exon VIII has any functional potential, it seems unlikely. The exon skipping results

Table II. Nucleotide percentages at GC-type compared to GT-type splice donors

Consensus GT splice[a]	-1 C/A	-2 A	-3 G	1 G	2 T	3 A/G	4 A	5 G	6 T
G	18	12	79	100	-	35	11	82	18
A	32	60	9	-	-	59	71	7	16
T	13	15	7	-	100	3	9	6	50
C	37	13	5	-	-	3	9	6	16

Consensus GC splice[b]	C	A	G	G	C	A	A	G	T
G	9.5	2.4	96.2	100	-	14.3	2.4	100	9.5
A	38.1	90.4	2.4	-	-	83.3	85.7	-	14.3
T	2.4	2.4	2.4	-		-	9.5	-	54.8
C	50	4.8	-	-	100	2.4	2.4	-	11.9

Significance of increased complementarity to U1 snRNA in GC splice donor sites (shaded boxes)								
p-values	<0.1 (ns)[c]	<0.001	-	-	<0.001	<0.05	<0.01	<0.1 (ns)[c]

[a] Data taken from reference (33).
Percentages are based on 3724 different slice donors.
[b] Data from this work and reference (43).
Percentages are based on 42 different splice donors.
[c] ns = not significantly different.

in a frameshift; consequently, the mRNA only encodes a very short protein. Presumably, this unconventionally spliced mRNA is the result of an erroneous splice event, rather than the product of a mechanism that creates an alternative, functional *ERCC6* protein.

All splice sites, with the exception of the splice donor at the 3′ end of exon VII, are in reasonable accordance with the consensus sequences $(Py)_nCAG/G$ and (C,A)AG/GTPuAGT for donor and acceptor, respectively (33). In the splice donor of intron VII, the 'invariable' GT dinucleotide is replaced by GC. 'GC' instead of 'GT' splice donors are very rare. Recently, Jackson (43) compared 26 'GC' splice donors from the literature, and noted that as a group they have a better 'overall' match to the (C/A)AG/GTAAGT consensus than the regular 'GT' donors. To assess the validity of this finding and to make conclusions statistically more significant, we have screened various data bases for additional 'GC' splice donors, and extended the list to a total of 42 (Table I). The comparison presented in Table II clearly confirms and strengthens the observation made by Jackson: 'GC' splice sequences display as a whole at all (remaining) positions a substantially higher level of complementarity to the U1 snRNA than the group of 'GT' donors. Particularly the G residue at position +5 seems invariant, whereas the 'G' at −1 is almost invariant. Most sites have 6−8 bases matching U1 snRNA (6 sites have 8 matches, 21 sites have 7 matches, 12 sites have 6 matches, the remaining 3 sites only have 5 bases matching to U1 snRNA). Apparently, the mismatch in base pairing with U1 snRNA caused by the GT to GC alteration needs to be compensated by a better match of the rest of the sequence, in order to permit the assembly of a functional spliceosome.

Three of the internal *ERCC6* exons (exons without either a cap site or a polyadenylation signal) are exceptionally large: 436, 708, and 745 nt (exons II, XVIII, and V respectively). Out of 1305 internal vertebrate exons examined by Hawkins, only 7 exceed 550 nt; the average length being 137 nt (35). A conceivable explanation for the scarcity of large internal exons, as proposed by Robberson and colleagues (44) is based on the idea that, in order to stably form a spliceosome, factors bound at the 5′ and 3′ border of an exon need to 'communicate'. The interaction between the different factors might become less stable if the exon length exceeds a certain limit. *In vitro* splicing experiments demonstrated splicing of an intron followed by a large exon (>300 nt) to be less efficient than splicing of the same intron preceding a small exon (<300 nt) (44). Whether this is true *in vivo* too, is unknown. If so, it might (partly) explain the low expression level of the *ERCC6* gene.

ACKNOWLEDGEMENTS

We thank Mirko Kuit for photography, and Bart Janssen (Dept. Clin. Genet.) for help in some of the experiments.

REFERENCES

1. Friedberg, E.C. (1985) DNA repair. W. H. Freeman and Company, San Francisco.
2. Bohr, V.A., Smith, C.A., Okumoto, D.S. and Hanawalt, P.C. (1985) *Cell*, 40, 359−369.
3. Mellon, I., Spivak, G. and Hanawalt, P.C. (1987) *Cell*, 51, 241−249.
4. Link Jr., C.J., Burt, R.K. and Bohr, V.A. (1991) *Cancer Cells*, 3, 427−436.
5. Venema, J., Bartosova, Z., Natarajan, A.T., van Zeeland, A.A. and Mullenders, L.H.F. (1992) *J. Biol. Chem.*, 267, 8852−8856.
6. Cleaver, J.E. and Kraemer, K.H. (1989) *In* Scriver, C.R., A.L. Beaudet, W.S. Sly, D. Valle (ed.), Xeroderma pigmentosum. McGraw-Hill Book Co., New York, 2949−2971.
7. De Weerd-Kastelein, E.A., Keijzer, W. and Bootsma, D. (1972) *Nature (London) New Biology*, 238, 80−83.
8. Vermeulen, W., Stefanini, M., Giliani, S., Hoeijmakers, J.H.J. and Bootsma, D. (1991) *Mutat. Res.*, 255, 201−208.
9. Shivji, M.K.K., Kenny, M.K. and Wood, R.D. (1992) *Cell*, 69, 367−374.
10. Venema, J., Van Hoffen, A., Natarajan, A.T., Van Zeeland, A.A. and Mullenders, L.H.F. (1990) *Nucl. Acids Res.*, 18, 443−448.
11. Kantor, G.J., Barsalou, L.S. and Hanawalt, P.C. (1990) *Mutat. Res.*, 235, 171−80.
12. Venema, J., van Hoffen, A., Karcagi, V., Natarajan, A.T., van Zeeland, A.A. and Mullenders, L.H.F. (1991) *Mol. Cell. Biol.*, 11, 4128−34.
13. Lehmann, A.R. (1987) *Cancer Rev.*, 7, 82−103.
14. Nance, M.A. and Berry, S.A. (1992) *Am. J. Med. Genet.*, 42, 68−84.
15. Tanaka, K., Kawai, K., Kumahara, Y., Ikenaga, M. and Okada, Y. (1981) *Somat. Cell Genet.*, 7, 445−455.
16. Lehmann, A.R. (1982) *Mutat. Res.*, 106, 347−356.
17. Venema, J., Mullenders, L.H.F., Natarajan, A.T., Van Zeeland, A.A. and Mayne, L.V. (1990) *Proc. Natl. Acad. Sci. USA*, 87, 4707−4711.
18. Robbins, J.H., Kraemer, K.H., Lutzner, M.A., Festoff, B.W. and Coon, H.G. (1974) *Ann. Intern. Med.*, 80, 221−248.
19. Busch, D., Greiner, C., Lewis, K., Ford, R., Adair, G. and Thompson, L. (1989) *Mutagenesis*, 4, 349−354.
20. Zdzienicka, M.Z., Schans, G.P.v.d. and Simons, J.W.I.M. (1988) *Mutat. Res.*, 194, 165−170.
21. Thompson, L.H., Shiomi, T., Salazar, E.P. and Stewart, S.A. (1988) *Somat. Cell. Mol. Genet.*, 14, 605−612.
22. Stefanini, M., Collins, A.R., Riboni, R., Klaude, M., Botta, E., Mitchell, D.L. and Nuzzo, F. (1991) *Cancer Res*, 51, 3965−3971.
23. Hoeijmakers, J.H.J. and Bootsma, D. (1990) *Cancer Cells*, 2, 311−320.
24. Flejter, W.L., McDaniel, L.D., Johns, D., Friedberg, E.C. and Schultz, R.A. (1992) *Proc. Natl. Acad. Sci. USA*, 89, 261−265.
25. Weeda, G., Van Ham, R.C.A., Vermeulen, W., Bootsma, D., Van der Eb, A.J. and Hoeijmakers, J.H.J. (1990) *Cell*, 62, 777−791.
26. Troelstra, C., Odijk, H., De Wit, J., Westerveld, A., Thompson, L.H., Bootsma, D. and Hoeijmakers, J.H.J. (1990) *Mol. Cell. Biol.*, 10, 5806−5813.
27. Troelstra, C., van Gool, A., de Wit, J., Vermeulen, W., Bootsma, D. and Hoeijmakers, J.H.J. (1992) *Cell*, 71, 939−953.
28. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
29. Fourney, R.M., Miyakoshi, J., III, R.S.D. and Paterson, M.C. (1988) *Focus*, 10, 5−7.
30. Espelund, M., Prentice Stacy, R.A. and Jakobsen, K.S. (1990) *Nucl. Acids Res.*, 18, 6157−6158.
31. Troelstra, C., Landsvater, R.M., Wiegant, J., Ploeg, M.v.d., Viel, G., Buys, C.H.C.M. and Hoeijmakers, J.H.J. (1992) *Genomics*, 12, 745−749.
32. Van Duin, M., Koken, M.H.M., van den Tol, J., ten Dijke, P., Odijk, H., Westerveld, A., Bootsma, D. and Hoeijmakers, J.H.J. (1987) *Nucl. Acids Res.*, 15, 9195−213.
33. Senapathy, P., Shapiro, M.B. and Harris, N.L. (1990) *Meth. Enzymol.*, 183, 252−278.
34. Ruskin, B., Krainer, A.R., Maniatis, T. and Green, M.R. (1984) *Cell*, 38, 317−331.
35. Hawkins, J.D. (1988) *Nucl. Acids Res.*, 16, 9893−9908.
36. Wickens, M. (1990) *Trends Biochem. Sci.*, 15, 277−281.
37. Birnstiel, M.L., Busslinger, M. and Strub, K. (1985) *Cell*, 41, 349−359.
38. Watson, R.J. (1988) *Oncogene*, 2, 267−272.
39. Weber, C.A., Salazar, E.P., Stewart, S.A. and Thompson, L.H. (1990) *EMBO J.*, 9, 1437−1447.
40. Weeda, G., Ma, L.B., van Ham, R.C., van der Eb, A.J. and Hoeijmakers, J.H. (1991) *Nucl. Acids Res.*, 19, 6301−8.
41. Dorit, R.L., Schoenbacher, L. and Gilbert, W. (1990) *Science*, 250, 1377−1382.
42. Iggo, R.D., Jamieson, D.J., MacNeill, S.A., Southgate, J., McPheat, J. and Lane, D.P. (1991) *Mol. Cell. Biol.*, 11, 1326−1333.
43. Jackson, I.J. (1991) *Nucl. Acids Res.*, 19, 3795−3798.
44. Robberson, B.L., Cote, G.J. and Berget, S.M. (1990) *Mol. Cell. Biol.*, 10, 84−94.
45. Takenaka, M., Noguchi, T., Inoue, H., Yamada, K., Matsuda, T. and Tanaka, T. (1989) *J. Biol. Chem.*, 264, 2363−2367.

46. Kotula, L., Laury-Kleintop, L.D., Showe, L., Sahr, K., Linnenbach, A.J., Forget, B. and Curtis, P.J. (1991) *Genomics*, 9, 131–140.
47. O'Malley, K.L., Mack, K.J., Gandelman, K.Y. and Todd, R.D. (1990) *Biochem.*, 29, 1367–1371.
48. Jongeneel, C.V., Sahli, R., McMaster, G.K. and Hirt, B. (1986) *J. Virol.*, 59, 564–573.
49. Hoefsloot, L.H., Hoogeveen-Westerveld, M., Reuser, A.J.J. and Oostra, B.A. (1990) *Biochem. J.*, 272, 493–497.
50. Abramson, R.D., Barbosa, P., Kalumuck, K. and O'Brien, W.E. (1991) *Genomics*, 10, 126–132.
51. Haviland, D.L., Haviland, J.C., Fleischer, D.T. and Wetsel, R.A. (1991) *J. Biol. Chem.*, 266, 11818–11825.
52. Heidenreich, R., Eisman, R., Surrey, S., Delgrosso, K., Bennett, J.S., Schwartz, E. and Poncz, M. (1990) *Biochem.*, 29, 1232–1244.
53. Byrd, A.D., Schardl, C.L., Songlin, P.J., Mogen, K.L. and Siegel, M.R. (1990) *Curr. Genet.*, 18, 347–354.
54. Girotti, M., Jones, R., Emery, D.C., Chia, W. and Hall, L. (1992) *Biochem. J.*, 281, 203–210.
55. Spivack, J.G., Woods, G.M. and Fraser, N.W. (1991) *J. Virol.*, 65, 6800–6810.
56. Barnum, S.R., Amiguet, P., Amiguet-Barras, F., Fey, G. and Tack, B.F. (1989) *J. Biol. Chem.*, 264, 8471–8474.
57. Noguiez, P., Barnes, D.E., Mohrenweiser, H.W. and Lindahl, T. (1992) *Nucl. Acids Res.*, 20, 3845–3850.
58. Rosbash, M. and Seraphin, B. (1991) *Trends Biochem. Sci.*, 16, 187–190.