GENOMICS 34, 399–409 (1996)
ARTICLE NO. 0303

Sequence Analysis of the *ERCC2* Gene Regions in Human, Mouse, and Hamster Reveals Three Linked Genes

JANE E. LAMERDIN,*[,1] STEPHANIE A. STILWAGEN,* MELISSA H. RAMIREZ,*
LISA STUBBS,† AND ANTHONY V. CARRANO*

**Human Genome Center, Biology and Biotechnology Research Program, L-452, Lawrence Livermore National Laboratory, Livermore, California 94550; and †Biology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-8077*

Received October 8, 1995; accepted April 9, 1996

The *ERCC2* (excision repair cross-complementing rodent repair group 2) gene product is involved in transcription-coupled repair as an integral member of the basal transcription factor BTF2/TFIIH complex. Defects in this gene can result in three distinct human disorders, namely the cancer-prone syndrome xeroderma pigmentosum complementation group D, trichothiodystrophy, and Cockayne syndrome. We report the comparative analysis of 91.6 kb of new sequence including 54.3 kb encompassing the human *ERCC2* locus, the syntenic region in the mouse (32.6 kb), and a further 4.7 kb of sequence 3′ of the previously reported *ERCC2* region in the hamster. In addition to *ERCC2*, our analysis revealed the presence of two previously undescribed genes in all three species. The first is centromeric (in the human) to *ERCC2* and is most similar to the kinesin light chain gene in sea urchin. The second gene is telomeric (in the human) to *ERCC2* and contains a motif found in ankyrins, some cell cycle proteins, and transcription factors. Multiple EST matches to this putative new gene indicate that it is expressed in several human tissues, including breast. The identification and description of two new genes provides potential candidate genes for disorders mapping to this region of 19q13.2.

INTRODUCTION

The ability to repair damage to its DNA is fundamental to an organism's survival. Both prokaryotes and eukaryotes have evolved elaborate mechanisms to accomplish this goal, one of which is the nucleotide excision repair (NER) pathway. The *ERCC2* (excision repair cross-complementing rodent repair group 2) gene product is involved in transcription-coupled NER as an integral member of the basal transcription factor BTF2/TFIIH complex (Schaeffer *et al.*, 1994; Drapkin *et al.*, 1994). Mutations in this gene lead to three very different human disorders: the cancer-prone syndrome xeroderma pigmentosum group D (XP-D) (Frederick *et al.*, 1994; Takayama *et al.*, 1995), trichothiodystrophy (TTD; Broughton *et al.*, 1994), and Cockayne syndrome (CS), whose clinical features include neurological abnormalities associated with nerve demyelination and severe growth retardation (Bootsma and Hoeijmakers, 1994). The diverse syndromes associated with defects in this one gene are indicative of its complex role in transcription and repair.

The human *ERCC2* gene maps to the long arm of human chromosome 19 to a region containing several other genes involved in DNA repair and metabolism, as depicted in Fig. 1. *ERCC1* (excision repair cross-complementing rodent repair group 1) and *ERCC2* are located within 250 kb of each other in 19q13.2–q13.3 and within 2 Mb of the *XRCC1* (X-ray repair cross-complementing rodent repair group 1) locus in 19q13.2 (Mohrenweiser *et al.*, 1989; Smeets *et al.*, 1990). The gene encoding DNA Ligase I (Barnes *et al.*, 1992) is located another 2.5 Mb telomeric of *ERCC1*. The muscle creatine kinase (*CKMM*) locus is centromeric of *ERCC2*, and both loci are located on the same 40-kb *Eco*RI fragment (Smeets *et al.*, 1990). The *XRCC1–CKMM* linkage group is conserved in the mouse (Brilliant *et al.*, 1994) and the hamster (Thompson *et al.*, 1989), where it maps to chromosomes 7 and 9, respectively. The *XRCC1–CKMM* loci are represented in two distinct linkage groups in the teleost fish, *Xiphophorus maculatus* (Walter *et al.*, 1991), with *CKMM* and *ERCC2* separated from the *XRCC1* block into linkage group U5.

The conservation of this synteny block among these species provides an ideal substrate for a comparative sequencing study designed to elucidate the genomic structure of these loci and to identify additional conserved functional motifs and genes. The genomic regions spanning the human *ERCC1* (Martin-Gallardo *et al.*, 1992), the human and mouse *XRCC1* (Lamerdin *et al.*, 1995), the fish *ERCC2* (Della Coletta *et al.*, 1995), and the hamster *ERCC2* (Kirchner *et al.*, 1994) loci

[1] To whom correspondence should be addressed at Human Genome Center, Biology and Biotechnology Research Program, L-452, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550. Telephone: (510) 423-3629. Fax: (510) 422-2282.

0888-7543/96 $18.00

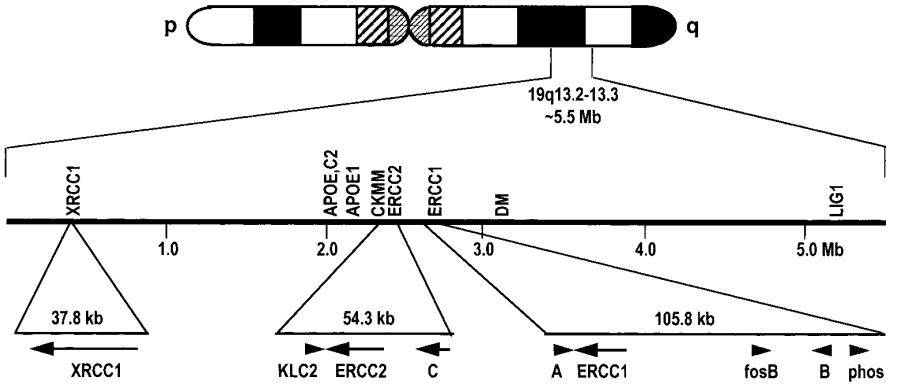

FIG. 1. A map of the 19q13.2–q13.3 region containing the three DNA repair genes *XRCC1*, *ERCC2*, and *ERCC1*. The map locations of five other genes are also shown. They include the apolipoprotein genes (*APOE*, *C2*, and *E1*), the muscle creatine kinase gene locus (*CKMM*), myotonic dystrophy (*DM*), and DNA Ligase I (*LIG1*). Sequenced regions are expanded. These include the 54.3-kb *ERCC2* gene region described in this paper, the *XRCC1* region (Lamerdin *et al.*, 1994), and the *ERCC1* region (Martin-Gallardo *et al.*, 1992). The arrows represent the direction of transcription of the genes identified by sequence analysis. *Gene C* identified in this work was named alphabetically after genes *A* and *B* identified by Martin-Gallardo *et al.* (1992).

have already been sequenced. We report here the analysis of 54.3 kb of sequence encompassing the human *ERCC2* locus, as well as 32.4 kb of the homologous region in the mouse and an additional 4.7 kb downstream of the previously sequenced *ERCC2* gene in the hamster (~16 kb).

MATERIALS AND METHODS

Cosmid subcloning and sequencing. Human (de Jong *et al.*, 1989) and hamster (Kirchner *et al.*, 1994) cosmids containing the *ERCC2* gene were identified by hybridization with 5′-region and/or 3′-region human cDNA probes (Weber *et al.*, 1990) as previously described (Kirchner *et al.*, 1994). An *ERCC2*-positive mouse cosmid was similarly obtained from a 129/Sv (male) library constructed in the pCos2 vector (Ehrich *et al.*, 1987), kindly provided by Dr. Anna-Maria Frischauf (ICRF, London, UK). The subcloning and sequencing of the ~16-kb hamster *ERCC2* essential gene region have been described (Kirchner *et al.*, 1994; Accession Nos. U04967 and U04968). The mouse (MXPD 1.11) and two overlapping human cosmids (f25251 and f19186) were sonicated, end-repaired with Klenow and T4 polymerase (Martin-Gallardo *et al.*, 1993), size-selected, and cloned into the *Hin*cII site of M13mp18. An additional 4.7-kb *Hin*dIII–*Eco*RI fragment distal to the *ERCC2* gene in the hamster cosmid was subcloned in a similar manner. DNA templates for sequencing were prepared using Qiagen 96-well format M13 kits according to the manufacturer's instructions (Qiagen, Chatsworth, CA). Single-stranded templates were sequenced on a Catalyst 800 Molecular Biology Labstation (Applied Biosystems Division (ABD), Perkin–Elmer, Foster City, CA), using a fluorescently labeled universal-21m13 primer and ABD *Taq* cycle sequencing kits. Resultant sequencing ladders were loaded on 4.75% polyacrylamide gels, and data were collected on ABD 373A DNA sequencers with a 35-cm well-to-read region distance. Double-strand sequence continuity was obtained primarily from reverse-strand reads generated by PCR of the insert from selected M13 clones as described (Muzny *et al.*, 1994) or by sequencing the RF form. Vector and ambiguous 3′-tailing sequence were identified and masked by the ABD Factura program, followed by assembly in the ABD Auto Assembler package on a Macintosh Power PC 8100/100 Hz, using the fast data finder option.

A total of 1859 clones from the overlapping human cosmids spanning a 54.3-kb region were sequenced to an average redundancy of 13-fold. The sequence was determined on both strands for 92% of the region; those areas not covered on both strands are flanked by or encompass *Alu* or long simple sequence repeats. For the mouse cosmid, 641 clones were sequenced, 557 of which covered the 32.6-kb insert; complete double-strand coverage was obtained at an average redundancy of 8-fold. The additional 4.7-kb fragment from the hamster cosmid was sequenced to an average redundancy of 11-fold, with all but 90 bp of a hairpin structure sequenced on both strands.

All of the sequences reported here were validated by comparison of the "digested" sequence to detailed restriction maps of the regions as well as by PCR analysis as previously described (Lamerdin *et al.*, 1995). The sequence of each cosmid/region has been deposited in the Genome Sequence Data Base (GSDB) under Accession Nos. L47234 for the human, L47235 for the mouse, and L47236 for the hamster.

Sequence analysis. Regions conserved in human, hamster, and mouse were identified using dblast and lax (Hardison and Miller, 1993) and compared to coding region predictions made by XGRAIL 1.2 (Uberbacher *et al.*, 1991). Repetitive elements in all three species were initially identified by comparison to a subset of known human (*Alu*, L1, THE, LTR, PE670) and mouse (B1, B2, L1) repeats using ALIGN (Intelligenetics, Mountain View, CA). Additional searches were performed against the GSDB (daily update) and the Repetitive Element Data Base (Jurka *et al.*, 1992) using FASTA. Protein sequence and pattern/motif searches were performed on translated consensus gene sequences using BEAUTY (BLAST-enhanced alignment utility; Worley *et al.*, 1995), querying the combined conserved sequence and annotated sequence databases on the Baylor College of Medicine (BCM) server. Multiple DNA and protein alignments were performed with the MAP program (Huang, 1994) on the BCM server.

Design and testing of polymorphic PCR primers. The following PCR primers were designed flanking a perfect $(CA)_{14}$ repeat ~10 kb downstream of the 3′-end of the *ERCC2* gene: her19-5f1 (GT strand), AGGCCAGGGTAAGTTGGTG; her19-5r2 (CA strand), TCACAGGGTTACACGGACAG. The CA-strand primer was synthesized with a 6-Fam dye phosphoramidite (ABD) on the 5′-end as recommended by the manufacturer. PCR conditions were as follows: 100 ng human genomic DNA (samples from 91 unrelated individuals obtained from the Coriell Institute), 10 pmol each primer, 250 μM each dNTP, 1.5 mM $MgCl_2$, 0.5 U AmpliTaq in a 25-μl reaction volume. Cycling was performed in a Perkin-Elmer Cetus thermal cycler with an initial denaturation of 5 min at 95°C, followed by 26 cycles at 95°C for 1 min, 58°C for 2 min, 72°C for 2 min, and a final extension at 72°C for 10 min. Each sample was diluted to a final volume of 40 μl, and a 1-μl aliquot was combined with 0.5 μl of Gene Scan-2500-Rox (ABD) and 4 μl of deionized formamide, then denatured for 3 min at 90°C prior to loading on a 6% denaturing polyacrylamide gel. Data were collected on an ABD 373A DNA sequencer, and fragment sizes were determined using ABD 672 Gene Scan software. Each allele was determined by differentiation of the major peak from the minor peaks created by polymerase stuttering (Ziegle *et al.*, 1992).

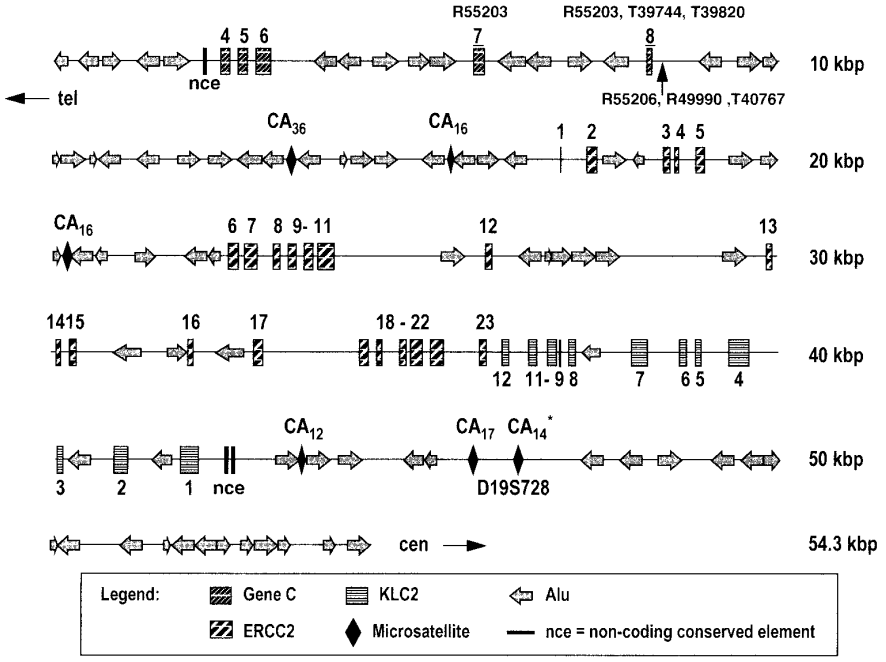

FIG. 2. Genomic structure of the human *ERCC2* gene region. The various sequence elements identified in the 54.3 kb of human sequence encompassing the *ERCC2* locus are shown. The figure is drawn relative to the direction of transcription of *ERCC2,* which is from telomere to centromere. *Alu* elements are represented by the solid arrows, whose direction indicates the orientation (e.g., 5′ → 3′) of the element, and are drawn to scale. Exons from the three identified genes are indicated by oblong patterned boxes as indicated in the figure key. Exon numbers are indicated above (transcription tel. → cen.) or below (transcription cen. → tel.) each box; all exons, except *ERCC2* exon 1 and *KLC2* exon 9, are drawn to scale. Matches to human ESTs were observed at exons 7 and 8 of *Gene C,* which are underlined. The names of the ESTs are noted above the exons (for 5′-end clones) or below the position to which they map (for 3′-end clones). Potentially polymorphic microsatellite repeats are indicated by a solid diamond with the type and number of repeats noted. The repeat indicated as D19S728 has been determined to be polymorphic experimentally. Noncoding sequence elements (nce) conserved between mouse and human are not drawn to scale.

RESULTS AND DISCUSSION

Genomic Structure of the Region

Sequence comparison of the *ERCC2* syntenic gene regions in the human (54.3 kb), mouse (32.6 kb), and hamster (20.6 kb) revealed the presence of three closely linked genes present in the same orientation in all three species, as indicated in Figs. 2–4. In the human, the *ERCC2* gene is transcribed from telomere to centromere, in the same orientation as the nearby *ERCC1* gene (Smeets *et al.,* 1990). Located just proximal to *ERCC2,* and transcribed from centromere to telomere, is a new kinesin light chain gene (*KLC2*). The closest known marker or gene in this syntenic group is *CKMM,* which is located within 25 kb and centromeric of the 3′-end of *ERCC2* (Smeets *et al.,* 1990). The third gene, arbitrarily named *Gene C,* was identified by sequence comparison of the mouse and human cosmids and currently has no known function. Its presence in the hamster cosmid was determined by PCR analysis with conserved primers designed in exons 5, 6, and 7 of the putative ORF (data not shown). Based on these data, the density of genes in this region of the human genome averages one per 18 kb, which is comparable to that observed near the *ERCC1* locus (one gene per 21 kb; Martin-Gallardo *et al.,* 1992).

The repeat sequence density within the regions studied is shown in Figs. 2–4 and in Table 1. In the human, the average density of *Alu* elements exceeds 1.4 *Alu*/kb, constituting 39.2% of the human sequence reported in this work. This density is comparable to that observed in the *ERCC1* region and slightly higher than that seen at the *BCR* or *ABL* loci on human chromosomes 22 and 9, respectively (Chissoe *et al.,* 1995). As can be seen in Fig. 2, these elements are heavily clustered, particularly in the segments upstream of the *ERCC2* gene and downstream of *KLC2.* One cluster of 4.5 contiguous *Alu* elements spans 1.36 kb with no intervening unique sequence. This dense clustering phenomenon has now been observed in numerous genomic regions of human chromosome 19 (unpublished data). No L1, THE, or LTR-like elements were observed in the human, mouse, or hamster *ERCC2* gene regions.

Several potentially polymorphic microsatellite repeats are present in the human sequence (Fig. 2). One CA_{14} repeat ~10 kb proximal to *ERCC2* is polymorphic: eight alleles were identified in DNA samples from 91 unrelated Caucasian individuals with an observed heterozygosity of 0.79. The primer sequences and allele frequencies for this repeat have been deposited in GDB and designated D19S728. No obvious minisatellite sequences were observed in the 54.3 kb of human sequence reported here.

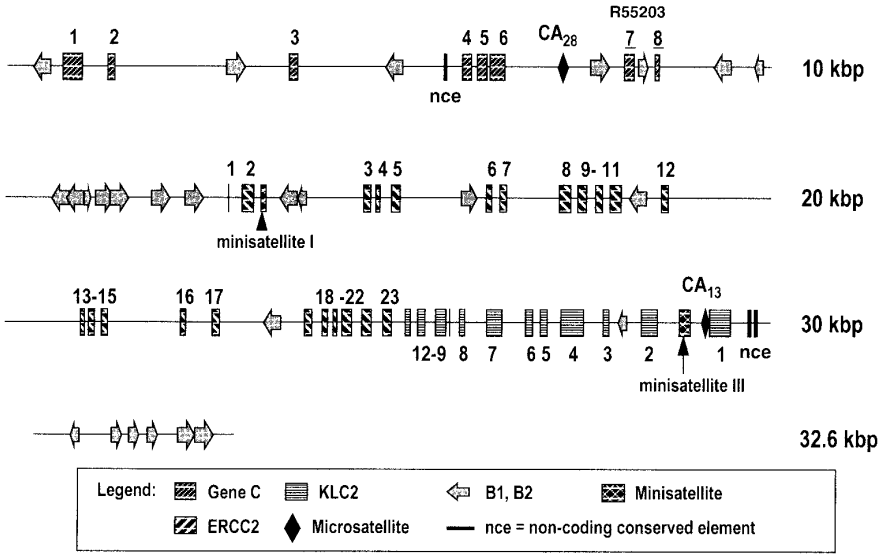

FIG. 3. Genomic structure of the mouse *ERCC2* gene region. The various sequence elements identified in the 32.6 kb of mouse sequence encompassing the *ERCC2* locus are shown. Representation of exons and direction of transcription are as described in the legend to Fig. 2. Exons 7 and 8 of *Gene C* in the mouse, which are underlined, also showed strong similarity to a human breast EST, which is noted above the exons. The retroposable B1 and B2 elements are indicated with solid arrows, whose direction indicates the orientation (e.g., 5′ → 3′) of the element, and are drawn to scale. In addition to microsatellite repeats, several different tandem repeats (see key) are indicated by number (minisatellites I–III) as in Table 1.

As depicted in Figs. 3 and 4, the repeat density of retroposable elements in the mouse and hamster *ERCC2* gene regions is much lower than that of the human, with an average of 0.7 B1, B2 elements per kilobase for the mouse (11% of the reported sequence), and 0.4 per kilobase for the hamster (5.9%). Interestingly, these elements are found clustered in the mouse in positions similar to those occupied by clusters of *Alu*s in the human.

We identified three unrelated minisatellites in the hamster and two in the mouse *ERCC2* regions, as indicated in Table 1. While the 191- and 63-bp elements appear to be novel, the 53-bp element is present in introns at various rodent loci, including the mouse *XRCC1* DNA repair gene, the zona pellucida glycoprotein ZP3 gene, the urokinase-type plasminogen activator locus, and the rat tropoelastin locus.

ERCC2 Region: Genomic Structure

The *ERCC2* gene is composed of 23 exons (Frederick *et al.*, 1994; Kirchner *et al.*, 1994; Della Coletta *et al.*, 1995) that span a genomic distance of 18.9 kb in the human, 12.3 kb in the mouse, 14.1 kb in the hamster, and 14.4 kb in the fish, *Xiphophorus maculatus* (Della

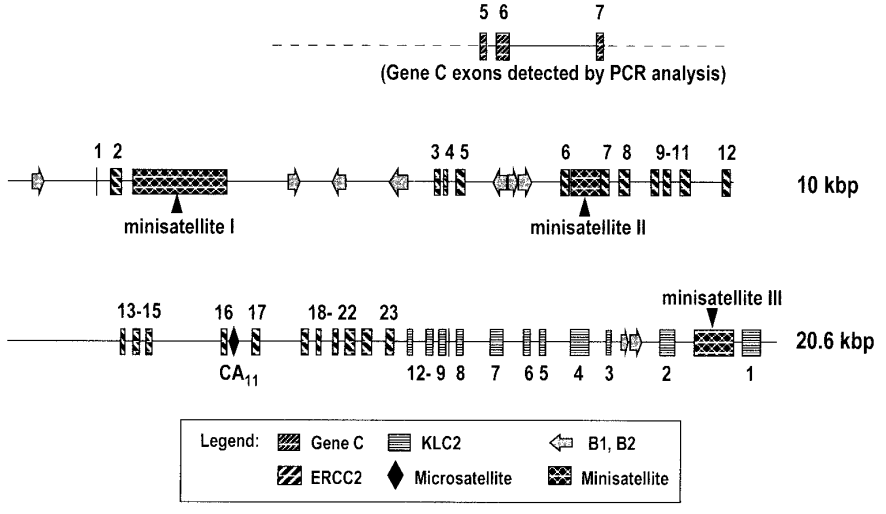

FIG. 4. Genomic structure of the hamster *ERCC2* gene region. The various sequence elements identified in the 20.6 kb of hamster sequence encompassing the *ERCC2* locus are shown. The solid lines indicate regions for which absolute distance is known, by either sequence analysis or PCR. The distance between *Gene C* and *ERCC2* (indicated by the dashed line) in the hamster is unknown. Refer to the legends of Figs. 2 and 3 for further description.

TABLE 1

Repeat Elements in the Human, Mouse, and Hamster ERCC2 Gene Regions

	Human[a]	Mouse	Hamster
Alu or B1/B2	74 (39.2%)	26 (11%)	9 (5.9%)
$CA_{(>10)}$	6 (0.4%)	2 (0.25%)	1 (0.1%)
Minisatellite I (191 bp)	—	0.33 (0.19%)	7 (6.5%)
Minisatellite II (63 bp)	—	—	7 (2.1%)
Minisatellite III (53 bp)	—	3 (0.49%)	8 (2.1%)
Total % repetitive:	39.6%	12.1%	16.7%

[a] Number of elements identified per cosmid (percentage of cosmid containing element).

TABLE 3

Levels of Conservation of the ERCC2 Protein from Four Vertebrate Species

	Hamster[a]	Mouse	Fish
Human	98% (99%)	97% (99%)	83% (94%)
Hamster	—	99% (99.8%)	84% (94%)
Mouse	—	—	84% (94%)

[a] Numbers are expressed as percentage identity (percentage similarity).

Coletta *et al.*, 1995), although the region required for gene function may be larger. Exon lengths are completely conserved in all four species, as shown in Table 2. The hamster gene exhibits the highest nucleotide identity to the human gene (90.1%) compared to its identity the mouse (89.8%) or the fish (76%). In contrast to the conserved exon lengths, intron lengths vary considerably (Table 2). The largest intron of the four species, ~4.6 kb, occurs in the hamster genomic sequence (intron 2) and is due to the expansion of the 191-bp tandem repeat. There appears to be little concordance in intron lengths between the mammalian species and the fish.

The *ERCC2* gene product encodes a 760 amino acid protein with ATP-dependent DNA helicase activity (Sung *et al.*, 1993). The levels of residue identity among the predicted proteins from the four species are shown in Table 3. The vertebrate *ERCC2* proteins exhibit significant identity to the *Saccharomyces cerevisiae* RAD3 and *Schizosaccharomyces pombe* RAD15 proteins (Weber *et al.*, 1990; Kirchner *et al.*, 1994; Della Coletta *et al.*, 1995), one of which has been shown experimentally to encode an ATP-dependent DNA helicase (Sung *et al.*, 1987). The three mammalian *ERCC2* proteins are completely conserved in the regions containing motifs typical of the functional domains of a superfamily of known and putative DNA and RNA helicases (Gorbalenya *et al.*, 1989). The fish *ERCC2* protein contains three nonconservative amino acid substitutions within one of these domains, two of which correspond to divergent amino acids between the mammalian proteins and the yeast homologs (data not shown).

Conservation of Noncoding Elements in the ERCC2 Gene Region

A multiple sequence alignment of the 5′-flanking region of *ERCC2*, including exon 1, in the three mamma-

TABLE 2

Intron/Exon Structure of the Fish, Hamster, Mouse, and Human *ERCC2* Gene Regions

Exon length (bp)		Intron length (bp)				
Exon	All 4 species	Intron	Fish[a]	Hamster	Mouse	Human
1	5					
2	100	1	129	201	218	302
3	78	2	930	~4600	1576	985
4	63	3	83	76	80	77
5	114	4	124	140	137	186
6	117	5	735	1301	1304	3471
7	117	6	1000	462	90	87
8	124	7	1142	342	756	290
9	97	8	72	91	101	92
10	134	9	771	85	84	115
11	169	10	944	70	82	74
12	119	11	1290	437	555	2100
13	70	12	103	1614	1840	3824
14	70	13	120	77	72	86
15	102	14	1129	81	80	102
16	64	15	76	909	1018	1541
17	122	16	1061	363	359	813
18	93	17	75	617	1168	1395
19	73	18	90	90	91	86
20	71	19	83	151	97	266
21	144	20	434	92	75	96
22	144	21	616	108	119	153
23	93	22	1162	185	169	491

[a] Fish intron sizes from Della Coletta *et al.* (1995).

```
      1              15 16             30 31             45 46             60 61             75
1 mus TTTCTCTGTTATTCT CTCTTAAACACTATT TTCACGACTTGGACA ACCAGAACAGAACCT GGAGGAGCGAAGCAG    75
2 cho CCAAGCCTCCAAGCT CTCCGTTACCCTCCC CTCCCCTCT---CCT CTCGGTCCTCACCGG GACAGCCCGGAAGAG   72
3 hum CTGAGACGATTTTTT TTTTTTTTTGAGACG AGGTGGAGGGGCGGG GGTTCTCGCTATCTT GCTCAAGCTGATCTC   75
                   pyrimidine-rich stretch "B"

      76             90 91            105 106            120 121            135 136            150
1 mus CAATGAACTCTATCG ATTCTCTCGACTCTC TTCTTTCTGTGTTCT GTT-CGAAGT---CC TCGGCGATCCTCA-G  145
2 cho GTGCGAGCCGGATCG ACCTCTATCGGCTCTC TTCCTTCCACGCGCT CCTGCGCCGT---CC TCCACGATCCTCAAG  144
3 hum GAACTCCTGGGTTCG AGTCTCTCGGCTCTT TCCCTTCCATGTTTT CTTTTTGATTGGCCC TCGACGATCCTCA-G  149
                            pyrimidine-rich stretch "A"

      151          165 166          180 181          195 196          210 211          225
1 mus CGATGCCTCGGGCAC CGCTTCTCTCCGA-A GT--CCGCCCTCTCA ccgccatgcgcaCGC GCCAATTTTCCGATT  217
2 cho CGATGCCTCGGGCCC CGCCTCCCCCCAAGA GC--CCGCCCTCCCG ccgccacgcgcaGGC GCGCGTCTTCCGATT  217
3 hum TGACGCCTCCCGCAC CGCCTCACCCGAGAG TCAGCCGCCCTCGCT TTTCCGTGCGCACGC GC-AGTATCCCGATT  223
                                       GC-box                    reverse CAAT box->

      226          240 241          255 256          270 271          285 286          300
1 mus GGC-----------A GAGTGACGGGAGGAC TCGGCCAATGGCGTC CCGAGGAAGGTGTGT CTCTGGGTCACGCC-  280
2 cho GGC-----------G GAGTGACGGGCAGGC TCGGCCAATAGCGTC GCGAGGAAGGCGGGG CTCCAGGCCCAGGC-  280
3 hum GGCTCTGCCCTAGCG GATTGACGGGCAGGT T-AGCCAATGGTCTC GtaatataGGTGGAG CGAGCCCTCGAGGAT  297
                                    forward CAAT box

      301          315 316          330 331          345 346          360 361        374
1 mus --CCCTC-------- ----GTTGAATATTC AGGAG-CGGGCGTGT GGACGCCGCGA---- -----TCAGCCGCC   330
2 cho --CCCTCCCCTCCTC CCTCGTTGAATATTC AAGAGGCGGGCGAGC GGACGCCGCGG---- -----GAAGACGCC   343
3 hum GTCCACGACCCGGCC TCTCGCTGAATATTC ATGAGGGAGGCGGGT CGACCCCGCTGCACA GTCCGGCCGGCGCC   371

      375          390
1 mus ATGAAGTGA    339
2 cho ATGAAGTGA    352
3 hum ATGAAGTGA    380
      exon 1
```

FIG. 5. Multiple alignment of the *ERCC2* promoter regions in mouse (339 nt), hamster (352 nt), and human (380 nt), including exon 1 aligned using the MAP program (Huang, 1994). Conserved canonical promoter elements in each species are underlined, such as the forward and reverse CAAT boxes and the GC box. A putative TATA box in the human sequence is indicated in lowercase letters at base 272 of the alignment. A putative α-Pal transcription factor binding site conserved in the rodents (but not in human) is indicated in lowercase letters at base 196 of the alignment. Two pyrimidine-rich stretches of sequence upstream of the promoter elements are indicated in italics as "stretch A" and "stretch B" and underlined.

lian species is shown in Fig. 5. Overall, the hamster (352 nt) and mouse (339 nt) 5′-flanking regions share 56.2 and 58.0% nucleotide identity, respectively, with the 380 nt of the putative human promoter. The rodent 5′-flanking regions are more similar to each other, sharing 67.1% identity over this region. Several classical eukaryotic promoter elements (Rice *et al.*, 1991) are completely conserved in all three species, including a reverse CAAT box (CCGATTGC) and a GC box (CCGCCC). A forward CAAT box also appears to be present in both rodent species (GGCCAATRG) as well as in the human (AGCCAATGG). A putative α-Pal transcription factor binding site previously identified in the hamster (Kirchner *et al.*, 1994) appears to be conserved in the mouse but not in the human sequence. A potential TATA box previously identified in the human 5′-flanking sequence (Weber *et al.*, 1990) is not present in the rodents. Although the fish *ERCC2* sequence also contains similar putative eukaryotic promoter elements (Della Coletta *et al.*, 1995), the locations and sequence context surrounding these elements are significantly diverged in the fish 5′-flanking region relative to that of the mammalian species, prohibiting its inclusion in the multiple alignment.

A pyrimidine-rich stretch of DNA (34 nt, 88% C + T) upstream of the canonical promoter elements was previously reported in human (Weber *et al.*, 1990) and is shown as "stretch A" in Fig. 5. The same regions in mouse and hamster are also enriched for pyrimidines (mouse: 34 nt, 82% C + T; hamster: 34 nt, 79.4% C + T). In addition, the hamster has another pyrimidine-rich stretch approximately 30 nt upstream, labeled "stretch B" in Fig. 5, which is 88% pyrimidine-rich over 42 nt (Kirchner *et al.*, 1994). "Stretch B" is also enriched for pyrimidines in the human and mouse 5′-flanking regions. A 30-nt CT-rich stretch (76.6%) located 419 bases upstream of the first exon is also observed in the fish *ERCC2* region. Similar pyrimidine-rich sequences have been identified in the 5′-flanking regions of the mouse and human *ERCC1* DNA repair genes (van Duin *et al.*, 1988). This region is postulated to play a role in the expression of these genes (Weber *et al.*, 1990), since its absence in a human genomic *ERCC2* clone containing all other promoter elements provided only transient UV resistance to transfected UV5 (repair-deficient) cells. In contrast, full correction was provided by clones containing the pyrimidine-rich region (Weber *et al.*, 1990).

The polyadenylation signal for the human *ERCC2* gene was previously identified within the 70 bp following the stop codon (Weber *et al.*, 1990). However, no consensus polyadenylation signal was identified in the

662 bp downstream of the stop codon in the hamster *ERCC2* sequence originally published by Kirchner *et al.* (1994). With the addition of ~4.7 kb of hamster sequence from this work, a near consensus polyadenylation signal (AATTAAA) has been identified ~1.2 kb distal to the stop codon, within intron 6 of the new KLC gene. A consensus polyadenylation signal (AATAAA) is found in the same intron in the mouse sequence, also ~1.2 kb distal to the stop codon of *ERCC2*. If this is the true polyadenylation site, and presuming similar transcription initiation sites, the rodent messages should be ~1 kb larger than the human. Transcript size is not known for the hamster gene, but the *ERCC2* transcript size in a mouse leukemia cell line (Sheibani and Eastman, 1990) was reported to be 4.1 kb, compared to the ~2.6-kb transcript in humans. If one allows for polyadenylation, this agrees fairly well with the predicted size as determined by sequence analysis of the mouse cosmid.

In contrast to the *XRCC1* locus (Lamerdin *et al.*, 1995), we did not observe any noncoding elements that were conserved in all three species within the span of the *ERCC2* gene other than the promoter region. At the *XRCC1* locus, these conserved elements were posited as potential regulatory elements for expression of the *XRCC1* gene. The lack of such elements at the *ERCC2* locus is not entirely unexpected, as a gene whose product is involved in an essential role such as transcription-coupled DNA repair should not require elaborate transcriptional control mechanisms for expression. Conservation of noncoding regions between mouse and hamster was detected in numerous introns of the *ERCC2* and *KLC2* genes (73–95% identity in regions 50 bp or larger), as well as in the 115 bp immediately upstream of the *KLC2* gene. Several introns are conserved virtually in their entirety (*ERCC2:* introns 3, 4, 10, 13, 18, and 21; *KLC2:* introns 7 and 8). There is also one partial B1 element in intron 2 of the *KLC2* gene that is conserved, in addition to several of the tandem repeats described above. The level of nucleotide identity observed in the noncoding regions is intriguing since the evolutionary distance between these two rodent species is nearly equivalent to that observed between mouse and human (~60 My).

KLC2 Region

Sequence comparison of all three mammalian species revealed the presence of 12 conserved exons <300 bp proximal to the *ERCC2* gene and in the opposite orientation. XGRAIL 1.2 predicted 10 of these in the human (as well as one alternate exon), 8 in the hamster, and 11 in the mouse sequence. The predicted human gene is 1512 bp and is 85% identical to both of the 1515-bp mouse and hamster genes. Translation of the putative human gene produces a 504 amino acid protein that has significant identity to the kinesin light chain (KLC) protein in human, rat, *Strongylocentrotus purpuratus* (SP, purple sea urchin), and other species. As shown in Fig. 6, the strongest residue identity (72%) to the *KLC2* gene product is observed in the conserved central domains of the sea urchin (SPKLC, Q05090; amino acids 209–410) and human (KLC1, Q07866; amino acids 197–404) KLC homologs. Conservation of this region, which is composed of a ~40-residue repeating structure (Cyr *et al.*, 1991; Cabeza-Arvelaiz *et al.*, 1993), is a hallmark of all KLCs identified to date. This repeating structure may play a role in proper folding of the protein and is believed to be essential for KLC function.

Slightly lower residue identity to the human and sea urchin KLCs was observed across the amino-terminal end of the protein (particularly residues 16–160), which constitutes the rod domain (Cabeza-Arvelaiz *et al.*, 1993). This region is indicated in Fig. 6 by a shaded box. This portion of the KLC protein has been demonstrated to interact with the tail of the kinesin heavy chain (KHC) in *Drosophila* (Gauger and Goldstein, 1993) through the interaction of coiled-coil domains. The level of identity (53.1%) in this region of the two human KLCs (KLC2 vs KLC1) is similar to that seen between very different species: e.g., the sea urchin and rat KLCs (Wedaman *et al.*, 1993) and the *Drosophila* and rat KLCs (Gauger and Goldstein, 1993). In addition, the previously identified human (Cabeza-Arvelaiz *et al.*, 1993) and rat (Cyr *et al.*, 1991) KLC isoforms appear to be true homologs, sharing up to 97% residue identity. This suggests that the human, hamster, and mouse *KLC2* homologs identified in this work represent a new member of this gene family.

Although the length of the carboxy terminus of the KLC2 protein is considerably shorter than KLC1 and SPKLC, portions of it share significant identity with one of the SPKLC isoforms (Q05090) primarily through conservation of charged amino acids, as indicated in Fig. 6 by the lightly shaded boxes at the bottom of the figure. It is interesting to note that KLC1 shares little similarity with either KLC2 or SPKLC at the C-terminus. The carboxy-terminal end of the KLC is thought to contribute to the kinesin molecule's fan-like tail (Scholey *et al.*, 1989; Cyr *et al.*, 1991) and is responsible for the binding of specific membrane-bound organelles. This end of the protein is the most diverse in all of the isoforms detected thus far, and some of the diversity has been demonstrated in the rat and sea urchin to be due to alternative splicing of the same gene (Cyr *et al.*, 1991; Wedaman *et al.*, 1993). Our identification of a second human gene may account for additional human isoforms, although primary transcripts for *KLC2* will need to be isolated to determine whether it is alternatively spliced. The striking differences in the C-terminus ends of KLC1 and KLC2 imply that the two gene products may bind to different substrates and thus could play very different roles in intracellular transport.

Analysis of the sequence 5′ of the first *KLC2* exon revealed no canonical CAAT or TATA elements in human, mouse, or hamster. Similarly, we detected no

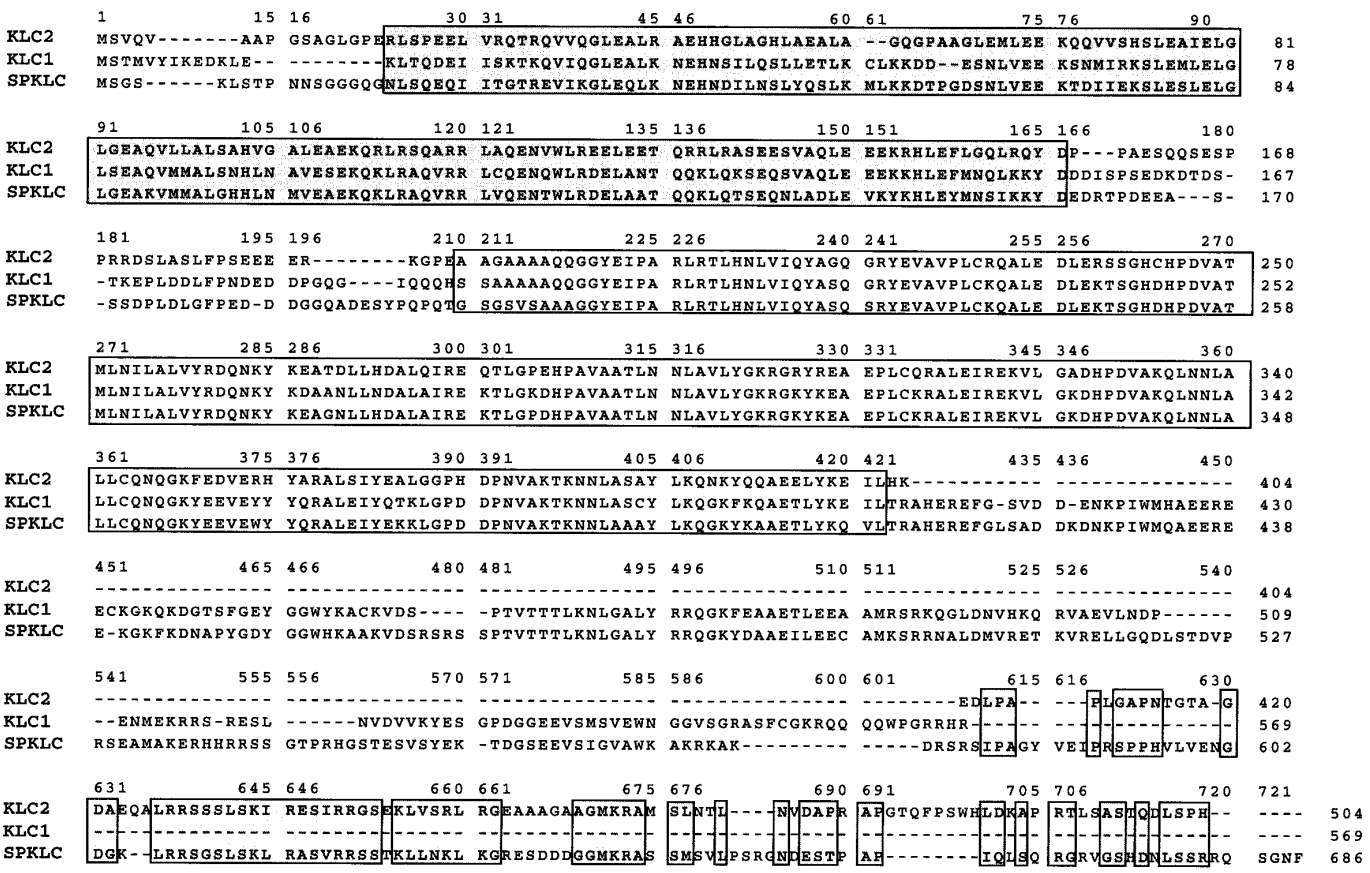

FIG. 6. Multiple alignment of the kinesin light chain proteins from human chromosome 19 (KLC2; 504 aa), human chromosome 14 (KLC1; 569 aa), and the sea urchin (SPKLC; 686 aa) aligned using the MAP program (Huang, 1994). The region of highest conservation among these three proteins is indicated with an open box, from positions 210–422 of the alignment. Residues indicated by the darker shaded box at the top (positions 24–166 of the alignment) interact with the stalk of the kinesin heavy chain (KHC). The C-terminus of the newly identified KLC2 shares significant identity with portions of the sea urchin C-terminus shown here, as indicated by the lighter shaded boxes at the bottom.

such elements in the ~276 bp of 5′-flanking sequence from the *KLC1* locus. Two sequence elements are conserved (45 bp, 75% identity; 56 bp, 87% identity) in the human and mouse ~250 bp upstream of the first *KLC2* exon. The second element contains two putative GC boxes (CCGCCC) spaced 9 bp apart, but it has not been demonstrated that they are involved in controlling expression of this gene. It is also not known if these elements are conserved in the hamster, as the sequence obtained in this work extends only 156 bp proximal to the first hamster *KLC2* exon. A similar lack of classical promoter elements was observed at the human and mouse *ERCC1* locus, which is expressed in all tissues and at all stages of development (van Duin *et al.*, 1988). The lack of promoter elements at the *KLC1* locus appears to be consistent with the observation that the *KLC1* gene may be constitutively expressed (Cabeza-Arvelaiz *et al.*, 1993) and may point to a similar mode of expression for the *KLC2* gene, although this remains to be determined experimentally.

Interestingly, analysis of the remaining 1.6 kb of sequence downstream of the *ERCC2* gene in the fish yielded no sequence or residue similarity to the mammalian *KLC2* gene. This is intriguing since the *ERCC2* and *CKMM* loci are believed to be tightly linked in the fish (Walter *et al.*, 1991). In the human, the two genes are within 25 kb of each other (Smeets *et al.*, 1991), and the *KLC2* gene falls between them. A similar physical distance between *ERCC2* and *CKMM* has been observed in the mouse by one of us (L.S.). If this region originated in a common vertebrate ancestor, the absence of tight physical linkage of a fish *KLC* to *ERCC2* would imply that (1) the fish homolog of the *KLC2* gene is present in this linkage group, but physically closer to *CKMM* than to *ERCC2*, (2) the fish homolog exists in a different linkage group, perhaps due to a chromosomal rearrangement event, or (3) the mammalian homologs have arisen since the divergence of fish and mammals approximately 400 million years ago.

Gene C

Analysis of the remaining 12.5 kb of sequence upstream of the mouse *ERCC2* gene with XGRAIL 1.2 predicted the presence of nine exons with marginal to excellent coding potential. Comparative sequence analysis of the remaining 17 kb of human sequence te-

lomeric of *ERCC2* indicated that four of these exons were conserved and two were false positives. One noncoding conserved element and one additional exon representing the end of the ORF were also detected. The beginning of the ORF in the mouse is indicated in Fig. 3 as exon 1. The human exons are numbered relative to the homologous exons in the mouse.

Multiple EST matches provide indirect evidence that *Gene C* is expressed in several tissues. Exons 7 and 8 in the human share 99.4 and 97.1% identity, respectively, with the 5′-end of a cDNA represented by human breast EST, R55203. The mouse exon 7 is 86.3% identical to the same EST. Human exon 8 also exhibits similar nucleotide identity to 5′-end ESTs from fetal spleen (T39744) and a normalized fetal liver/spleen library (T39820). The 3′-ends of these particular cDNA clones did not match any of the 54.3 kb of sequence reported here, but the 3′-ends of clones represented by ESTs R55206 and R49990 (95 and 95.2% identity, respectively) from human breast and T40767 (98.5% identity) from fetal spleen, all from oligo(dT)-primed libraries, mapped just distal to the end of the ORF, further defining the 3′-end of this gene. In addition, a classical polyadenylation signal is present in the human ~20 bp 5′ to the start of sequence identity with two of the three 3′-end ESTs, which begins following the poly(A) tail. All EST matches to this sequence were from clones sequenced by the WashU-Merck EST Project (Hillier *et al.*, 1995).

The human and mouse homologs of *Gene C* exhibit 84% nucleotide identity over the 666 bp constituting the portion of the gene present in the human. *Gene C* in the mouse encodes a 376-aa protein that is 89.6% identical to the human homolog over 223 aa; most of the variability is at the C-terminus. The central portion of the predicted protein (amino acids 185–270) contains a motif found in several proteins, including a potassium channel transport protein, AKT1 (61% similarity over 58 residues and 45% similarity over a second 79-aa region of the AKT1 protein), and a nonerythroid ankyrin protein expressed in brain (53% similarity over 79 residues). The region of similarity to brain ankyrin is composed of a 33 amino acid repeat domain also present in proteins involved in cell differentiation, cell cycle control, and transcription (Otto *et al.*, 1991; Tse *et al.*, 1991). Indeed, the *Gene C* protein motif also exhibits similarity to regions of the human NF-κB transcription factor (58% similarity over 59 residues) and the *Caenorhabditis elegans* sex determining protein, FEM-1 (50% similarity over 59 aa). This core 33-aa repeat is contained within the region of ankyrin involved in membrane binding and thus may play a role in protein–protein interactions.

Due to the small size of the predicted protein, the presence of a consensus splice junction at the start of "exon 1," and the lack of any canonical promoter elements in the 265 bp upstream, it is unlikely that the entire coding region of *Gene C* is present in the mouse sequence described here. Future comparative sequence analysis extending 5′ of this locus will allow us to elucidate the true structure of this gene and potentially provide additional clues to its function.

We have described the comparative analysis of 108 kb of sequence encompassing the human, mouse, and hamster *ERCC2* gene regions; a portion of the hamster sequence has been previously described (Kirchner *et al.*, 1994). This region is highly conserved over evolutionary time, with the order and orientation of at least three genes preserved in the mammalian lineage within this linkage group. The level of nucleotide sequence conservation within two of these genes is quite high: *ERCC2* is ~90% and *KLC2* is ~85% between the human and rodent homologs. It seems plausible that the strong conservation of these two genes reflects high selective pressure operating on these essential genes.

The *KLC2* gene identified in this work is now the third distinct human *KLC* gene in GenBank (however, one contains only promoter and exon 1 sequence) and the only one for which the genomic structure of the gene has been determined. Future comparison of this genomic region with those containing other *KLC* genes may shed some light on the prevalence of alternative splicing of these genes as the primary mechanism responsible for generating diverse KLC isoforms or whether a larger gene "family" may exist (or a combination of the two). The latter seems plausible given the ever-growing number of KHCs and kinesin-like proteins (KLPs) in the literature and the diverse roles kinesin plays in cellular processes (e.g., secretion, endocytosis, and axonal transport). We have already identified the 3′-end of another *KLC* gene linked to a different DNA repair gene (unpublished data) on human chromosome 14 that also appears to be distinct from the *KLC1* gene provisionally mapped by the Cabeza-Arvelaiz group to the same chromosome. It is interesting, given the level of C-terminal diversity of known KLCs from a variety of species, that the *KLC2* gene product shares significant similarity to SPKLC at the C-terminus. It is tempting to speculate that these two proteins may share binding affinity for a similar substrate (and different from that of KLC1), e.g., a specific receptor on a membrane-bound organelle. The sequence depicted here provides the molecular tools to aid in the isolation of full-length cDNAs to determine the expression patterns of this gene and to pinpoint the regions of the molecule required for interaction with membrane-bound organelles. These clones may also help us to elucidate how the KLC protein recognizes specific "cargo" molecules and to delineate its role in specific cellular functions.

Comparative genomic sequence analysis between mouse and human is a powerful tool for the identification of both coding and noncoding conserved elements. Knowledge of the coding regions in regions of synteny in multiple species provides more complete information with which to ascertain the structure of a gene or closely linked genes. Furthermore, comparative analysis allows the identification of putative regulatory and

structural elements that are conserved through evolutionary time. These elements are often quite small and are less likely to be identified by existing computer algorithms. Further comparative sequence analysis within the ~2-Mb region between *XRCC1* and *ERCC2* will provide a clearer picture of its genomic structure (e.g., gene content and repeat density) and determine the prevalence of conserved regulatory elements.

ACKNOWLEDGMENTS

The authors thank Dr. Christine Weber for providing the pCXPD-3 cosmid and helpful discussions relating to this work and Dr. Elbert Branscomb for critical reading of this manuscript. We also thank Bill Dunn, Aaron Adamson, Mishelle Montgomery, and Subha Basu for excellent technical assistance. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48 and under Contract DE-AC0584OR21400 with Lockheed-Martin Energy Systems, Inc., at the Oak Ridge National Laboratory.

REFERENCES

Barnes, D. E., Kodama, K.-I., Tynan, K., Trask, B. J., Christensen, M., DeJong, P. J., Spurr, N. K., Lindahl, T., and Mohrenweiser, H. W. (1992). Assignment of the gene encoding DNA Ligase I to human chromosome 19q13.2-13.3. *Genomics* 12: 164–166.

Bootsma, D., and Hoeijmakers, J. H. J. (1994). The molecular basis of nucleotide excision repair syndromes. *Mutat. Res.* 307: 15–23.

Brilliant, M. H., Williams, R. W., Conit, C. J., Angel, J. M., Oakey, R. J., and Holdener, B. C. (1994). Mouse Chromosome 7. *Mamm. Genome* 5: S104–S123.

Broughton, B. C., Steingrimsdottir, H., Weber, C. A., and Lehmann, A. R. (1994). Mutations in the xeroderma pigmentosum group D DNA repair/transcription gene in patients with trichothiodystrophy. *Nature Genet.* 7: 189–194.

Cabeza-Arvelaiz, Y., Shih, L.-C. N., Hardman, N., Asselbergs, F., Bilbe, G., Schmitz, A., White, B., Siciliano, M. J., and Lachman, L. B. (1993). Cloning and genetic characterization of the human kinesin light-chain (KLC) gene. *DNA Cell Biol.* 12: 881–892.

Chissoe, S. L., Bodenteich, A., Wang, Y.-F., Wang, Y.-P., Burian, D., Clifton, S. W., Crabtree, J., Freeman, A., Iyer, K., Jian, L., Ma, Y., McLaury, H.-J., Pan, H.-Q., Sarhan, O. H., Toth, S., Wang, Z., Zhang, G., Heisterkamp, N., Groffen, J., and Roe, B. A. (1995). Sequence analysis of the human *ABL* gene, the *BCR* gene, and the regions involved in the Philadelphia chromosome translocation. *Genomics* 27: 67–82.

Cyr, J. L., Pfister, K. K., Bloom, G. S., Slaughter, C. A., and Brady, S. T. (1991). Molecular genetics of kinesin light chains: Generation of isoforms by alternative splicing. *Proc. Natl. Acad. Sci. USA* 88: 10114–10118.

de Jong, P. J., Yokobata, K., Chen, C., Lohman, F., Pederson, L., McNinch, J., and Van Dilla, M. (1989). Human chromosome-specific partial digest libraries in λ and cosmid vectors. *Cytogenet. Cell Genet.* 51: 985.

Della Coletta, L., Rolig, R. L., Fossey, S., Morizot, D. C., Nairn, R. S., and Walter, R. B. (1995). Characterization of the *Xiphophorus* Fish (Teleoster: Poeciliidae) *ERCC2/XPD* locus. *Genomics* 26: 70–76.

Drapkin, R., Reardon, J. T., Ansari, A., Huang, J. C., Zawel, L., Ahn, K., Sancar, A., and Reinberg, D. (1994). Dual role of TFIIH in excision repair and in transcription by RNA polymerase II. *Nature* 368: 769–772.

Ehrich, E., Craig, A., Poutska, A., Frischauf, A.-M., and Lehrach, H. (1987). A family of cosmid vectors with the multi-copy R6K replication origin. *Gene* 57: 229–237.

Frederick, G. D., Amirkhan, R. H., Schultz, R. A., and Freidberg, E. C. (1994). Structural and mutational analysis of the xeroderma pigmentosum group D (XPD) gene. *Hum. Mol. Genet.* 3: 1783–1788.

Gauger, A. K., and Goldstein, L. S. (1993). The Drosophila kinesin light chain. Primary structure and interaction with kinesin heavy chain. *J. Biol. Chem.* 268: 13657–13666.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P., and Blinov, V. M. (1989). Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. *Nucleic Acids Res.* 17: 4713–4730.

Hardison, R., and Miller, W. (1993). Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Mol. Biol. Evol.* 10: 73–102.

Hillier, L., Clark, N., Dubuque, T., Elliston, K., Hawkins, M., Holman, M., Hultman, M., Kucaba, T., Le, M., Lennon, G., Marra, M., Parsons, J., Rifkin, L., Rohlfing, T., Soares, M., Tan, F., Trevaskis, E., Waterston, R., Williamson, A., Wohldmann, P., and Wilson, R. (1995). The WashU-Merck EST Project.

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

Jurka, J., Walichiewicz, J., and Milosavljevic, A. (1992). Prototypic sequences for human repetitive DNA. *J. Mol. Evol.* 35: 286–291.

Kirchner, J. M., Salazar, E. P., Lamerdin, J. E., Montgomery, M. A., Carrano, A. V., and Weber, C. A. (1994). Cloning and molecular characterization of the chinese hamster nucleotide excision repair gene homologous to the human *ERCC2* gene. *Genomics* 23: 592–599.

Lamerdin, J. E., Montgomery, M. A., Stilwagen, S. A., Scheidecker, L., Tebbs, R. S., Thompson, L. H., and Carrano, A. V. (1995). Genomic sequence comparison of the human and mouse *XRCC1* DNA repair gene regions. *Genomics* 25: 547–554.

Martin-Gallardo, A., McCombie, W. R., Gocayne, J. D., FitzGerald, M. G., Wallace, S., Lee, B. M. B., Lamerdin, J., Trapp, S., Kelley, J. M., Liu, L-I., Dubnick, M., Johnston-Dow, L. A., Kerlavage, A. R., deJong, P., Carrano, A., Fields, C., and Venter, J. C. (1992). Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nature Genet.* 1: 34–39.

Martin-Gallardo, A., Lamerdin, J., and Carrano, A. V. (1994). Shotgun sequencing. *In* "Automated DNA Sequencing & Analysis" (M. Adams, C. Fields, and J. C. Venter, Eds.), pp. 37–41, Academic Press, London.

Mohrenweiser, H. W., Carrano, A. V., Fertitta, A., Perry, B., Thompson, L. H., Tucker, J. D., and Weber, C. A. (1989). Refined mapping of the three DNA repair genes. *ERCC1, ERCC2,* and *XRCC1,* on human chromosome 19. *Cytogenet. Cell Genet.* 52: 11–14.

Muzny, D. M., Richards, S., Shen, Y., and Gibbs, R. A. (1994). PCR based strategies for gap closure in large-scale sequencing projects. *In* "Automated DNA Sequencing & Analysis" (M. Adams, C. Fields, and J. C. Venter, Eds.), pp. 182–190, Academic Press, London.

Otto, E., Kunimoto, M., McLaughlin, T., and Bennett, V. (1991). Isolation and characterization of cDNAs encoding human brain ankyrins reveal a family of alternatively spliced genes. *J. Cell Biol.* 114: 241–253.

Rice, P. M., Elliston, K., and Gribskov, M. (1991). DNA. *In* "Sequence Analysis Primer" (M. Gribskov and J. Devereux, Eds.), pp. 1–59, Stockton Press, New York.

Schaeffer, L., Moncollin, V., Roy, R., Staub, A., Mezzina, M., Sarasin, A., Weeda, G., Hoeijmakers, J. H., and Egly, J. M. (1994). The *ERCC2*/DNA repair protein is associated with the class II BTF2/TFIIH transcription factor. *EMBO J.* 13: 2388–2392.

Scholey, J. M., Heuser, J., Yang, J. T., and Goldstein, L. S. B. (1989). Identification of globular mechanochemical heads of kinesin. *Nature* 338: 355–357.

Sheibani, N., and Eastman, A. (1990). Analysis of various mRNA potentially involved in cisplatin resistance of murine leukemia L1210 cells. *Cancer Lett.* 52: 179–185.

Smeets, H., Bachinski, L., Coerwinkel, M., Schepens, J., Hoeijmakers, J., van Duin, M., Grzeschik, K.-H., Weber, C. A., de Jong, P., Siciliano, M. J., and Wieringa, B. (1990). A long-range restriction map of the human chromosome 19q13 region: Close physical linkage between *CKMM* and the *ERCC1* and *ERCC2* genes. *Am. J. Hum. Genet.* 46: 492–501.

Sung, P., Prakash, L., Matson, S. W., and Prakash, S. (1987). RAD3 protein of *Saccharomyces cerevisiae* is a DNA helicase. *Proc. Natl. Acad. Sci. USA* 84: 8951–8955.

Sung, P., Bailly, V., Weber, C., Thompson, L. H., Prakash, L., and Prakash, S. (1993). Human xeroderma pigmentosum group D gene encodes a DNA helicase. *Nature* 365: 852–855.

Takayama, K., Salazar, E. P., Lehmann, A., Stefanini, M., Thompson, L. H., and Weber, C. A. (1995). Defects in the DNA repair and transcription gene *ERCC2* in the cancer-prone disorder xeroderma pigmentosum group D. *Cancer Res.* 55: 5656–5663.

Thompson, L. H., Bachinski, L. L., Stallings, R. L., Dolf, G., Weber, C. A., Westerveld, A., and Siciliano, M. J. (1989). Complementation of repair gene mutations on the hemizygous chromosome 9 in CHO: A third repair gene on human chromosome 19. *Genomics* 5: 670–679.

Tse, W. T., Menninger, J. C., Yang-Feng, T. L., Francke, U., Sahr, K. E., Lux, S. E., Ward, D. C., and Forget, B. G. (1991). Isolation and chromosomal localization of a novel non-erythroid ankyrin gene. *Genomics* 10: 858–866.

Uberbacher, E. C., and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* 88: 11261–11265.

van Duin, M., van den Tol, J., Warmerdam, P., Odijk, H., Meijer, D., Westerveld, A., Bootsma, D., and Hoeijmakers, J. H. J. (1988). Evolution and mutagenesis of the mammalian excision repair gene. *ERCC-1. Nucleic Acids Res.* 16: 5305–5322.

Walter, R. B., Harless, J., Svensson, R. T., Kallman, K. D., Morizot, D. C., and Nairn, R. S. (1991). Linkage assignment of a DNA sequence (*ERCC2L1*) homologous to a human DNA repair gene in *Xiphophorus* fishes: Implications for the evolutionary derivation of human chromosome 19. *Genomics* 10: 1083–1086.

Weber, C. A., Salazar, E. P., Stewart, S. A., and Thompson, L. H. (1990). *ERCC2:* cDNA cloning and molecular characterization of a human nucleotide excision repair gene with high homology to yeast *RAD3. EMBO J.* 9: 1437–1447.

Wedaman, K. P., Knight, A. E., Kendrick-Jones, J., and Scholey, J. M. (1993). Sequences of sea urchin kinesin light chain isoforms. *J. Mol. Biol.* 231: 155–158.

Worley, K. C., Wiese, B. A., and Smith, R. F. (1995). An enhance BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res.* 5: 173–184.

Ziegle, J. S., Su, Y., Corcoran, K. P., Nie, L., Mayrand, P. E., Hoff, L. B., McBride, L. J., Kronick, M. N., and Diehl, S. R. (1992). Application of automated DNA sizing technology for genotyping microsatellite loci. *Genomics* 14: 1026–1031.