Genomics **65,** 274–282 (2000)
doi:10.1006/geno.2000.6174, available online at http://www.idealibrary.com on IDEAL®

Cloning and Characterization of HARP/SMARCAL1: A Prokaryotic HepA-Related SNF2 Helicase Protein from Human and Mouse

Matthew A. Coleman,*[,1] Jonathan A. Eisen,†[,2] and Harvey W. Mohrenweiser*

**Biology and Biotechnology Research Program, L-452 Lawrence Livermore National Laboratory, Livermore, California 94551; and †Department of Biological Sciences, Stanford University, Stanford, California 94305*

Received December 13, 1999; accepted March 2, 2000

The SNF2 gene family consists of a large group of proteins involved in transcriptional regulation, maintenance of chromosome integrity, and various aspects of DNA repair. We cloned a novel SNF2 family human cDNA, with sequence identity to the *Escherichia coli* RNA polymerase-binding protein HepA and named the human hepA-related protein (HHARP/SMARCAL1). In addition, the mouse ortholog (Mharp/Smarcal1) was cloned, and the *Caenorhabditis elegans* ortholog (CE-HARP) was identified in the GenBank database. Phylogenetic analysis indicates that the HARP proteins share a high level of sequence similarity to the seven motif helicase core region (SNF2 domain) with identifiable orthologs in other eukaryotic species, except for yeast. Purified His-tagged HARP/SMARCAL1 protein exhibits single-stranded DNA-dependent ATPase activity, consistent with it being a member of the SNF2 family of proteins. Both the human and the mouse genes consist of 17 exons and 16 introns. The human gene maps to chromosome 2q34–q36, and the mouse gene is localized to the syntenic region of chromosome 1 (between markers *Gls* and *Acrg*). HARP/SMARCAL1 transcripts are ubiquitously expressed in human and mouse tissues, with testis presenting the highest levels of mRNA expression in humans. © 2000 Academic Press

INTRODUCTION

Several complexes of proteins are involved in the remodeling of chromatin to change nucleosome compaction for gene regulation, replication, recombination, and DNA repair (Eisen and Lucchesi, 1998; Tsukiyama and Wu, 1997; Wolffe and Kurumizaka, 1998). A hallmark component of several such multiprotein complexes is a single protein subunit that is a member of the conserved SNF2 family of proteins (Cairns, 1998; Pollard and Peterson, 1998; Tsukiyama and Wu, 1997). Members of the SNF2 family of proteins have been identified in organisms ranging from *Escherichia coli* to *Homo sapiens* (Eisen *et al.*, 1995). They are currently subdivided into 11 evolutionarily and functionally distinct subfamilies, with a wide range of biological functions, but all containing the conserved SNF2 domain (Eisen *et al.*, 1995). The SNF2 domain is defined by the existence of seven motifs (I, Ia, and II–VI) with sequence similarity to those motifs found in DNA and RNA helicases (Bork and Koonin, 1993; Gorbalenya *et al.*, 1989; Koonin, 1993).

The conserved seven helicase motifs comprise the SNF2 domain of approximately 400 amino acid residues (Eisen *et al.*, 1995). Motifs I, Ia, and II of the SNF2 domain make up the nucleotide-binding site and are characterized by a phosphate-binding loop, often referred to as the Walker A and B boxes (Koonin, 1993). This phosphate-binding loop is highly conserved among enzymes that hydrolyze ATP and GTP, as seen with the helicase II superfamily (Gorbalenya *et al.*, 1989). Motif II of the SNF2 domain includes the signature amino acid sequence D-E-x-D (where x is any amino acid), which is referred to as the DEAD box or Walker B box of the superfamily II helicases. The properties of motif III are currently unknown. Motifs IV, V, and VI are involved in DNA and ATP binding, but their exact roles are not yet clear (Korolev *et al.*, 1998). Recently, motif IV has been suggested to have a role in coupling protein conformational changes with ATP hydrolysis and DNA binding (Hall *et al.*, 1998). Despite the presence of the characteristic "helicase" motif, no protein in the SNF2 family has yet demonstrated helicase function. In general, members of the SNF2 family utilize the energy of ATP hydrolysis to stabilize or perturb protein–DNA interactions, while tracking along DNA (Pazin and Kadonaga, 1997). It is this ability to translocate and make and break protein–DNA interactions and thus remodel chromatin through nucleosome restructuring that defines the critical function of the SNF2 protein family.

We describe here the cloning and initial character-

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession Nos. AF082179, AF210833–AF210842, AF08884, and AF209773.

[1] To whom correspondence should be addressed at BBRP, L-448, LLNL, 7000 East Avenue, Livermore, CA 94550. Telephone: (925) 423-7687. Fax: (925) 422-2282. E-mail: coleman16@llnl.gov.

[2] Current address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850.

ization of a novel gene from human and mouse that shares homology with the prokaryotic HEPA1 SNF2 subfamily. The HepA-related protein (HARP/SMARCAL1)[3] subfamily of genes represents the 12th subfamily of the highly conserved SNF2 family of proteins. Although we identified a *Caenorhabditis elegans* ortholog, it is interesting that the HARP subfamily contains no known ortholog in yeast, suggesting a unique higher order eukaryotic function.

MATERIALS AND METHODS

Identification and Cloning of HARP Subfamily Genes

Identification of HHARP/SMARCAL1 cDNA. Human ESTs 121187 and 277802 were selected from the National Center for Biotechnology Information (NCBI) database based on a tBlastn comparison of the amino acid sequence to the SNF2 domain of several human repair proteins such as RAD54 (22% identity) and CSB (24% identity) (Altschul *et al.*, 1990). The ESTs were obtained from the IMAGE consortium collection at Lawrence Livermore National Laboratory and sequenced using previously described techniques (Wilson *et al.*, 1998). To extend the 5′- and 3′-ends of the cDNA coding region, RACE–PCR was performed several times using the Advantage PCR protocol and adaptor-ligated human testis cDNA (Clontech, Palo Alto, CA). The oligos used were as follows—RACE reaction 1: primer 1, 5′-ACTCCAGCTCTCTGAAAGGGC-3′, primer 2, 5′-CACGAGCTTGGGGTCCACTTC-3′; RACE reaction 2: primer 1, 5′-CAATCACCGCCTCCGAATAGC-3′, primer 2, 5′-GCATGCACTTCCCTGTCACAA-3′; and RACE reaction 3; primer 1, 5′-ATGGAGGTGCCCGAGCTAGTC-3′, primer 2, 5′-TGCTTGGCCTGGAATGGGTTG-3′. The RACE–PCR products were cloned into the pGemTA vector (Promega, Madison, WI) and sequenced as above. The human cDNA sequence was submitted to GenBank and assigned Accession No. AF082179.

Identification of Mharp/SMARCAL1 cDNA. To identify an orthologous mouse cDNA, we screened a mouse pachytene spermatocyte cDNA library (Caldwell *et al.*, 1996) constructed in the Lambda ZAP vector (Stratagene, La Jolla, CA) using a PCR product from human EST 121187 as the probe. Following successive rounds of hybridization of nylon filter plaque lifts followed by plaque purification, probe-positive cDNA isolates were converted to phagemids according to the manufacturer's *in vivo* excision protocol (Stratagene). Phagemid DNA was isolated and sequenced. RACE was once again performed to obtain the 5′ and 3′ cDNA sequences using adaptor-ligated mouse testis cDNA. The oligos used for the mouse cDNA were as follows—RACE reaction 1: primer 1, 5′-GCTGGGCAGAGACTTAAACAC-3′, primer 2, 5′-GGAAGCGATCCCCAGTCTTTA-3′; RACE reaction 2: primer 1, 5′-CAATCACCGCCTCCGAATAGC-3′, primer 2, 5′-GCATGCACTTCCCTGTCACAA-3′. The mouse cDNA sequence was submitted to GenBank and assigned Accession No. AF08884.

Identification of the HARP/SMARCAL1 subfamily of genomic clones. Arrayed human and mouse BAC libraries (Research Genetics, Huntsville, AL) were screened by hybridizing with the entire sequence of EST 121187 as described under Northern and Southern Blot Analysis. A total of nine positive human BAC clones (37m5, 41n8, 114f8, 139c11, 139k14, 210f11, 210g11, 248l22, and 367h24), and five positive mouse BAC clones (122I16, 73p5, 196g3, 260f4, and 122j17) were found. The mouse BAC 122j17 with a 123-kb insert and the human BAC 367h24 with a 162-kb insert were sequenced as previously described (Brookman *et al.*, 1996). The human genomic sequence from BAC 367h24 was assigned GenBank Accession Nos. AF210833–AF210842, and the mouse genomic sequence from BAC 122j17 was assigned GenBank Accession No. AF209773.

Chromosomal Localization of Genes

Mapping of the HHARP/SMARCAL1 gene. The entire sequence of EST 121187 (*HHARP/SMARCAL1*) was compared with the Sequenced-Tagged Site (STS) database at NCBI to identify positively mapped EST sequences (Olson *et al.*, 1989). Data from mapping of the human and mouse genes were compared using NCBI's Human and Mouse Homology Mapping Web site (http:3/27/00/www.ncbi.nlm.nih.gov/Homology/) (DeBry and Seldin, 1996).

Interspecific backcross mapping of the mouse Harp/Smarcal1 gene. A 2.0-kb *Mharp/Smarcal1* cDNA probe was mapped in mouse. This was accomplished by tracing the segregation of restriction fragment length variants in progeny of a *Mus musculus* × *Mus spretus* interspecific backcross (Doyle *et al.*, 1996; Stubbs *et al.*, 1996).

Northern and Southern Blot Analyses

Nylon blots (MTN1, MTN2, RMTN, and MMTN1; Clontech) containing mRNA from various human, rat, or mouse tissues were prehybridized for 1 h at 65°C in Express Hybridization Solution (Clontech) containing salmon sperm DNA. A 1.2-kb PCR product from the human EST 121187 was labeled with [α-^{32}P]dCTP (Amersham, Arlington Heights, IL) using the Megaprime DNA Labeling System and hybridized at 65°C for 1–2 h. Blots were washed once at room temperature for 30 min and twice at 65°C for 30 min in 50 mM Na_2PO_4 (pH 7.4), 0.5% SDS, 1.0 mM EDTA. Hybridization signals were visualized with a Molecular Dynamics Storm 860 PhosphorImager or by autoradiography. Band intensities are compared against the *β-actin* cDNA control supplied by Clontech. Southern blot analysis of a Clontech interspecies zoo-blot using ^{32}P-labeled human EST 121187 PCR product was performed under reduced stringency conditions.

Computer Software and Bioinformatics

The alignment of protein sequences was accomplished using ClustalW and Clustalx multiple sequence alignment programs (Baylor College of Medicine). Some of the computer-generated alignments were optimized by minor manual adjustment. The amino acid sequence alignment figure was produced with Boxshade (http://ulrec3.unil.ch/software/BOX_form.html). Phylogenetic trees were generated from sequence alignments as previously reported (Eisen, 1998; Eisen *et al.*, 1995). Prosite, Profile Scan (http://www.isrec.isb-sib.ch/profile.html), and PSORT II (http://psort.nibb.ac.jp:8800/) were used to identify amino acid sequence motifs with significant homology (Bairoch *et al.*, 1996; Bucher *et al.*, 1996).

Protein Expression, Purification, and ATPase Assays

The *Mharp/Smarcal1* coding region was reverse transcribed and PCR amplified from mouse total RNA using primers 5′-CCCCCATATGTCCTTGCCACTTACAGAGGAGCA-3′ and 5′-CCCGGATCCAAAGGGAGAGGAAAAGCTGTC-3′. The RT-PCR product, which lacked coding sequence for the first 107 amino acids, was digested with *Nde*I and *Bam*HI and then directionally cloned into pET28a (Novagen, Madison, WI) plasmid and sequenced. The plasmid pEM-HARP contained the *Mharp/Smarcal1* cDNA fragment that had been cloned into plasmid pET28a with a 6xHis tag at the N-terminus. Protein expression was performed in BL834 (DE3) *E. coli* cells (Novagen), in the presence of 30 μg/ml of kanamycin. Cultures were grown to an OD of 0.6, and isopropyl-beta-D-thiogalactopyranoside was added to a final concentration of 2.5 mM. Inductions proceeded for 4 h at 37°C, before cells were collected by centrifugation at 10,000*g* for 20 min. Cell pellets were washed once with buffer A (50 mM Na_2PO_4 (pH 8.0), 200 mM NaCl, and 5% glycerol) and processed immediately or frozen (−80°C) until needed. Processing entailed resuspending pellets in cold buffer A and sonicating with three bursts of 15 s using a Misonix XL sonicator. Insoluble material was removed by centrifugation at 20,000*g* for 30 min; lysates containing

[3] *SMARCAL1* (SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily a-like 1) is the HGM locus designation for the *HHARP* gene.

Mharp/Smarcal1 protein, as analyzed by Western blotting (see below), were retained. Imidazole was added to the lysate to a final concentration of 20 mM. The Ni-NTA resin (Qiagen, Santa Clarita, CA) was equilibrated in buffer A before the soluble lysate was added. This mixture was incubated on ice with agitation for 1 h. The Ni-NTA lysate mixture was then poured onto a 5-ml column and washed with 5 bed-volumes of buffer A containing 20 mM imidazole. Bound protein was eluted with a step gradient of 20–500 mM imidazole in buffer A. Fractions of 1 ml were collected and assayed by Western blotting using a histidine-specific primary antibody coupled with a HRP-conjugated secondary antibody (Santa Cruz Biotechnology, Santa Cruz, CA) and ECL Enhanced Chemiluminescence (Amersham). The major Mharp/Smarcal1 protein containing fractions were dialyzed in buffer A to remove imidazole, pooled, and passed over a S10 cationic-exchange column using a Biologic Workstation (Bio-Rad, Hercules, CA). The protein was eluted with a gradient of 25–500 mM NaCl in buffer A. Two milliliter fractions were collected and analyzed by Western blot.

Fractions positive for the Mharp/Smarcal1 protein were assayed for ATPase activity using previously described methods (Matson and Richardson, 1983). Briefly, 3 μl of a Mharp/Smarcal1 positive protein fraction, as assayed by Western blot, was added to 5 μl of 4× ATPase buffer (160 mM Tris–HCl (pH 7.5), 16 mM $MgCl_2$, and 4 mM DTT), 4 μl of 4× ATP (2.5 mM ATP + [α-^{32}P]ATP), 2 μl of 20 mM M13 ssDNA, and 6 μl of H_2O at 4°C. The assay mix was incubated at 37°C for 1 h. The reaction was terminated by addition of 5 μl of 0.25 mM EDTA. A 5-μl aliquot was spotted on a thin-layer chromatography plate (Polyethylene Imine Cellulose, PEL-F; J. T. Baker, Phillipsburg, NJ) and placed in a 1 M formic acid and 0.8 M lithium chloride buffer at room temperature. Chromatographs were visualized and quantitated with a Molecular Dynamics Storm 860 PhosphorImager.

RESULTS

Identification of Novel Members of the SNF2 Family

We isolated cDNAs from both human and mouse encoding proteins with a high level of sequence similarity to the seven motifs of the helicase core region of genes in the SNF2 family. This was accomplished using search programs from NCBI to identify ESTs of interest and then applying computational and molecular biology tools to characterize the cDNA clones. The full-length *HHARP/SMARCAL1* cDNA, encoding a 954-amino-acid protein with a predicted molecular mass of 105,912 Da and a theoretical p*I* of 9.17, was isolated using RT-PCR. The *Mharp/Smarcal1* cDNA, encoding a 910-amino-acid protein with a predicted molecular mass of 101,043 Da and a theoretical p*I* of 9.24, was similarly isolated. The identity between the two full-length proteins is 72% (Fig. 1A). The human and mouse proteins show 80% identity across the highly conserved SNF2 domain. A search of the nonredundant protein database at NCBI revealed that *C. elegans* has an open reading frame, C16A3.Cel (Accession No. U41534) (Wilson *et al.*, 1994), encoding an 880-amino-acid protein (CEHARP) that shares homology with the human and mouse proteins. The CEHARP protein is 37% identical to HHARP/SMARCAL1 and 41% identical to Mharp/Smarcal1 over the entire length of the respective protein sequence (Fig. 1A). All other SNF2 family members available in the GenBank database had identities that are less than 28% when the entire amino acid sequences were compared to any of the HARP/SMARCAL1 protein sequences. Human and mouse HARP/SMARCAL1 proteins are 51% identical at the N-terminus while the identity of mouse vs CEHARP is 31% and human vs CEHARP is 28% for this region. A Blast search failed to find any additional proteins with homology to the N-terminus of these proteins. The N-terminal HARP/SMARCAL1 domain is more evolutionarily divergent than the conserved SNF2 domain (C-terminal domain).

Within the SNF2 domain of the three HARP proteins, motif I includes the expected SNF2 pattern of GxGK[S/T]x (where x is any amino acid). This motif corresponds to the Walker A site for helicases (Gorbalenya and Koonin, 1993). The mouse and human HARP proteins have a threonine at the [S/T] position, while a serine is inserted for CEHARP. In contrast, detailed analysis of the Walker B site in motif II reveals conservation of a serine residue for all three of the HARP/SMARCAL1 proteins. A DESH amino acid sequence in motif II is conserved in all three organisms (Fig. 1B). The overall spatial arrangement is conserved for helicase motifs I–VI. All three HARP/SMARCAL1 proteins exhibit a bipartite nuclear localization signal (bipartite-NLS) located within motif IV of the SNF2 domain, with the consensus amino acid sequence being RRLKSDVLSQLPAKQRKI for human and mouse. The human and mouse proteins contain a second bipartite-NLS sequence that is located at the N-terminus and is encoded by the amino acid motif KKIEENRQKALARRAEKL, which was not found in CEHARP. A prediction for nuclear localization of the HARP/SMARCAL1 proteins, using the PSORT II algorithm, indicates that both HHARP/SMARCAL1 (89% reliability) and Mharp/Smarcal1 (71% reliability) should be localized to the nucleus, whereas CEHARP was predicted to be cytoplasmic (56% reliability). The CEHARP protein has most likely evolved a mechanism for nuclear import that is different from human or mouse. The HARP/SMARCAL1 proteins do not have defined amino acid motifs such as a bromodomain, chromodomain, or zinc fingers that play a role in transcription or repair and are characteristically found in other related SNF2 proteins (Eisen *et al.*, 1995).

Molecular phylogenetic analysis of the HARP/SMARCAL1 SNF2 domains segregated the three HARP/SMARCAL1 proteins into a single new subfamily (Fig. 2) within the SNF2 family of proteins (Eisen *et al.*, 1995). This approach also provided additional verification that *C. elegans* cDNA (C16A3.Cel) is a related member of the HARP/SMARCAL1 subfamily. No yeast or prokaryotic orthologs were found to be part of this novel subfamily of SNF2 proteins. In contrast, every other eukaryotic SNF2 subfamily has a yeast protein member. The nearest segregated evolutionary link to the HARP/SMARCAL1 subfamily, based on phylogenetic analysis, is the prokaryotic HEPA1 subfamily (Fig. 2). The nearest eukaryotic phylogenetic links to the HARP/SMARCAL1 subfamily are the RAD16 and RAD54 subfamilies.

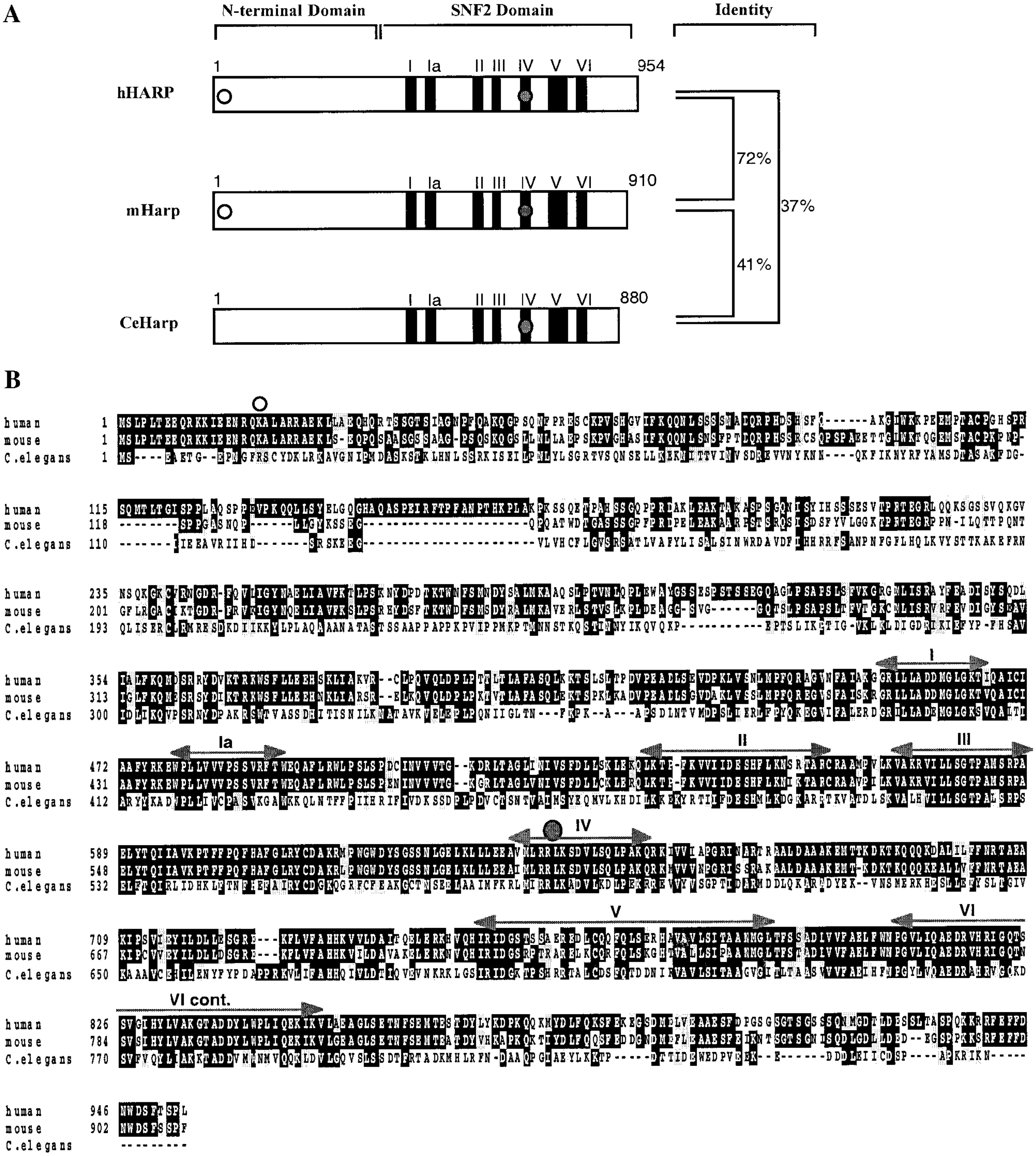

FIG. 1. Schematic diagram of motifs of the protein domains identified in the HARP/SMARCAL1 subfamily of proteins. (**A**) The seven motifs of the SNF2 domains are indicated above the schematic. The nuclear localization signals are indicated by an open circle (NLS shared by HHARP/SMARCAL1 and Mharp/Smarcal1) or a closed circle (NLS shared by all three proteins) within the respective diagrams. The identity between proteins is shown. (**B**) Comparison of the predicted amino acid sequence of the SNF2 HARP/SMARCAL1 subfamily. The seven putative helicase motifs are indicated by arrows above the sequence. Amino acids with white type and black backgrounds are identical, amino acids shaded in gray are conserved, and amino acids with black type and white backgrounds are nonconserved.

A comparison of cDNA to genomic sequence reveals that both human and mouse genes contain 17 exons that range in size from 37 to 869 bp (Table 1). Intron sizes are shown only for the mouse gene, and they range in size from 549 to 9857 bp (Table 1). Exon 1 consists only of 5′-untranslated sequence in both organisms. Exon 2 contains the translational start site, while exon 17 encodes the translational stop codon and the 3′-untranslated region. The SNF2 domain is encoded by exons 6 to 17. An interspecies Southern blot, using a human 1.2-kb probe (EST 121187) that contained the SNF2 domain in the C-terminal half of

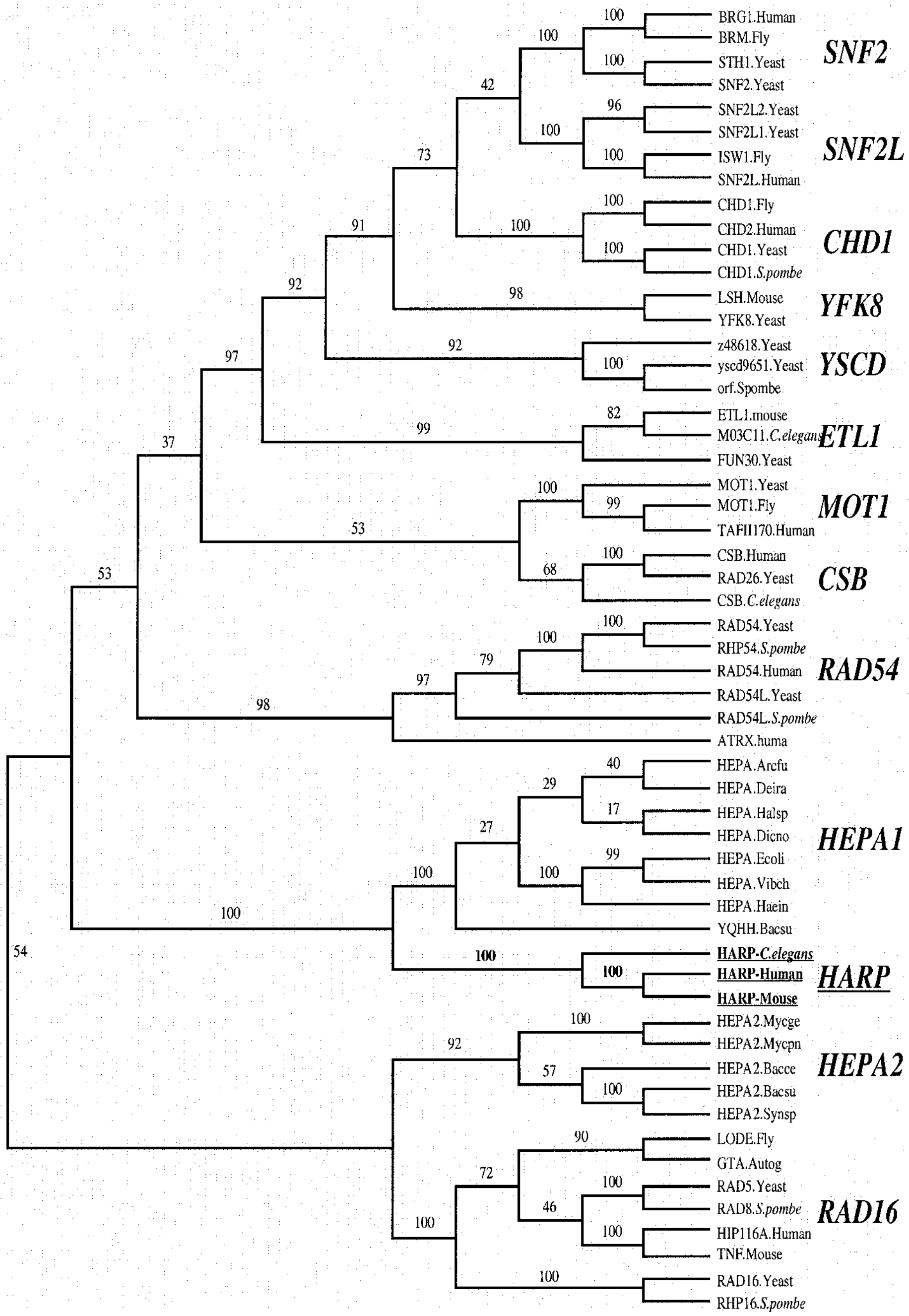

FIG. 2. Organization of the SNF2 gene subfamilies based on phylogenetic bootstrapping. A molecular phylogenetic approach was used to determine the relationship of the HARP/SMARCAL1 proteins (shown underlined and in boldface type) to other members of the SNF2 family of proteins. The boldface type indicates subfamilies within the SNF2 family of proteins. Members of each subfamily are indicated in a smaller font. Bootstrap values indicate the percentage of times a subfamily was grouped together.

HHARP/SMARCAL1, showed cross-hybridization with eukaryotic genomes, including rat, cat, and cow, and was consistent with a single HARP gene in each species (data not shown).

Human and Mouse HARPSMARCAL1 Genes Localize to Syntenic Chromosomes

The marker STSG8125, combined by NCBI, was found to contain sequence overlapping ESTs T08650 (HIBBH71), T91835 (116570), T96959 (121187), and N47591 (277802), corresponding to the *HHARP/SMARCAL1* cDNA. Marker STSG8125 is localized to chromosome 2 in a region between markers D2S164 and D2S163 (222–225.6 cM) in the radiation hybrid panel database and corresponds to the 2q34–q36 region of the chromosome. The mouse *Mharp/Smarcal1* gene was localized, using interspecific backcross mapping (Stubbs *et al.,* 1996), to mouse chromosome 1 between markers *Gls* (phosphate-activated glutami-

TABLE 1

Genomic Structure of the Harp Gene

Exon No.	Exon size[a] (bp)	Intron size[b]	AA at splice site[c]
1	37/167	2197	—
2	869/768	646	Asp
3	51	3962	Met
4	233/212	549	Asp
5	51	3913	Ile
6	186	1403	Asn
7	155/153	1236	Gln
8	162/163	2439	Asp
9	67	9857	Lys/Val
10	142	1373	Met
11	220/216	3484	Thr/Lys
12	70	3454	Ile/Glu
13	102	6327	His
14	186	1453	Gly/Val
15	102	1172	Pro
16	98	3346	Lys/Asp
17	249/316	—	—

[a] The first number is for human sequence and the second is for mouse sequence.

[b] Intron size is only for mouse genomic sequence.

[c] Amino acids located at the putative splice site that formed at the acceptor and donor region of each exon.

nase) and *Acrg* (γ-subunit of the acetylcholine receptor; data not shown). Markers *Gls* and *Acrg* are separated by 23.3 cM with the mouse *Harp/Smarcal1* gene located approximately 14.6 cM from *Gls* and approximately 8.7 cM from *Acrg.* Comparison between mouse and human chromosomal locations revealed that the mouse *Gls* and *Acrg* genes are arranged in a similar order as are the human markers *GLS* and *CHRNG* (*CHRNG* is the human gene corresponding to the mouse *Acrg* gene) on human chromosome 2q34–q36 (DeBry and Seldin, 1996). A comparison of the syntenic regions identified 26 genes in common for human and mouse. One discordant localization was noted, as the mouse syntenic region contained the *Xpg* gene that is located at 13q32–q33 in humans.

Expression Pattern of HARP/SMARCAL1 mRNA Is Ubiquitous

A major ~3.4-kb transcript was detected in all tissues examined using multitissue Northern (MTN) blots from human, mouse, and rat (Fig. 3A). The level of expression is similar in all tissue types, with the exception of testis. For both human and rat, HARP/SMARCAL1 mRNA levels are elevated in testis compared to the eight other tissues examined (rat data not shown). In mouse, testis mRNA levels appear comparable to the transcript levels of kidney, heart, liver, and brain tissue, which are slightly higher than the levels detected in spleen, lung, and skeletal muscle (Fig. 3B). Of special note is the additional 7.0-kb transcript that is abundant in mouse brain (Fig. 3B) but not detected in other mouse tissues. The 7.0-kb transcript is unique to mouse brain as only the 3.4-kb transcript was seen in the human (Fig. 3A) and rat brain tissues (rat data not shown).

The HARP/SMARCAL1 Proteins Exhibit DNA-Dependent ATPase Activity

Other members of the SNF2 family of proteins exhibit DNA-dependent ATPase activity. To test for this enzymatic function in the HARP/SMARCAL1 subfamily, we generated a recombinant form of the Mharp/Smarcal1 protein. The expression plasmid construct (pEMHARP) produced a truncated protein, where the first 107 amino acids from the N-terminus were absent and replaced with a 6xHis tag. The mHarp-6x His protein was purified by nickel affinity and followed by S10 cation-exchange chromatography (Fig. 4A); the mHarp eluted in fractions 8–15 (Figs. 4A and 4B). All fractions containing Mharp/Smacal1 protein were tested for ATPase activity in the presence and in the absence of ssDNA. The highest levels of ATPase activity in the presence of ssDNA were seen in fractions 11 and 12 (Fig. 4C). No ATPase activity was detected in the absence of ssDNA. Mock fractions consisting of potential contaminating proteins from extracts of BL834(DE3) bacterial cells harboring the pET28a

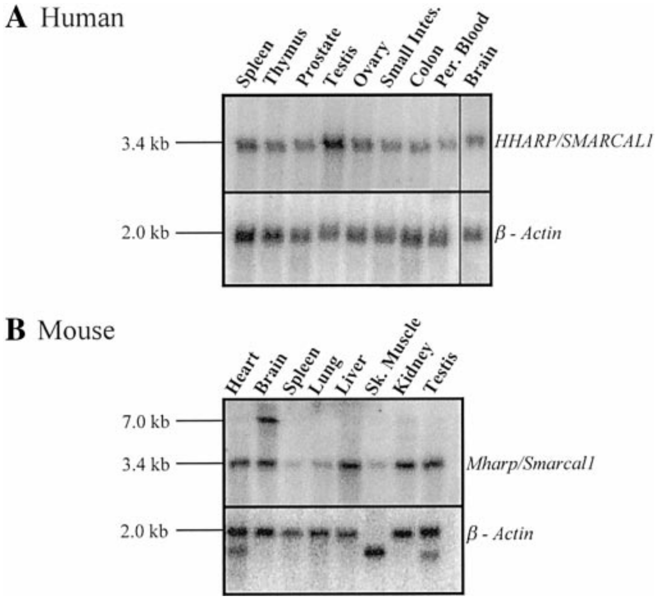

FIG. 3. The tissue expression pattern of *HHARP/SMARCAL1* and *Mharp/Smarcal1* transcripts. (**A**) Northern blot analysis of human mRNA expression in several adult tissue types (see Materials and Methods). The brain sample is from Clontech's human MTN II blot. A *β-actin* cDNA probe of the same blot was used as a control. Size markers are indicated on the left of the autoradiogram. (**B**) Mouse tissue distribution of *Mharp/Smarcal1* mRNA using a mouse-specific HARP probe. The same blot was later probed with *β-actin* cDNA as a control.

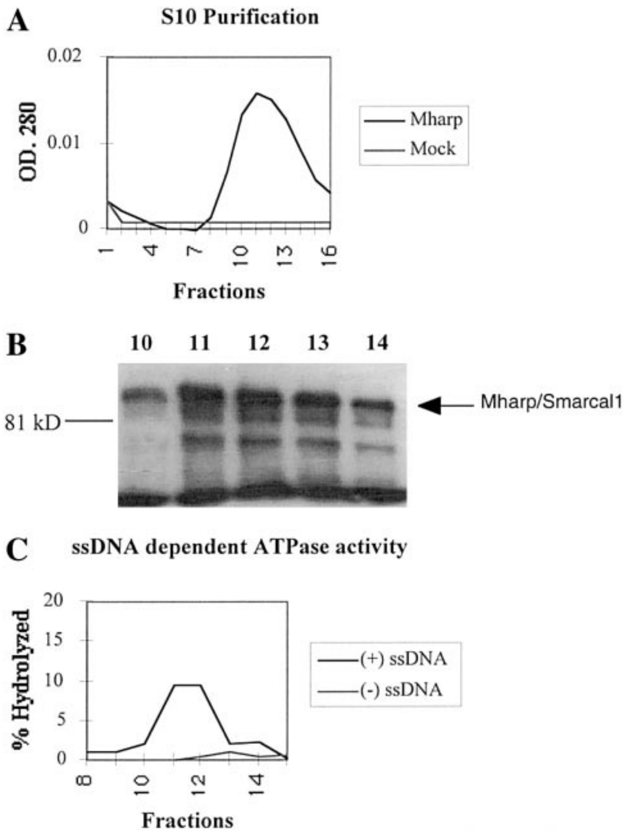

FIG. 4. Purification and activity of a Mharp/Smarcal1 His-tagged protein. (**A**) The Mharp/Smarcal1 and Mock protein fraction elution profiles from a cationic S10 column are shown (see Materials and Methods for details). (**B**) A Western blot of the peak protein fractions isolated from the S10 column. The position of the Mharp/Smarcal1 protein is indicated with an arrow (Western blotting of the Mock fraction detected no bands (data not shown)). Numbers at the top correspond to the eluted S10 column fractions. (**C**) Fractions were assayed for ATPase activity in the presence ((+) ssDNA) or in the absence of ssDNA ((−) ssDNA).

vector failed to show ATPase activity, with or without ssDNA.

DISCUSSION

Using EST database searches to identify novel genes encoding proteins with homology to the SNF2 domain, we cloned and characterized novel conserved human and mouse genes, which we have termed the HepA-related protein (HARP/SMARCAL1) because of sequence similarity to the *E. coli* HepA protein from the SNF2 HEPA1 subfamily. A *C. elegans* HARP/SMARCAL1 ortholog was also identified by searching the GenBank database. Together, these three genes constitute a distinct subfamily within the SNF2 family. All three cDNAs encode proteins of similar size and share conserved SNF2 regions in the C-terminal end of the proteins. Currently, the HEPA1 subfamily is thought to represent the oldest evolutionary members of the SNF2 family (Eisen *et al.*, 1995). The phylogenetic relatedness of the eukaryotic HARP/SMARCAL1 subfamily to the prokaryotic HEPA1 subfamily (Fig. 2) suggests that the HARP/SMARCAL1 subfamily is one of the older eukaryotic SNF2 subfamilies and may have arisen by gene duplication (Eisen *et al.*, 1995). Alternatively, the HARP/SMARCAL1 gene could have originated by a lateral transfer event from bacteria to eukaryotes (Eisen *et al.*, 1995; Muzzin *et al.*, 1998).

The genes for human and mouse are located on syntenic chromosomes providing further evidence that these genes are functional orthologs. Of interest is the occurrence of chromosomal aberrations such as translocations and deletions within the region 2q34–q36.1 in several cancers in humans (Krajinovic *et al.*, 1998a,b). Loss of function for SNF2 family members has been found to be directly associated with several health-related syndromes, such as Cockayne syndrome and α-thalassemia (Gibbons *et al.*, 1995; Troelstra *et al.*, 1992). The potential involvement of HARP/SMARCAL1 in oncogenesis through a mechanism of transcriptional initiation of tumor suppressor genes or oncogenes remains to be determined.

The HARP/SMARCAL1 transcripts are ubiquitously expressed in all tissue types in the species studied, suggesting that the encoded proteins are involved in normal cellular functions or housekeeping activities such as transcriptional regulation. Many of the other SNF2 genes, such as ISWI and SNF2, which are involved in regulation of transcription, show similar ubiquitous patterns of tissue expression. The levels and tissue distribution of expression may also be consistent with a role for HARP/SMARCAL1 proteins in DNA repair as their transcripts are primarily elevated in testis tissues in human and rats. Other SNF2-related repair proteins such as the human and mouse HELLS and RAD54 also show elevated transcript levels in the testis (Geiman *et al.*, 1998; Kanaar *et al.*, 1996).

One measurable functional feature of the HARP/SMARCAL1 subfamily of proteins is the ATPase domain. Within the ATPase domain, the DESH signature appears as a distinct feature of this subfamily, as it has not been seen in other helicases. It is speculated that the serine residue may have a role in Mg^{2+} coordination and, as such, could change HARP/SMARCAL1 protein affinity for the Mg-NTP molecule (Gorbalenya and Koonin, 1993). The ATPase activity of the HARP/SMARCAL1 proteins is stimulated by ssDNA. The ssDNA-dependent ATPase activity was associated with only the fractions containing Mharp/Smarcal1-6xHis protein, demonstrating that the HARP/SMARCAL1 proteins encode a functional NTP-binding site capable of interacting with nucleic acids.

Further insight into HARP/SMARCAL1 protein function may be gained from the characteristics of the closely related prokaryotic HEPA1 subfamily. The HEPA1 subfamily is best represented by the *E. coli* HepA 110-kDa protein, which plays a role in both repair and transcription (Muzzin *et al.*, 1998; Sukhodolets and Jin, 1998). When the *hepA* gene is disrupted in *E. coli*, the cells become UV sensitive (Muzzin *et al.*, 1998; Sukhodolets and Jin, 1998). The HepA protein also interacts with the RNA polymerase core enzyme (Sukhodolets and Jin, 1998). One can imagine that HARP/SMARCAL1 proteins may play a similar role in higher eukaryotes involving an interaction with RNA polymerase complexes to influence transcription, for example, genes encoding mammalian transcription factors interact with RNA polymerase II holoenzymes to activate or regulate transcriptional initiation (Mitchell and Tjian, 1989).

Although a clearly defined biological function for the HARP/SMARCAL1 subfamily is not yet known, we have identified them as proteins that represent a unique eukaryotic subfamily of the SNF2 family that do not contain a representative protein from yeast. They are ubiquitously expressed in human and mouse and exhibit an ATPase function in the presence of ssDNA that is consistent with SNF2 proteins. Biochemical and genetic studies to identify the protein complexes that include HARP will be crucial to understanding the possible role of HARP/SMARCAL1 in transcription, replication, and various aspects of DNA repair.

ACKNOWLEDGMENTS

We thank the IMAGE Consortium, Aaron Adamson, and Drs. Irene Jones, Mark Shannon, James George, Ian McConnell, and David Wilson III for technical help as well as for preparation of the manuscript. We also thank the Joint Genome Institute and Jane Lamerdin for assistance and helpful advice in the sequencing of BAC clones. This research was performed under the auspices of USDOE by LLNL Contract No. W-7405-ENG-48.

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Bairoch, A., Bucher, P., and Hofmann, K. (1996). The PROSITE database, its status in 1995. *Nucleic Acids Res.* **24:** 189–196.

Bork, P., and Koonin, E. V. (1993). An expanding family of helicases within the 'DEAD/H' superfamily. *Nucleic Acids Res.* **21:** 751–752.

Brookman, K., Lamerdine, J., Thelen, M., Hwang, M., Reardon, J., Sancar, A., Zhou, Z., Walter, C., Parris, C., and Thompson, L. (1996). ERCC4 (XPF) encodes a human nucleotide excision repair protein with eukaryotic recombination homologs. *Mol. Cell. Biol.* **16:** 6553–6562.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). A flexible search technique based on generalized profiles. *Comput. Chem.* **20:** 3–24.

Cairns, B. (1998). Chromatin remodeling machines: Similar motors, ulterior motives. *Trends Biochem. Sci.* **1:** 20–25.

Caldwell, K., Wiltshire, T., and Handel, M. (1996). A genetic strategy for differential screening of meiotic germ-cell cDNA libraries. *Mol. Reprod. Dev.* **43:** 403–413.

DeBry, R., and Seldin, M. (1996). Human/mouse homology relationships. *Genomics* **33:** 337–351.

Doyle, J., Hellevuo, K., and Stubbs, L. (1996). The gene encoding adenylyl cyclase VII is located in central mouse chromosome 8. *Mamm. Genome* **7:** 320–321.

Eisen, A., and Lucchesi, J. (1998). Unraveling the role of helicases in transcription. *BioEssays* **20:** 634–641.

Eisen, J. (1998). A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res.* **26:** 4291–4300.

Eisen, J. A., Sweder, K. S., and Hanawalt, P. C. (1995). Evolution of the SNF2 family of proteins: Subfamilies with distinct sequences and functions. *Nucleic Acids Res.* **14:** 2715–2723.

Geiman, T., Durum, S., and Muegge, K. (1998). Characterization of gene expression, genomic structure, and chromosomal localization of Hells (Lsh). *Genomics* **54:** 477–483.

Gibbons, R., Picketts, D., Villard, L., and Higgs, D. (1995). Mutations in a putative global transcriptional regulator cause X-linked mental retardation with alpha-thalassemia (ATR-X syndrome). *Cell* **80:** 837–845.

Gorbalenya, A., and Koonin, E. (1993). Helicases: Amino acid sequence comparisons and structure–function relationships. *Curr. Opin. Struct. Biol.* **3:** 419–429.

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.* **12:** 4713–4730.

Hall, M., Ozsoy, A., and Matson, S. (1998). Site-directed mutations in motif VI of *Escherichia coli* DNA helicase II result in multiple biochemical defects: Evidence for the involvement of motif VI in the coupling of ATPase and DNA binding activities via conformational changes. *J. Mol. Biol.* **277:** 257–271.

Kanaar, R., Troelstra, C., Swagemakers, S., Essers, J., Smit, B., Franssen, J., Pastink, A., Bezzubova, O., Buerstedde, J., Clever, B., Heyer, W., and Hoeijmakers, J. (1996). Human and mouse homologs of the *Saccharomyces cerevisiae* RAD54 DNA repair gene: Evidence for functional conservation. *Curr. Biol.* **6:** 828–838.

Koonin, E. V. (1993). A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic Acids Res.* **21:** 2541–2547.

Korolev, S., Yao, N., Lohman, T. M., Weber, P. C., and Waksman, G. (1998). Comparisons between the structures of HCV and Rep helicases reveal structural similarities between SF1 and SF2 superfamilies of helicases. *Protein Sci.* **3:** 605–610.

Krajinovic, M., Richer, C., Gorska-Flipot, I., Gaboury, L., Novakovic, I., Labuda, D., and Sinnett, D. (1998a). Genomic loci susceptible to replication errors in cancer cells. *Br. J. Cancer* **78:** 981–985.

Krajinovic, M., Richer, C., Lukovic, L., Labuda, D., and Sinnett, D. (1998b). Chromosomal assignment of loci susceptible to replication errors by radiation hybrid mapping. *Mutat. Res.* **382:** 81–83.

Matson, S., and Richardson, C. (1983). DNA-dependent nucleoside 5′-triphosphatase activity of the gene 4 protein of bacteriophage T7. *J. Biol. Chem.* **258:** 14009–14016.

Mitchell, P. J., and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245:** 371–378.

Muzzin, O., Campbell, E., Xia, L., Severinova, E., Darst, S., and Severinov, K. (1998). Disruption of *Escherichia coli* hepA, an RNA polymerase-associated protein, causes UV sensitivity. *J. Biol. Chem.* **24:** 15157–15161.

Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989). A common language for physical mapping of the human genome. *Science* **245:** 1434–1435.

Pazin, M. J., and Kadonaga, J. T. (1997). SWI2/SNF2 and related proteins: ATP-driven motors that disrupt protein–DNA interactions? *Cell* **6:** 737–740.

Pollard, K., and Peterson, C. (1998). Chromatin remodeling: A marriage between two families? *BioEssays* **20:** 771–780.

Stubbs, L., Carver, E., Shannon, M., Kim, J., Geisler, J., Generoso, E., Stanford, B., Dunn, W., Mohrenweiser, H., Zimmermann, W., Watt, S., and Ashworth, L. (1996). Detailed comparative map of human chromosome 19q and related regions of the mouse genome. *Genomics* **35:** 499–508.

Sukhodolets, M., and Jin, D. (1998). RapA, a novel RNA polymerase-associated protein, is a bacterial homolog of SWI2/SNF2. *J. Biol. Chem.* **273:** 7018–7023.

Troelstra, C., Gool, A. V., Wit, J. D., Vermeulen, W., Bootsma, D., and Hoeijmakers, J. (1992). ERCC6, a member of a subfamily of putative helicases, is involved in Cockayne's syndrome and preferential repair of active genes. *Cell* **71:** 939–953.

Tsukiyama, T., and Wu, C. (1997). Chromatin remodeling and transcription. *Curr. Opin. Genet. Dev.* **2:** 182–191.

Wilson, D., Carney, J., Coleman, M., Adamson, A., Christensen, M., and Lamerdin, J. (1998). Hex1: A new human Rad2 nuclease family member with homology to yeast exonuclease 1. *Nucleic Acids Res.* **26:** 3762–3768.

Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., and Cooper, J. (1994). 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368:** 32–38.

Wolffe, A., and Kurumizaka, H. (1998). The nucleosome: A powerful regulator of transcription. *Prog. Nucleic Acid Res. Mol. Biol.* **61:** 379–422.