The EMBO Journal vol.13 no.8 pp.1831 – 1843, 1994

Purification and cloning of a nucleotide excision repair complex involving the xeroderma pigmentosum group C protein and a human homologue of yeast RAD23

Chikahide Masutani[1], Kaoru Sugasawa[1,2], Junn Yanagisawa[3,4], Tadao Sonoyama[3], Michio Ui[3,5], Takemi Enomoto[3], Koji Takio[6], Kiyoji Tanaka[7], Peter J.van der Spek[8], Dirk Bootsma[8], Jan H.J.Hoeijmakers[8] and Fumio Hanaoka[9]

[1]Cellular Physiology Laboratory, [2]Biodesign Research Group and [6]Department of Research Fundamentals Technology, The Institute of Physical and Chemical Research (RIKEN), Wako-shi, Saitama 351-01, [3]Department of Physiological Chemistry, Faculty of Pharmaceutical Sciences, University of Tokyo, Bunkyo-ku, Tokyo 113, [7]Institute for Molecular and Cellular Biology, Osaka University, 1-3 Yamada-oka, Suita, Osaka 565, Japan and [8]Department of Cell Biology and Genetics, Medical Genetic Centre, Erasmus University, PO Box 1738, 3000 DR Rotterdam, The Netherlands

[4]Present address: La Jolla Cancer Research Foundation, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

[5]Present address: Ui Special Laboratory, The Institute of Physical and Chemical Research (RIKEN), 2-1 Hirosawa, Wako-shi, Saitama 351-01, Japan

[9]Corresponding author

Communicated by D.Bootsma

Complementation group C of xeroderma pigmentosum (XP) represents one of the most common forms of this cancer-prone DNA repair syndrome. The primary defect is located in the subpathway of the nucleotide excision repair system, dealing with the removal of lesions from the non-transcribing sequences ('genome-overall' repair). Here we report the purification to homogeneity and subsequent cDNA cloning of a repair complex by *in vitro* complementation of the XP-C defect in a cell-free repair system containing UV-damaged SV40 minichromosomes. The complex has a high affinity for ssDNA and consists of two tightly associated proteins of 125 and 58 kDa. The 125 kDa subunit is an N-terminally extended version of previously reported XPCC gene product which is thought to represent the human homologue of the *Saccharomyces cerevisiae* repair gene *RAD4*. The 58 kDa species turned out to be a human homologue of yeast RAD23. Unexpectedly, a second human counterpart of RAD23 was identified. All RAD23 derivatives share a ubiquitin-like N-terminus. The nature of the XP-C defect implies that the complex exerts a unique function in the genome-overall repair pathway which is important for prevention of skin cancer.

Key words: nucleotide excision repair/RAD mutants/ repair complex/ubiquitin/XP

Introduction

DNA repair plays a key role in the prevention of carcinogenesis and mutagenesis. Nucleotide excision repair (NER) is the principal pathway for eliminating a broad spectrum of structurally unrelated lesions such as ultraviolet (UV)-induced cyclobutane pyrimidine dimers and [6-4] photoproducts, as well as bulky chemical adducts and certain cross-links (for review see Friedberg, 1985). At least five steps can be discerned in the reaction mechanism of NER: damage recognition, incision of the damaged strand on both sides of the lesion (Huang *et al.*, 1992), excision of the lesion-containing oligonucleotide, synthesis of new DNA using the undamaged strand as a template, and ligation. Although the molecular mechanism underlying NER is now well understood in the bacterium *Escherichia coli* (Van Houten, 1990; Hoeijmakers, 1993a; Sancar and Hearst, 1993), the mechanism of NER in mammals has not yet been clarified. The high level of sophistication of the NER system is illustrated by the existence of distinct subpathways. One of these deals with the preferential elimination of lesions that thwart ongoing transcription (transcription-coupled repair), a second subpathway effects the slower repair of the rest of the genome ('genome-overall' repair) (Hanawalt and Mellon, 1993).

The association of a DNA repair defect with a human cancer-prone syndrome, xeroderma pigmentosum (XP) was first reported by Cleaver (1968). XP is a rare, autosomal recessive disease associated with a high incidence of sunlight-induced skin abnormalities including cancers (Cleaver and Kraemer, 1989). Complementation tests by cell fusion have provided evidence for the existence of at least seven NER-deficient complementation groups: XP-A to XP-G.

Another important category of mammalian mutants is the class of laboratory-induced, UV-sensitive rodent cell lines. At least 11 NER complementation groups have been identified (Riboni *et al.*, 1992; Collins, 1993). DNA-mediated gene transfer has led to the cloning of human genes that correct the mutations in rodent complementation groups. These human genes are named 'excision repair cross complementing rodent repair deficiency' (*ERCC*) genes, followed by a number referring to the corrected complementation group. With the exception of *ERCC1* (van Duin *et al.*, 1989), all other cloned *ERCC* genes appeared to be also responsible for one of the XP defects or for one of the forms of another NER disorder, Cockayne's syndrome (CS). Thus *ERCC2*, *ERCC3*, *ERCC5* and *ERCC6* were found to be identical to the genes causing XP-D, XP-B, XP-G and CS-B, respectively (Weeda *et al.*, 1990; Fletjer *et al.*, 1992; Troelstra *et al.*, 1992; O'Donovan and Wood, 1993; for a recent review see Hoeijmakers, 1993b). Hence, a considerable overlap exists between the rodent mutants and the human disorders. In addition, phenotypic correction of XP-cells by genomic or cDNA transfection has resulted in the cloning of the genes implicated in XP-A (the *XPAC* gene, for XP-A correcting; Tanaka *et al.*, 1990) and XP-C (the *XPCC* gene; Legerski and Peterson, 1992).

Sequence analysis has revealed a striking evolutionary

Table I. Purification of XP-C correcting protein from HeLa cells

	Protein (mg)	Activity (units)	Specific activity (units/mg)	Purification (fold)
Nuclear extract	1390	38 160	27.5	1
Phosphocellulose	634	22 360	35.3	1.28
Single-stranded DNA cellulose	1.86	15 680	8430	306
CM cosmogel	0.40	14 250	35 625	1295
Mono Q	0.23	12 780	55 565	2020

conservation. For all mammalian NER genes cloned to date, (presumed) yeast counterparts have been found (Hoeijmakers, 1993b; A.van Gool, C.Troelstra and J.H.J.Hoeijmakers, unpublished data). In *Saccharomyces cerevisiae* a minimum of 11 distinct NER mutants has been identified, collectively designated the *RAD3* epistasis group. The degree of similarity between the human and yeast genes strongly suggests that the NER pathways in both extremes of the eukaryotic spectrum are largely superimposable and mechanistically very related. However, for several yeast NER genes a mammalian equivalent is still lacking.

Another powerful tool for unravelling the molecular mechanism of excision repair is an *in vitro* system based on cell-free extracts capable of performing NER on a damaged naked DNA template. We have recently adapted this system originally developed by Wood *et al.* (1988) and Sibghat-Ullah *et al.* (1989) to the use of SV40 minichromosomes (Sugasawa *et al.*, 1993; Masutani *et al.*, 1993). Using this system as an assay, we report here the purification to homogeneity of a 125 kDa XP-C correcting protein from HeLa cells, the cloning of the corresponding cDNA, as well as the co-purification and cDNA cloning of a tightly associated protein of 58 kDa. The latter turned out to be homologous to the yeast RAD23 NER protein, thus filling in one of the remaining gaps in the parallels between yeast and man. Interestingly, a second human homologue of RAD23 was identified as well. Both human homologues of RAD23 (designated HHR23A and HHR23B) harbour a ubiquitin-like N-terminal domain. The XPCC–HHR23B complex is suspected to play a selective role in the genome-overall NER subpathway, since the repair defect in XP-C is limited to the genome-overall system (Kantor *et al.*, 1990; Venema *et al.*, 1990, 1991).

Results

Purification of the XP-C correcting protein from HeLa cells

A cell-free system for DNA repair was constructed in which UV-damaged SV40 minichromosomes can be repaired during an incubation with extracts from human cells (Sugasawa *et al.*, 1993). The system contains UV-irradiated or unirradiated SV40 minichromosomes as well as unirradiated pUC19 supercoiled DNA. The following evidence indicates that DNA synthesis with UV-irradiated chromosomes is due to excision repair of UV-induced damage: (i) it is defective in extracts from all excision-deficient XP complementation groups, (ii) it is stimulated by the addition of T4 endonuclease V to XP extracts, (iii) it is complemented by mixing XP cell extracts of different complementation groups, (iv) it is complemented by the addition of purified XP-A complementing (*XPAC*) gene product to XP-A cell extracts, and (v) it is inhibited by the addition of antiserum raised against *XPAC* protein to repair-proficient cell extracts (Masutani *et al.*, 1993).

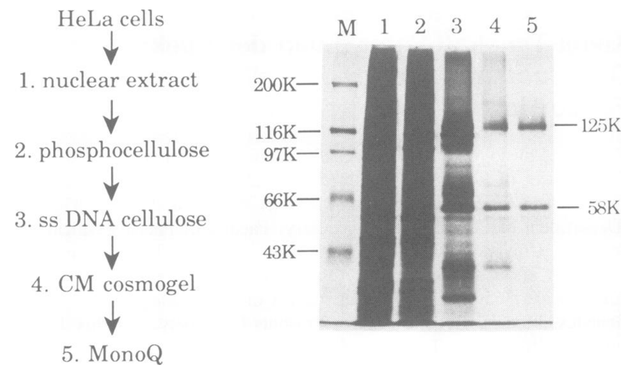

Fig. 1. Purification of the XP-C correcting protein from HeLa cells. The purification procedure is shown on the left. A sample at each purification step (indicated by numbers) was subjected to SDS–PAGE (8% polyacrylamide) and stained with silver. The marker proteins used were myosin, β-galactosidase, phosphorylase B, bovine serum albumin and ovalbumin (lane M).

One important use of cell-free systems is for fractionation and biochemical identification of factors involved in the reactions. We used our cell-free DNA repair system for purification of a protein that corrects DNA repair defects of XP-C cell extracts. Activity that complements the repair defect of XP4PASV (group C) cell extracts was assayed in the cell-free system. XP-C complementing activity was detected in nuclear extracts from HeLa cells and purified by successive column chromatographies on phosphocellulose, single-stranded DNA–cellulose, FPLC CM cosmogel and FPLC Mono Q (HR5/5) (for details see Materials and methods). The XP-C correcting activity bound strongly to a single-stranded DNA–cellulose column (being eluted between 0.6 and 1.5 M KCl), suggesting that the protein associates with DNA in cells. The purification procedure yielded a good recovery of the activity and ~2000-fold increase in its activity over that of the starting material (Table I). After FPLC Mono Q column chromatography, two polypeptides with apparent molecular masses of 125 and 58 kDa (p125 and p58) were detected by SDS–PAGE (Figure 1). As shown in Figure 2A, the XP-C correcting activity was eluted from a Sephacryl S-300 column at a position corresponding to a molecular weight of 500–550 kDa as estimated by a linear extrapolation. The two polypeptides were co-eluted with the activity (Figure 2B), indicating that these polypeptides form a physical complex and are associated with the XP-C correcting activity. Although the estimated molecular weight is much bigger than the sum of 125 and 58 kDa, it is unlikely that it is due to the protein aggregation, because we employed the gel

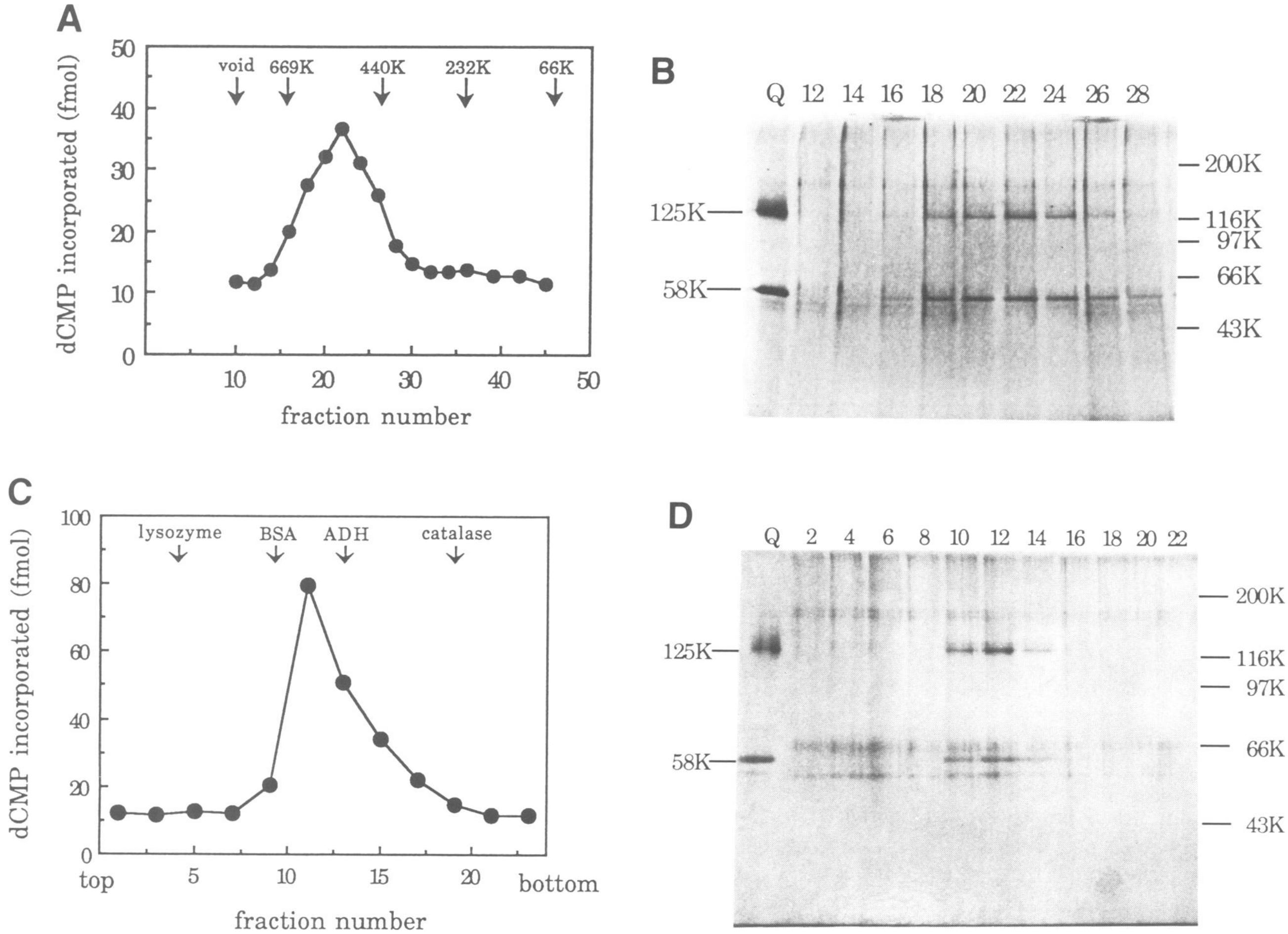

Fig. 2. Physical properties of the XP-C correcting protein. (**A**) Purified XP-C correcting protein was subjected to Sephacryl S-300 column chromatography and the XP-C correcting activity in eluted fractions was assayed as described in Materials and methods. The incorporation of radioactive materials into UV-irradiated mini-chromosomes was quantified. The positions of elution of marker proteins (thyroglobulin, ferritin, catalase and bovine serum albumin) fractionated under identical conditions are indicated by their molecular weights. (**B**) Samples (20 μl) of fractions around the peak of activity in panel A were subjected to SDS–PAGE (8% polyacrylamide) and stained with silver. (**C**) Purified XP-C correcting protein was subjected to glycerol density gradient centrifugation and assayed for XP-C correcting activity. The sedimentation positions of marker proteins in a parallel gradient are indicated. (**D**) Samples (20 μl) of fractions in panel C were subjected to SDS–PAGE (8% polyacrylamide) and stained with silver.

filtration in the presence of 0.3 M KCl, 10% glycerol and 0.01% Triton X-100. In fact, the activity sedimented at 6.2S (Figure 2C) on glycerol density gradient centrifugation under the same solution condition as in the gel filtration except for the various concentrations of glycerol. Again the p125 and p58 polypeptides co-migrated with the activity (Figure 2D). The molecular weight of the p125–p58 protein complex was estimated to be 110 kDa from the sedimentation position in the glycerol gradient, much smaller than that predicted from the results of gel filtration analysis and even smaller than the sum of 125 and 58 kDa, suggesting that the XP-C correcting protein is laminar in shape. We note that neither of the two proteins has the 93 kDa molecular weight predicted for the XPCC gene product, encoded by the cDNA cloned recently by Legerski and Peterson (1992). The purified XP-C correcting protein was tested for various enzymatic activities. No detectable DNA polymerase, DNA helicase, DNA ligase, DNA exonuclease or DNA endonuclease activity with UV-irradiated or unirradiated DNA was found under the conditions described in Materials and methods.

Specificity of complementation by the XP-C correcting protein fraction

The specificity of the activity of the p125–p58 protein preparation to complement defects of XP-C cell extracts was examined. Addition of 10 ng of the purified XP-C correcting protein to extracts from two XP-C cells (XP4PASV and XP3KA) induced a correction of the UV-specific repair synthesis to a level comparable with that of a repair-proficient cell extract (Figure 3A and B). In contrast, no significant increase in UV-dependent incorporation in the SV40 minichromosomes was observed in cell-free extracts from any of the six remaining excision-deficient XP complementation groups (Figure 3B).

Isolation of the cDNA encoding the p125 subunit

To clone the cDNA for the p125–p58 protein complex, the two polypeptides were separated from each other by gel filtration in the presence of guanidine–HCl (separation under physiological conditions failed). CNBr cleavage yielded completely different peptide profiles for the two proteins. Thus it is unlikely that p58 is a proteolytic product of the

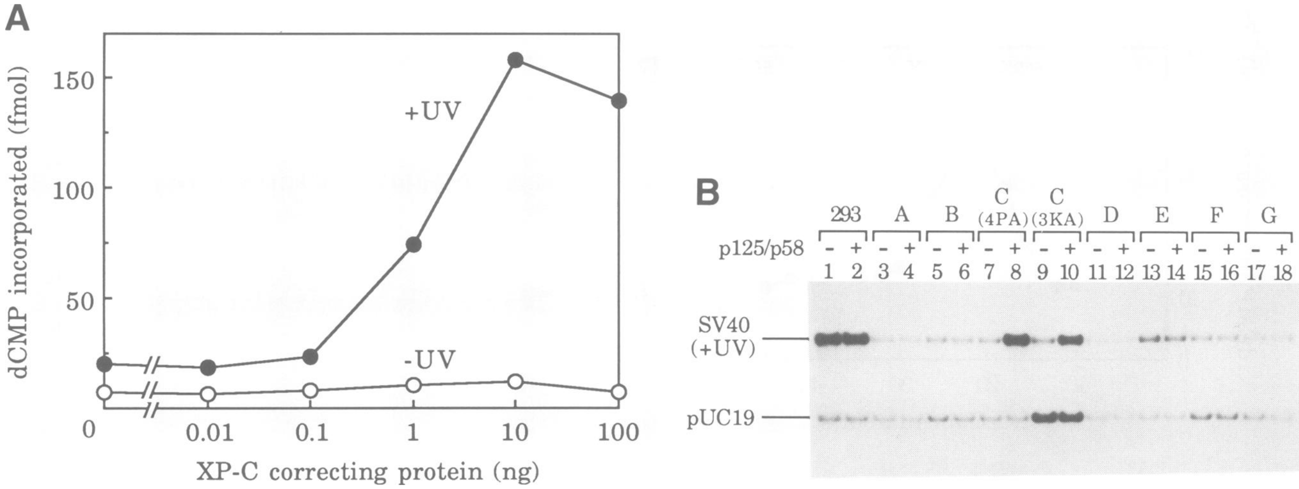

Fig. 3. Specificity of complementation by p125−p58 XP-C correcting protein. (**A**) Dose−response of XP-C correction. UV-irradiated (closed circles) or unirradiated (open circles) SV40 minichromosomes were incubated in standard reaction mixtures with XP4PASV cell extracts with increasing amounts of the p125−p58 protein complex. The incorporation into minichromosomes was quantified. (**B**) Complementation group specificity. UV-irradiated SV40 mini-chromosomes were incubated in the standard reaction mixture in the presence (even numbered lanes) or absence (odd numbered lanes) of the p125−p58 protein complex (10 ng) with 80 μg of protein of 293 (lanes 1 and 2), XP2OSSV (XP-A) (lanes 3 and 4), CRL1199 (XP-B) (lanes 5 and 6), XP4PASV (XP-C) (lanes 7 and 8), XP3KA(XP-C) (lanes 9 and 10), XP6BESV (XP-D) (lanes 11 and 12), XP2RO (XP-E) (lanes 13 and 14), XP2YOSV (XP-F) (lanes 15 and 16) or XP3BRSV (XP-G) (lanes 17 and 18) cell extracts. Purified DNA products were linearized with *Eco*RI and then subjected to 1% agarose gel electrophoresis as described in Materials and methods. An autoradiogram of the gel is shown. Although a higher level of DNA synthesis was observed with pUC19 plasmid DNA and the XP3KA cell extract than with other extracts, it must be an independent phenomenon from the repair event because the synthesis did not change on addition of p125−p58 protein complex in spite of the increase on addition of UV-irradiated chromosomes (lanes 9 and 10).

p125 subunit (data not shown). One partial amino acid sequence of p125 and two of p58 were determined, none of which matched with the predicted amino acid sequence of the previously cloned *XPCC* gene. The sequence of p125 was >50 amino acids in length. Since the same sequence was obtained for the undigested p125 polypeptide, it represents the N-terminal sequence of p125.

To prepare a DNA probe for screening cDNA libraries, two sets of oligonucleotide mixtures were synthesized according to the determined amino acid sequence of p125 (see Materials and methods) and used for the RT-PCR with poly(A)$^+$ RNA from HeLa cells. A PCR product of the expected length (132 bp) and nucleotide sequence was obtained and used for screening a λgt10 cDNA library prepared from HeLa cells. A positive clone with a 3.6 kb insert was obtained and its complete nucleotide sequence was determined (Figure 4). The first ATG, preceded by an in-frame stop codon, initiates an open reading frame (ORF) encoding 940 amino acids. The N-terminal part was entirely consistent with the experimentally determined partial amino acid sequence. In view of the different N-terminus, it was unexpected that at position 266 the sequence was found to be identical to the reported sequence for the *XPCC* gene (Legerski and Peterson, 1992). The predicted amino acid sequence of the p125 polypeptide was joined in-frame with the deduced ORF of the XPCC protein. As a result, there were 117 additional amino acids at the N-terminus of the XPCC protein and the calculated molecular mass increased from 93 to 106 kDa. Therefore, the deduced product of the reported *XPCC* gene is probably part of the p125 protein truncated in the N-terminal region. We infer that the p125 polypeptide represents the full-length XPCC gene product.

Legerski and Peterson (1992) reported that the putative XPCC protein shares limited homology with the *RAD4* gene product of *S.cerevisiae*. We could not find any significant additional homology with RAD4 or other proteins or any functional motifs in the newly identified N-terminal region of the p125.

Cloning and sequence analysis of the cDNA encoding the p58 subunit

To obtain a cDNA clone for the p58, an oligonucleotide mixture was synthesized according to one of the two determined amino acid sequences of p58 (see Materials and methods) and was used for screening a λgt10 cDNA library prepared from HeLa cells. A positive clone with a 2.9 kb insert was obtained and its complete nucleotide sequence was determined (Figure 5). An ORF encoding 409 amino acids including both determined amino acid sequences was found. Although the calculated molecular mass of the protein was only 43 kDa, we concluded that the clone includes the full length of the coding region of the p58 polypeptide because a termination codon (TGA) was found in frame in an upstream region of the putative initiation codon. Consistent with this notion is our finding that the protein overproduced in *E.coli* by the cloned cDNA migrates at the same position as the p58 protein (unpublished results).

Searches in various databases for sequence homology to the p58 ORF revealed several interesting features:

(i) At the nucleotide sequence level, two expressed sequence tags (ESTs) with unknown function representing partial human cDNA clones of brain and a liver cell line [accession numbers M85669 (Adams *et al.*, 1992) and D12303 (Okubo *et al.*, 1992)] were—with the exception of a few sequence uncertainties—identical to the corresponding part of the p58 cDNA sequence. These cDNAs are therefore expected to be derived from the p58 gene.

(ii) Amino acid sequence comparison uncovered significant resemblance between the N-terminal 79 amino acids of p58

```
TCGAAGGGGC GTGGCCAAGC GCACCGCCTC GGGGCGGGGC CGGCGTTCTA GCGCATCGCG GCCGGGTGCG TCACTCGCGA AGTGGAATTT    90
GCCCAGACAA GCAACATGGC TCGGAAACGC GCGGCCGGCG GGGAGCCGCG GGGACGCGAA CTGCGCAGCC AGAAATCCAA GGCCAAGAGC   180
                M  A   R  K  R   A  A  G   G  E  P   R  G  R   E  L  R   S  Q  K   S  K  A   K  S    (25)
AAGGCCCGGC GTGAGGAGGA GGAGGAGGAT GCCTTTGAAG ATGAGAAACC CCCAAAGAAG AGCCTTCTCT CCAAAGTTTC ACAAGGAAAG   270
K  A  R  R   E  E  E   E  E  D   A  F  E  D   E  K  P   P  K  K   S  L  L  S   K  V  S   Q  G  K    (55)
AGGAAAAGAG GCTGCAGTCA TCCTGGGGGT TCAGCAGATG GTCCAGCAAA AAAGAAAGTG GCCAAGGTGA CTGTTAAATC TGAAAACCTC   360
R  K  R  G   C  S  H   P  G  G   S  A  D  G   P  A  K   K  K  V   A  K  V  T   V  K  S   E  N  L    (85)
AAGGTTATAA AGGATGAAGC CCTCAGCGAT GGGGATGACC TCAGGGACTT TCCAAGTGAC CTCAAGAAGG CACACCATCT GAAGAGAGGG   450
K  V  I  K   D  E  A   L  S  D   G  D  D  L   R  D  F   P  S  D   L  K  K  A   H  H  L   K  R  G   (115)
GCTACCATGA ATGAAGACAG CAATGAAGAA GAGGAAGAAA GTGAAAATGA TTGGGAAGAG GTTGAAGAAC TTAGTGAGCC TGTGCTGGGT   540
A  T  M  N   E  D  S   N  E  E   E  E  S   E  N  D   W  E  E   V  E  E  L   S  E  P   V  L  G      (145)
GACGTGAGAG AAAGTACAGC CTTCTCTCGA TCTCTTCTGC CTGTGAAGCC AGTGGAGATA GAGATTGAAA CGCCAGAGCA GGCGAAGACA   630
D  V  R  E   S  T  A   F  S  R   S  L  L  P   V  K  P   V  E  I   E  I  E  T   P  E  Q   A  K  T   (175)
AGAGAAAGAA GTGAAAAGAT AAAACTGGAG TTTGAGACAT ATCTTCGGAG GGCGATGAAA CGTTTCAATA AAGGGGTCCA TGAGGACACA   720
R  E  R  S   E  K  I   K  L  E   F  E  T  Y   L  R  R   A  M  K   R  F  N  K   G  V  H   E  D  T   (205)
CACAAGGTTC ACCTTCTCTG CCTGCTAGCA AATGGCTTCT ATCGAAATAA CATCTGCAGC CAGCCAGATC TGCATGCTAT TGGCCTGTCC   810
H  K  V  H   L  L  C   L  L  A   N  G  F  Y   R  N  N   I  C  S   Q  P  D  L   H  A  I   G  L  S   (235)
ATCATCCCAG CCCGCTTTAC CAGAGTGCTG CCTCGAGATG TGGACACCTA CTACCTCTCA AACCTGGTGA AGTGGTTCAT TGGAACATTT   900
I  I  P  A   R  F  T   R  V  L   P  R  D  V   D  T  Y   Y  L  S   N  L  V  K   W  F  I   G  T  F   (265)
ACAGTTAATG CAGAACTTTC AGCCAGTGAA CAAGATAACC TGCAGACTAC ATTGGAAAGG AGATTTGCTA TTTACTCTGC TCGAGATGAT   990
T  V  N  A   E  L  S   A  S  E   Q  D  N  L   Q  T  T   L  E  R   R  F  A  I   Y  S  A   R  D  D   (295)
GAGGAATTGG TCCATATATT CTTACTGATT CTCCGGGCTC TGCAGCTCTT GACCCGGCTG GTATTGTCTC TACAGCCAAT TCCTCTGAAG  1080
E  E  L  V   H  I  F   L  L  I   L  R  A  L   Q  L  L   T  R  L   V  L  S  L   Q  P  I   P  L  K   (325)
TCAGCAACAG CAAAGGGAAA GAAACCTTCC AAGGAAAGAT TGACTGCGGA TCCAGGAGGC TCCTCAGAAA CTTCCAGCCA AGTTCTAGAA  1170
S  A  T  A   K  G  K   K  P  S   K  E  R  L   T  A  D   P  G  G   S  S  E  T   S  S  Q   V  L  E   (355)
AACCACACCA AACCAAAGAC CAGCAAAGGA ACCAAACAAG AGGAAACCTT TGCTAAGGGC ACCTGCAGGC CAAGTGCCAA AGGGAAGAGG  1260
N  H  T  K   P  K  T   S  K  G   T  K  Q  E   E  T  F   A  K  G   T  C  R  P   S  A  K   G  K  R   (385)
AACAAGGGAG GCAGAAAGAA ACGGAGCAAG CCCTCCTCCA GCGAGGAAGA TGAGGGCCCA GGAGACAAGC AGGAGAAGGC AACCCAGCGA  1350
N  K  G  G   R  K  K   R  S  K   P  S  S  S   E  E  D   E  G  P   G  D  K  Q   E  K  A   T  Q  R   (415)
CGTCCGCATG GCCGGGAGCG GCGGGTGGCC TCCAGGGTGT CTTATAAAGA GGAGAGTGGG AGTGATGAGG CTGGCAGCGG CTCTGATTTT  1440
R  P  H  G   R  E  R   R  V  A   S  R  V  S   Y  K  E   E  S  G   S  D  E  A   G  S  G   S  D  F   (445)
GAGCTCTCCA GTGGAGAAGC CTCTGATCCC TCTGATGAGG ATTCCGAACC TGGCCCTCCA AAGCAGAGGA AAGCCCCCGC TCCTCAGAGG  1530
E  L  S  S   G  E  A   S  D  P   S  D  E  D   S  E  P   G  P  P   K  Q  R  K   A  P  A   P  Q  R   (475)
ACAAAGGCTG GGTCCAAGAG TGCCTCCAGG ACCCATCGTG GGAGCCATCG TAAGGACCCA AGCTTGCCAG TGGCATCCTC AAGCTCTTCA  1620
T  K  A  G   S  K  S   A  S  R   T  H  R  G   S  H  R   K  D  P   S  L  P  V   A  S  S   S  S  S   (505)
AGCAGTAAAA GAGGCAAGAA AATGTGCAGC GATGGTGAGA AGGCAGAAAA AAGAAGCATA GCTGGTATAG ACCAGTGGCT AGAGGTGTTC  1710
S  S  K  R   G  K  K   M  C  S   D  G  E  K   A  E  K   R  S  I   A  G  I  D   Q  W  L   E  V  F   (535)
TGTGAGCAGG AGGAAAAGTG GGTATGTGTA GACTGTGTGC ACGGTGTGGT GGGCCAGCCT CTGACCTGTT ACAAGTACGC CACCAAGCCC  1800
C  E  Q  E   E  K  W   V  C  V   D  C  V  H   G  V  V   G  Q  P   L  T  C  Y   K  Y  A   T  K  P   (565)
ATGACCTATG TGGTGGGCAT TGACAGTGAC GGCTGGGTCC GAGATGTCAC ACAGAGGTAC GACCCAGTCT GGATGACAGT GACCCGCAAG  1890
M  T  Y  V   V  G  I   D  S  D   G  W  V  R   D  V  T   Q  R  Y   D  P  V  W   M  T  V   T  R  K   (595)
TGCCGGGTTG ATGCTGAGTG GTGGGCCGAG ACCTTGAGAC CATACCAGAG CCCATTTATG GACAGGGAGA AGAAAGAAGA CTTGGAGTTT  1980
C  R  V  D   A  E  W   W  A  E   T  L  R  P   Y  Q  S   P  F  M   D  R  E  K   K  E  D   L  E  F   (625)
CAGGCAAAAC ACATGGACCA GCCTTTGCCC ACTGCCATTG GCTTATATAA GAACCACCCT CTGTATGCCC TGAAGCGGCA TCTCCTGAAA  2070
Q  A  K  H   M  D  Q   P  L  P   T  A  I  G   L  Y  K   N  H  P   L  Y  A  L   K  R  H   L  L  K   (655)
TATGAGGCCA TCTATCCCGA GACAGCTGCC ATCCTTGGGT ATTGTCGTGG AGAAGCGGTC TACTCCAGGG ATTGTGTGCA CACTCTGCAT  2160
Y  E  A  I   Y  P  E   T  A  A   I  L  G  Y   C  R  G   E  A  V   Y  S  R  D   C  V  H   T  L  H   (685)
TCCAGGGACA CGTGGCTGAA GAAAGCAAGA GTGGTGAGGC TTGGAGAAGT ACCCTACAAG ATGGTGAAAG GCTTTTCTAA CCGTGCTCGG  2250
S  R  D  T   W  L  K   K  A  R   V  V  R  L   G  E  V   P  Y  K   M  V  K  G   F  S  N   R  A  R   (715)
AAAGCCCGAC TTGCTGAGCC CCAGCTGCGG GAAGAAAATG ACCTGGGCCT GTTTGGCTAC TGGCAGACAG AGGAGTATCA GCCCCCAGTG  2340
K  A  R  L   A  E  P   Q  L  R   E  E  N  D   L  G  L   F  G  Y   W  Q  T  E   E  Y  Q   P  P  V   (745)
GCCGTGGACG GGAAGGTGCC CCGGAACGAG TTTGGGAATG TGTACCTCTT CCTGCCCAGC ATGATGCCTA TTGGCTGTGT CCAGCTGAAC  2430
A  V  D  G   K  V  P   R  N  E   F  G  N  V   Y  L  F   L  P  S   M  M  P  I   G  C  V   Q  L  N   (775)
CTGCCCAATC TACACCGCGT GGCCCGCAAG CTGGACATCG ACTGTGTCCA GGCCATCACT GGCTTTGATT TCCATGGCGG CTACTCCCAT  2520
L  P  N  L   H  R  V   A  R  K   L  D  I  D   C  V  Q   A  I  T   G  F  D  F   H  G  G   Y  S  H   (805)
CCCGTGACTG ATGGATACAT CGTCTGCGAG GAATTCAAAG ACGTGCTCCT GACTGCCTGG GAAAATGAGC AGGCAGTCAT TGAAAGGAAG  2610
P  V  T  D   G  Y  I   V  C  E   E  F  K  D   V  L  L   T  A  W   E  N  E  Q   A  V  I   E  R  K   (835)
GAGAAGGAGA AAAAGGAGAA GCGGGCTCTA GGGAACTGGA AGTTGCTGGC CAAAGGTCTG CTCATCAGGG AGAGGCTGAA GCGTCGCTAC  2700
E  K  E  K   K  E  K   R  A  L   G  N  W  K   L  L  A   K  G  L   L  I  R  E   R  L  K   R  R  Y   (865)
GGGCCCAAGA GTGAGGCAGC AGCTCCCCAC ACAGATGCAG GAGGTGGACT CTCTTCTGAT GAAGAGGAGG GGACCAGCTC TCAAGCAGAA  2790
G  P  K  S   E  A  A   A  P  H   T  D  A  G   G  G  L   S  S  D   E  E  E  G   T  S  S   Q  A  E   (895)
GCGGCCAGGA TACTGGCTGC CTCCTGGCCT CAAAACCGAG AAGATGAAGA AAAGCAGAAG CTGAAGGGTG GGCCCAAGAA GACCAAAAGG  2880
A  A  R  I   L  A  A   S  W  P   Q  N  R  E   D  E  E   K  Q  K   L  K  G  G   P  K  K   T  K  R   (925)
GAAAAGAAAG CAGCAGCTTC CCACCTGTTC CCATTTGAGA AGCTGTGAGC TGAGCGCCCA CTAGAGGGGC ACCCACCAGT TGCTGCTGCC  2970
E  K  K  A   A  A  S   H  L  F   P  F  E  K   L  *                                                (940)
CCACTACAGG CCCCACACCT GCCCTGGGCA TGCCCAGCCC CTGGTGGTGG GGGGTTCTCT GCTGAGAAGG CAAACTGAGG CAGCATGCAC  3060
GGAGGCGGGG TCAGGGGAGA CGAGGCCAAG CTGAGGAGGT GCTGCAGGTC CCGTCTGGCT CCAGCCCTTG TCAGATTCAC CCAGGGTGAA  3150
GCCTTCAAAG CTTTTTGCTA CCAAAGCCCA CTCACCCTTT GAGCTACAGA ACACTTTGCT AGGAGATACT CTTCTGCCTC CTAGACCTGT  3240
TCTTTCCATC TTTAGAAACA TCAGTTTTTG TATGGAAGCC ACCGGGAGAT TTCTGGATGG TGGTGCATCC GTGAATGCGC TGATCGTTTC  3330
TTCCAGTTAG AGTCTTCATC TGTCCGACAA GTTCACTCGC CTCGGTTGCG GACCTAGGAC CATTTCTCTG CAGGCCACTT ACCTTCCCCT  3420
GAGTCAGGCT TACTAATGCT GCCCTCACTG CCTCTTTGCA GTAGGGGAGA GAGCAGAGAA GTACAGGTCA TCTGCTGGGA TCTAGTTTTC  3510
CAAGTAACAT TTTGTGGTGA CAGAAGCCTA AAAAAAGCTA AAATCAGA                                                3558
```

Fig. 4. Nucleotide and predicted amino acid sequence of the p125/XPCC. Top numbers on the right are those of nucleotide residues and lower ones (in parentheses) are those of amino acids. A termination codon, TAG, in the 5′ untranslated region is boxed. Two circled methionines are putative initiation codons for the p125 of the XP-C correcting protein and the previously reported XPCC protein (Legerski and Peterson, 1992), respectively. The arrow indicates the start position of the reported sequence. The asterisk indicates the termination codon, TGA, for this ORF. Doubly underlined amino acids represent a peptide sequence derived from the purified p125 polypeptide. Boxed nucleotides (nucleotide positions: 286, 1601, 2166 and 3024) are different from those in the sequence reported by Legerski and Peterson (1992). The GenBank accession number for human XPCC (p125) is D21089.

```
TAGCGATTCC CTGCTTGTCT CGCCGACCCC CTCGCGCCTT CTGCAGACTC CGTGGCTGGC GCTCGGCGCG TGAGGAAGCA CGGCGGCCCG   90
AGTTCGCGGG GAAGGCCGCA GTCGCGGAGG CAGCGGCGCG GTCCGGGGCA CGGGCTGGGG GAGAGGCCGC TCCGCTGGGC GAATCTGACA  180
AGCCCCCACC CCCACCGCCT TCCTCCCCAG AGCGCGAGGA GCGCGGGCGA CCCCGGGGCC CCGCCAGGCC ACAGACCCCG CCCAGCGGCC  270
AGCACCCGGC GCAGGCCCGG CAGCCGAGCT GCGCGGCGGC ACCATGCAGG TCACCCTGAA GACCCTCCAG CAGCAGACCT TCAAGATAGA  360
                                              M Q V    T L K    T L Q    Q Q T F    K I D  (16)
CATTGACCCC GAGGAGACGG TGAAAGCACT GAAAGAGAAG ATTGAATCTG AAAAGGGGAA AGATGCCTTT CCAGTAGCAG GTCAAAAATT  450
 I D P    E E T V    K A L    K E K    I E S E    K G K    D A F    P V A G    Q K L  (46)
AATTTATGCA GGCAAAATCC TCAATGATGA TACTGCTCTC AAAGAATATA AAATTGATGA GAAAAACTTT GTGGTGGTTA TGGTGACCAA  540
 I Y A    G K I L    N D D    T A L    K E Y K    I D E    K N F    V V V M    V T K  (76)
ACCCAAAGCA GTGTCCACAC CAGCACCAGC TACAACTCAG CAGTCAGCTC CTGCCAGCAC TACAGCAGTT ACTTCCTCCA CCACCACAAC  630
 P K A    V S T P    A P A    T T Q    Q S A P    A S T    T A V    T S S T    T T T  (106)
TGTGGCTCAG GCTCCAACCC CTGTCCCTGC CTTGGCCCCC ACTTCCACAC CTGCATCCAT CACTCCAGCA TCAGCGACAG CATCTTCTGA  720
 V A Q    A P T P    V P A    L A P    T S T P    A S I    T P A    S A T A    S S E  (136)
ACCTGCACCT GCTAGTGCAG CTAAACAAGA GAAGCCTGCA GAAAAGCCAG CAGAGACACC AGTGGCTACT AGCCCAACAG CAACTGACAG  810
 P A P    A S A A    K Q E    K P A    E K P A    E T P    V A T    S P T A    T D S  (166)
TACATCGGGT GATTCTTCTC GGTCAAACCT TTTTGAAGAT GCAACGAGTG CACTTGTGAC GGGTCAGTCT TACGAGAATA TGGTAACTGA  900
 T S G    D S S R    S N L    F E D    A T S A    L V T    G Q S    Y E N M    V T E  (196)
GATCATGTCA ATGGGCTATG AACGAGAGCA AGTAATTGCA GCCCTGAGAG CCAGTTTCAA CAACCCTGAC AGAGCAGTGG AGTATCTTTT  990
 I M S    M G Y E    R E Q    V I A    A L R A    S F N    N P D    R A V E    Y L L  (226)
AATGGGAATC CCTGGAGATA GAGAAAGTCA GGCTGTGGTT GACCCCCCTC AAGCAGCTAG TACTGGGGCT CCTCAGTCTT CAGCAGTGGC 1080
 M G I    P G D R    E S Q    A V V    D P P Q    A A S    T G A    P Q S S    A V A  (256)
TGCAGCTGCA GCAACTACGA CAGCAACAAC TACAACAACA AGTTCTGGAG GACATCCCCT TGAATTTTTA CGGAATCAGC CTCAGTTTCA 1170
 A A A    A T T T    A T T    T T T    S S G G    H P L    E F L    R N Q P    Q F Q  (286)
ACAGATGAGA CAAATTATTC AGCAGAATCC TTCCTTGCTT CCAGCGTTAC TACAGCAGAT AGGTCGAGAG AATCCTCAAT TACTTCAGCA 1260
 Q M R    Q I I Q    Q N P    S L L    P A L L    Q Q I    G R E    N P Q L    L Q Q  (316)
AATTAGCCAA CACCAGGAGC ATTTTATTCA GATGTTAAAT GAACCAGTTC AAGAAGCTGG TGGTCAAGGA GGAGGAGGTG GAGGTGGCAG 1350
 I S Q    H Q E H    F I Q    M L N    E P V Q    E A G    G Q G    G G G G    G G S  (346)
TGGAGGAATT GCAGAAGCTG GAAGTGGTCA TATGAACTAC ATTCAAGTAA CACCTCAGGA AAAAGAAGCT ATAGAAAGGT TAAAGGCATT 1440
 G G I    A E A G    S G H    M N Y    I Q V T    P Q E    K E A    I E R L    K A L  (376)
AGGATTTCCT GAAGGACTTG TGATACAAGC GTATTTTGCT TGTGAGAAGA ATGAGAATTT GGCTGCCAAT TTTCTTCTAC AGCAGAACTT 1530
 G F P    E G L V    I Q A    Y F A    C E K N    E N L    A A N    F L L Q    Q N F  (406)
TGATGAAGAT TGAAAGGGAC TTTTTTATAT CTCACACTTC ACACCAGTGC ATTACACTAA CTTGTTCACT GGATTGTCTG GGATGACTTG 1620
 D E D    *                                                                             (409)
GGCTCATATC CACAATACTT GGTATAAGGT AGTAGATTGT TGGGGGTGGG GAGGGAGGGA TCTAGGATAC AGGGCAGGGA TAAATACAGT 1710
GCATGTCTGC TTCAATTAGC AGATGCCGCA ACTCCACACA GTGTGTAAAA TATATACAAC CAAAAATCAG CTTTTGCAGG TCTTTATTTC 1800
TTCTGTAAAA CAGTAGGTAA CTTTTCCTAG GTTTCACTCT TTTTAGTGTA CTAGATCCAG AAACTTAGTG TAATGCCCTG CTTTATATAT 1890
CTTTGACTTA ACATTGGTTT CAGAAAGAAT CTTAGCTACC TAGAATTTAC AGTCTCTGTT TCATGGCAAC ACTGGATAAT GGCTTTGTGA 1980
AATTTAAAAA ATTTTTGTAG CGACTGTAAA CAGAAATGCC AAATTGATGG TTAATTGTTG CTGCTTCAAA AATAAGTATA AAATTAATAT 2070
GTAAGGAAGC CCATTCTTTC ATGTTAAATA CTTGGGGTGG GAGGGGAGAA AGGGAACCTT TTCTTAAAAT GAAAATAATT ACTGCTATTT 2160
TAAAATTTCT TGATCATTGA ATGTGAGACC CTTCTAACAT GATTTGAGAA GCTGTACAAG TATAGGCAGA GTTATTTTCC TGTTTACATT 2250
TTTTTTTTGT TTTGGGGAAA AAATTGGTAG GTGTCTAATT ACTGTTTACT TCATTGTTAT ATTGCAGTAA AAGTTTTAAA ACAACCATTG 2340
CATGTTTGCT TTTGATGTAT CCCTTTGTGA AATTAGCACT TTTGGGGCCA ATGGAGAAAT GCAGCATTCA CTCTCCCTGT CTTTTCCCCT 2430
TCCCTCAGCA GAAACGTGTT TATCAGCAAG TCGTGAGTCA AACTGCTGCC TTTTAAAAAA CCCACAAAAT GCTGATTCAG TTCAAAATTA 2520
ATGCAAATGT TTCAAAACTG GGTTTCTGAT ATTTGTAAAT GTGTTTCTTT ATTAGATAAG AGTGTATTAC CATTAAAGTC ATTAGTATAA 2610
TATTGCTTTC AAAAAGAAAT GGTAGACAAA ACTATAATCC AGCATCTTTT ATTGCATTGG AAAGACTGGC AAAGTCTTTT GGATGGGTTG 2700
GGAGATGTGG CTGGAAAGTA CTTTGGAAAA TATACAATCA AGATATCTCA TGGCATATTA AAAGAAAAAT CTTAATAGCA GTGTTGGCTT 2790
TTATTTGGAT TTTTTCATCT CAGTTTTTTC TGTGGAATCT CCTTCATTGG CATTGTTATT TAATCATAAA CGGGGCAGAT GTCTACTTGT 2880
TCAGTTTTTC AAATCTGTTT                                                                              2905
```

Fig. 5. Nucleotide and predicted amino acid sequence of the p58/HHR23B. Top numbers on the right are those of nucleotide residues and lower ones (in parentheses) are those of amino acids. An in-frame termination codon, TGA, in the 5′ untranslated region is boxed. The asterisk indicates the termination codon, TGA, for this ORF. Doubly underlined amino acids represent peptide sequences derived from the purified p58 polypeptide. Putative polyadenylation signals (ATTAAA) in the 3′ untranslated region are shown by bold boxes. Three ATTTA sequence motifs (mRNA degradation signals) are underlined. The GenBank accession number for human XPCC (p58/HHR23B) is D21090.

and ubiquitin and a similar domain in various ubiquitin fusion proteins (see below).

(iii) Interestingly, the p58 amino acid sequence appeared to share extensive overall sequence homology with the *S.cerevisiae RAD23* gene (Melnick and Sherman, 1993; sequence prior to publication kindly provided by S.Prakash, Galveston), a member of the *RAD3* NER epistasis group for which no human homologue has yet been identified. The *RAD23* gene is identical to the sygg-orf29 sequence, identified on chromosome 5 as part of the yeast genome sequencing project (accession number L10830).

(iv) Finally, using the BLAST algorithm (Altschul *et al.*, 1990), which is able to detect amino acid sequence homologies translated from all six possible frames, we identified several human partial cDNAs which exhibited some homology to the amino acid sequence of p58, when uncertainties in the sequence are taken into account. These cDNAs were derived from heart (accession number M77024) and a T lymphoblastoid cell-line (accession numbers Z15569, Z12748 and Z15568). Because of the presence of some sequence ambiguities and to find out whether this cDNA shared additional sequence similarity to p58, we decided to isolate the corresponding full-length cDNA by RT-PCR using total HeLa RNA combined with library screening.

The nucleotide and deduced amino acid sequence of the cDNA encoded by this p58-related gene, that we termed tentatively *HHR23A* for **h**uman **h**omologue of *RAD23* A is presented in Figure 6. The ORF, starting from the first ATG encodes an acidic protein (pI 4.4) of 363 amino acids, with a calculated molecular mass of 40 kDa. Also this protein synthesized in *E.coli* migrates well above its predicted molecular weight (P.J.van der Spek, unpublished results). The 3′ UTR harbours a canonical AATAAA polyadenylation signal 12 bp before the start of the poly(A) tail. As shown

```
GGGATCCCGG GGCCGCCGCG TCGCTCGGGC CCCGCCATGG CCGTCACCAT CACGCTCAAA ACGCTGCAGC AGCAGACCTT CAAGATCCGC    90
                                     M  A   V  T  I   T  L  K   T  L  Q  Q   Q  T  F   K  I  R    (18)
ATGGAGCCTG ACGAGACGGT GAAGGTGCTA AAGGAGAAGA TAGAAGCTGA GAAGGGTCGT GATGCCTTCC CCGTGGCTGG ACAGAAACTC   180
M  E  P  D   E  T  V   K  V  L   K  E  K  I   E  A  E   K  G  R   D  A  F  P   V  A  G   Q  K  L    (48)
ATCTATGCCG GCAAGATCTT GAGTGACGAT GTCCCTATCA GGGACTATCG CATCGATGAG AAGAACTTTG TGGTCGTCAT GGTGACCAAG   270
I  Y  A  G   K  I  L   S  D  D   V  P  I  R   D  Y  R   I  D  E   K  N  F  V   V  V  M   V  T  K    (78)
ACCAAAGCCG GCCAGGGTAC CTCAGCACCC CCAGAGGCCT CACCCACAGC TGCCCCAGAG TCCTCTACAT CCTTCCCGCC TGCCCCCACC   360
T  K  A  G   Q  G  T   S  A  P   P  E  A  S   P  T  A   A  P  E   S  S  T  S   F  P  P   A  P  T   (108)
TCAGGCATGT CCCATCCCCC ACCTGCCGCC AGAGAGGACA AGAGCCCATC AGAGGAATCC GCCCCCACGA CGTCCCCAGA GTCTGTGTCA   450
S  G  M  S   H  P  P   P  A  A   R  E  D  K   S  P  S   E  E  S   A  P  T  T   S  P  E   S  V  S   (138)
GGCTCTGTTC CCTCTTCAGG TAGCAGCGGG CGAGAGGAAG ACGCGGCCTC CACGCTAGTG ACGGGCTCTG AGTATGAGAC GATGCTGACG   540
G  S  V  P   S  S  G   S  S  G   R  E  E  D   A  A  S   T  L  V   T  G  S  E   Y  E  T   M  L  T   (168)
GAGATCATGT CCATGGGCTA TGAGCGAGAG CGGGTCGTGG CCGCCCTGAG AGCCAGCTAC AACAACCCCC ACCGAGCCGT GGAGTATCTG   630
E  I  M  S   M  G  Y   E  R  E   R  V  V  A   A  L  R   A  S  Y   N  N  P  H   R  A  V   E  Y  L   (198)
CTCACGGGAA TTCCTGGGAG CCCCGAGCCG GAACACGGTT CTGTCCAGGA GAGCCAGGTA TCGGAGCAGC CGGCCACGGA AGCAGCAGGA   720
L  T  G  I   P  G  S   P  E  P   E  H  G  S   V  Q  E   S  Q  V   S  E  Q  P   A  T  E   A  A  G   (228)
GAGAACCCCC TGGAGTTCCT GCGGGACCAG CCCCAGTTCC AGAACATGCG GCAGGTGATT CAGCAGAACC CTGCGCTGCT GCCCGCCCTG   810
E  N  P  L   E  F  L   R  D  Q   P  Q  F  Q   N  M  R   Q  V  I   Q  Q  N  P   A  L  L   P  A  L   (258)
CTCCAGCAGC TGGGCCAGGA GAACCCTCAG CTTTTACAGC AAATCAGCCG GCACCAGGAG CAGTTCATCC AGATGCTGAA CGAGCCCCCT   900
L  Q  Q  L   G  Q  E   N  P  Q   L  L  Q  Q   I  S  R   H  Q  E   Q  F  I  Q   M  L  N   E  P  P   (288)
GGGGAGCTGG CGGACATCTC AGATGTGGAG GGGGAGGTGG GCGCCATAGG AGAGGAGGCC CCGCAGATGA ACTACATCCA GGTGACGCCG   990
G  E  L  A   D  I  S   D  V  E   G  E  V  G   A  I  G   E  E  A   P  Q  M  N   Y  I  Q   V  T  P   (318)
CAGGAGAAAG AAGCTATAGA GAGGTTGAAG GCCCTGGGCT TCCCAGAGAG CCTGGTCATC CAGGCCTATT TCGCGTGTGA AAAAAATGAG  1080
Q  E  K  E   A  I  E   R  L  K   A  L  G  F   P  E  S   L  V  I   Q  A  Y  F   A  C  E   K  N  E   (348)
AACTTGGCTG CCAACTTCCT CCTGAGTCAG AACTTTGATG ACGAGTGATG CCAGGAAGCC AGGCCACCGA AGCCCCCACC CTACCCTTAT  1170
N  L  A  A   N  F  L   L  S  Q   N  F  D  D   E  *                                                (363)
TCCATGAAAG TTTTATAAAA GAAAAAATAT ATATATATTC ATGTTTATTT AAGAAATGGA AAAAAAAATC AAAAATCTTA AAAAAACAAG  1260
CAAACAGTCC AGCTTCCTGT CCTCCTAAAG TGGCCCCTGT TCCCATCTCC CGGGCCAGAC AGCTGTCCCC CCGTCCTCCT CCCCAGCCCA  1350
GCCTGCTCAG AGAAGCTGGC AGGACTGGGA GGCGACAGAT GGGCCCCTCT TGGCCTCTGT CCCAGCTCTC TGCAGCCAGA CGGAAAGGCG  1440
GCTGCTTGCC TCTCCATCCT CCGAAAAACC CCTGAGGACC CCCCCCCATC CTCTTCTAGG ATGAGGGGAA GCTGGAGCCC CAACTTTGAT  1530
CCTCCATTGG AGTGGCCCAA ATCTTTCCAT CTAGGGCAAG TCCTGAAAGG CCCAAGGCCC CCTCCCAGTC TGGCCTTGGC CTCCAGCCTG  1620
GAGAAGGGCT AACATCAGCT CATTGTCAAG GCCACCCCCA CCCCAGAACA GAACCGTGTC TCTGATAAAG GTTTTGAAGT GAATAAAGTT  1710
TTAAAAACTA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA                                                       1750
```

Fig. 6. Nucleotide and predicted amino acid sequence of the HHR23A. Top numbers on the right indicates the numbering of nucleotides; the numbers in parentheses correspond to those of the amino acids. The sequence 5′ proximal to the ATG matches perfect with the optimal translation initiation sequence (Kozak, 1991). The asterisk indicates the termination of the ORF. The polyadenylation site present in the 3′ untranslated region before the poly(A) tail is boxed. The GenBank accession number for HHR23A is D21235.

by the amino acid sequence alignment in Figure 7A, the two human proteins exhibit a high overall homology to each other (57% identity, 76% similarity) and to the yeast *RAD23* gene product (30–34% identity, 41% homology). Furthermore, it is worth noting that regions rich in S, T, P and A amino acids are found at two locations. The first starts immediately following the ubiquitin-like domain: residues 79–144; 84% of which is S, T, P, or A (figures for p58). The second runs from residues 241 to 272, 87% of which is made up of these residues (figures for p58). Finally, a glycine-rich stretch is present in p58 between residues 336–348.

The alignment of all three RAD23 homologues with (human) ubiquitin and with similar domains in other ubiquitin-like fusion proteins is presented in Figure 7B. The level of homology to ubiquitin is very similar for all three polypeptides (25–31% identity, 55–59% similarity) and is in the same range as that of other ubiquitin hybrid polypeptides. We conclude that both human proteins belong to the family of ubiquitin-fusion proteins and represent homologues of RAD23. Consistent with the designation HHR23A, we term the p58 HHR23B. Apart from the ubiquitin motif, no other functional domains could be identified in the HHR23/RAD23 sequence using the PROSITE software package or comparison to other proteins.

Discussion

XP-C correcting protein

XP-C is one of the most common forms of XP (Kraemer *et al.*, 1987). Group C patients display the (for XP obligate) features of hypersensitivity to sunlight (UV) and other cutaneous manifestations, including predisposition to skin cancer, but a second hallmark, accelerated neurodegeneration, is absent. Recently, the NER defect in XP-C was pinpointed to the genome-overall subpathway; 'transcription-coupled repair' functions normally in these cells (Kantor *et al.*, 1990; Venema *et al.*, 1990, 1991). This provides a plausible explanation for the relatively high cellular resistance to UV. Furthermore, transcription-coupled NER may be important for counteracting neurodegeneration. However, since this repair process is limited only to the transcribed strand of active genes, it has no effect on mutagenesis in the non-transcribed strand nor in the rest of the genome. Presumably, this explains why XP-C patients cannot effectively avert sunlight-induced skin cancer. Here we have purified a protein complex that based on the nature of the XP-C mutation is expected to operate specifically in the 'genome-overall' repair pathway. More recently, six distinct mutations including point mutations, deletions and insertions were detected in the *XPCC* gene of five XP-C cell lines (Li *et al.*, 1993). Thus a defect in the p125 subunit gives rise to cancer proneness. The complex consists of two tightly associated polypeptides: a 125 kDa species representing the XP-C gene product and a 58 kDa protein, which turned out to be a human homologue of *S.cerevisiae RAD23*, one of the remaining yeast NER genes for which no human counterpart was known. Unexpectedly, a second human equivalent of *RAD23* appeared to exist. All *RAD23* homologues share an N-terminal ubiquitin-like domain.

A DNA-dependent ATPase, designated ATPase Q1, was previously found to be altered in XP-C cells in terms of its

A

```
RAD23         1  M.VSLTFKNFKKEKVPLDLEPSNTILETKTKLAQSISCEESQI...KLIYSGKVLQDSKT
HHR23A        1  MAVTITLKTLQQQTFKIRMEPDETVKVLKEKIEAEKGRDAFPVAGQKLIYAGKILSDDVP
HHR23B p58    1  M..QVTLKTLQQQTFKIDIDPEETVKALKEKIESEKGKDAFPVAGQKLIYAGKILNDDTA

RAD23        57  VSECGLKDGDQVVFMVSQKKSTKTKVTEPPIAPESATTPGRENSTEASPSTDASAAPAAT
HHR23A       61  IRDYRIDEKNFVVVMVTKTKAGQGTSAPPEASPTAA..PESSTSFPPAPTSGMSHPPPAA
HHR23B p58   59  LKEYKIDEKNFVVVMVTKPKAVSTPAPATTQQSAPASTTAVTSSTTTTVAQAPTPVPALA

RAD23       117  APEGSQPQEEQTATTERTESASTPGF..................................
HHR23A      119  REDKS.PSEESAPTTSPESVSGSVP...........................SSGSSGREE
HHR23B p58  119  PTSTPASITPASATASSEPAPASAAKQEKPAEKPAETPVATSPTATDSTSGDSSRSNLFE

RAD23       143  .....VV.GTERNETIERIMEMGYQREEVERALRAAFNNPDRAVEYLLMGIPENLRQPEP
HHR23A      152  DAASTLVTGSEYETMLTEIMSMGYERERVVAALRASYNNPHRAVEYLLTGIP...GSPEP
HHR23B p58  179  DATSALVTGQSYENMVTEIMSMGYEREQVIAALRASFNNPDRAVEYLLMGIP...GDRES

RAD23       197  QQQTAAAAEQPSTAATTAEQPAEDDLFAQAAQGGNASSGALGTTGGATDAAQGGPPGSIG
HHR23A      209  EH....GSVQES...QVSEQPA.........................TEAA.GENPLEFL
HHR23B p58  236  Q................AVVDPPQAASTGAPQSSAVAAAAATTTATTTTSSGGHPLEFL

RAD23       257  LTVEDLLSLRQVVSGNPEALRPLLENISARYPQLREHIMANPEVFVSMLLEAVGDNMQDV
HHR23A      236  RDQPQFQNMRQVIQQNPALLPALLQQLGQENPQLLQQISRHQEQFIQMLNEPPGEL....
HHR23B p58  280  RNQPQFQQMRQIIQQNPSLLPALLQQIGRENPQLLQQISQHQEHFIQMLNEPVQEA....

RAD23       317  MEGADDMVEGEDIEVTGEAAAAGLGQGEGEGSFQVDYTPEDDQAISRLCELGFERDLVIQ
HHR23A      292  .........ADISDVEGEVGAIGEEAPQ.MNYIQV..TPQEKEAIERLKALGFPESLVIQ
HHR23B p58  336  ........GGQGGGGGGGSGGIAEAGSGHMNYIQV..TPQEKEAIERLKALGFPEGLVIQ

RAD23       377  VYFACDKNEEAAANILFSDHAD
HHR23A      340  AYFACEKNENLAANFLLSQNFDDE
HHR23B p58  386  AYFACEKNENLAANFLLQQNFDED
```

B

```
UBIQ. hum.  MQ..IFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQ...QRLIFAGKQLEDGRT
RAD23       M.VSLTFKNFKKEKVPLDLEPSNTILETKTKLAQSISCEESQI...KLIYSGKVLQDSKT
HHR23A      MAVTITLKTLQQQTFKIRMEPDETVKVLKEKIEAEKGRDAFPVAGQKLIYAGKILSDDVP
HHR23B p58  MQ..VTLKTLQQQTFKIDIDPEETVKALKEKIESEKGKDAFPVAGQKLIYAGKILNDDTA

NEDD8       ML..IKVKTLTGKEIEIDIEPTDKVERIKERVEEKEGIPPQQ...QRLIYSGKQMNDEKT
An1a        ME..LFIETLTGTCFELRVSPYETVTSVKSKIQRLEGIPVAQ...QHLIRNNMELEDECS
An1b        ME..LFIETLTGTCFELRVSPYETVTSVKSKIQRLEGIPVAQ...QHLIWNNMELEDECS
GdX         MQ..LTVKALQGRECSLQVPEDELVSTLKQLVSEKLNVPVRQ...QRLLFKGKALADGKR
BAT3        LE..VLVKTLDSQTRTFIVGAQMNVKEFKEHIRASVSIPSEK...QRLIYQGRVLQDDKK
fau         MQ..LFVRAQELH..TFEVTGQETVAQIKAHVASLEGIAPED...QVVLLAGAPLEDEAT

UBIQ. hum.  LSDYNIQKESTLHLVLRLRGG*
RAD23       VSECGLKDGDQVVFMVSQKKS-->
HHR23A      IRDYRIDEKNFVVVMVTKTKA-->
HHR23B p58  LKEYKIDEKNFVVVMVTKPKA-->

NEDD8       AADYKILGGSVLHLVLALRGG-->
An1a        LSGYNISEGCTLKMVLAMRGG-->
An1b        LSDYNISEGCTLKMVLAMRGG-->
GdX         LSDYSIGPNSKLNLVVKPLEK-->
BAT3        LQEYNVGG.KVIHLVERAPPQ-->
fau         LGQCGVEALTTLEVAGRMLGG-->
```

Fig 7. Sequence alignment of the yeast and human homologues of RAD23 with each other and with ubiquitin. (**A**) Conserved sequences between yeast RAD23, HHR23A and p58/HHR23B. The amino acid sequence of the human HHR23A and p58/HHR23B proteins are compared with yeast RAD23. (**B**) Alignment of ubiquitin, RAD23, HHR23A, p58/HHR23B and ubiquitin-like sequences. The N-terminal conserved regions of the RAD23, HHR23A, p58/HHR23B and the ubiquitin-like domain in the NEDD8, AN1A, AN1B, GdX, BAT3 and fau proteins are compared with ubiquitin. Sequences used in this figure are NEDD8 (Kumar *et al.*, 1992), AN1A, AN1B (Linnen *et al.*, 1993), GdX (Toniolo *et al.*, 1988), BAT3 (Banerji *et al.*, 1990) and fau (Kas *et al.*, 1992). The amino acid sequence is given in the one letter code. Identical amino acids are presented by black boxes, whereas similar residues (A, S, T, P; D, E, N, Q; R, K; I, L, M, V; F, Y, W) are given in grey boxes.

elution position from a FPLC Mono Q column (Yanagisawa *et al.*, 1992). However, the XP-C correcting protein described here differs from the ATPase Q1 for the following reasons. First, we could not detect any DNA helicase activity in the XP-C correcting protein while the ATPase Q1 has relatively weak but detectable helicase activity. Second, the

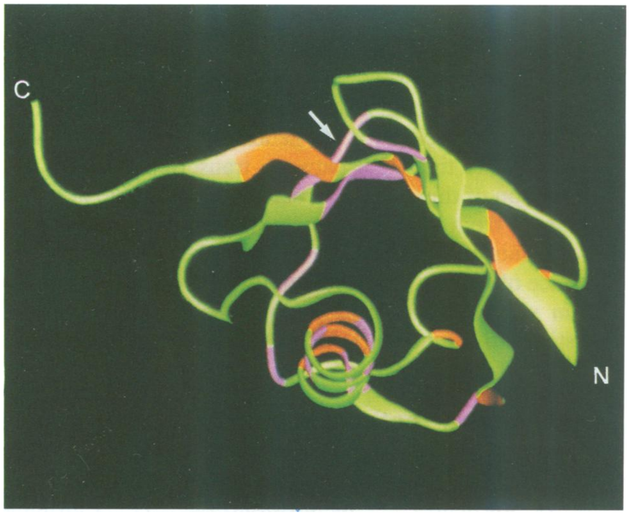

Fig. 8. Selective conservation of HHR23 residues in the 'core' of ubiquitin. The computer drawing shows a model for the tertiary structure of ubiquitin including the presence of one α-helix and four β-sheets. Secondary structure prediction revealed a similar pattern for the N-terminus of RAD23, HHR23A, p58/HHR23B as for ubiquitin (data not shown). The diagram shows in purple the residues of ubiquitin which are identical with those of RAD23, HHR23A and p58/HHR23B as well as with those of many other ubiquitin-like domains: K^{6};P^{19}, T^{22}, K^{27}, K^{29}, L^{43}, I^{44}, G^{47}, K^{48}, L^{50}, and D^{52}. Similar residues (I^{3}, V^{17}, I^{23}, I^{26}, I^{30}, L^{56}, I^{61}, L^{67}, L^{69}, V^{70}) are indicated in orange. It is apparent that intrapolation of these conserved residues into the structure of ubiquitin reveals selective conservation of the core of the protein. Particularly, the inside of the helix seems strongly conserved. The invariant K^{48} is indicated by an arrow.

molecular weights of the two polypeptides purified in this work are different from that of purified ATPase Q1 (73 kDa on SDS−PAGE). Third, a cDNA clone for ATPase Q1 is different from the cDNA clones of p125 and p58. Fourth, purified or partially purified ATPase Q1 cannot complement the repair defects of XP-C cell extracts in our cell-free system (C.Masutani, unpublished observations). Despite the above facts, the alteration of elution of ATPase Q1 from Mono Q column was observed with all five independent XP-C cell lines examined. At present, we do not know why two apparently different proteins, the XP-C correcting protein and the ATPase Q1, are altered in XP-C cells. A possible explanation is a direct or indirect effect of the XPCC protein (complex) on the physical properties of ATPase Q1, e.g. by post-translational protein modification. We are now examining this or other possibilities.

Parallels with yeast

Since the *S.cerevisiae RAD4* gene is likely to be the yeast equivalent of XP-C (Legerski and Peterson, 1992), one inference from our observations is that the yeast RAD23 and RAD4 proteins are likely to interact with each other. Intriguing discrepancies emerge when these parallels are extrapolated to the corresponding mutants and genes. *Rad4* and *rad23Δ* mutants are very different. *RAD4* is one of the seven *RAD* genes that appear to be absolutely required for NER, since *rad4* mutants do not show detectable incisions during incubation after UV exposure (Friedberg, 1988). In contrast, *rad23Δ* mutants exhibit only a partial NER defect, supporting the idea that this gene does not play an essential role in the NER process (Perozzi and Prakash, 1986). Furthermore, both genes differ in their transcriptional response to UV. Transcription of the *RAD23* gene is enhanced upon UV irradiation and during meiosis (Madura and Prakash, 1990) but that of *RAD4* is not (Fleer *et al.*, 1987). Although this damage-induced expression may be similar to the SOS response in bacteria, its functional significance in yeast still needs to be established. Therefore, it will be of interest to examine whether the *RAD23* response is evolutionarily conserved. In view of the likely participation of both yeast proteins in the same complex, it is surprising that the mutant phenotypes are so different. One would assume that absence of one component would render the entire complex non-functional. Indeed we cannot separate the two human partners without inactivating the XP-C correcting activity. One possibility is that—like in man—a second *RAD23*-like gene is hidden in the yeast genome and that this related gene takes over part of the functions of *RAD23*. An alternative, although perhaps not so likely option is that *RAD4* is not the real yeast *XPCC* equivalent. One argument in favour of this idea is the prediction that a true yeast *XPCC* mutant should be specifically defective in the 'genome-overall' NER subpathway. When the relative contribution of this NER subpathway to survival is similar in yeast and man, one would expect a milder phenotype for an XP-C-like yeast mutant than actually revealed by *rad4*. Unfortunately, the degree of homology between the XPCC gene product and the RAD4 protein is not conclusive.

Dual genes for RAD23 in man

Why do two homologues of RAD23 exist in man? All NER genes analysed to date appear to be unique. The only precedent of a repair gene duplication are the human homologues of *RAD6*, *HHR6A* and *HHR6B*, which are implicated in post-replication repair (Koken *et al.*, 1991). Concerning *HHR23A* and *HHR23B*, we have found that both genes are expressed in the same cells. In the XPCC purification scheme, however, only the HHR23B protein is found in a complex with p125/XPCC. It is possible that a second form of this complex involving HHR23A exists that has been missed. Alternatively, the HHR23A component may have dissociated from the complex during purification, or HHR23A is engaged in another complex with the human homologue of RAD4, when this gene is not the XP-C counterpart. Unfortunately, no human mutant defective in HHR23A has been identified so far. Transfection and microinjection experiments of this gene into any of the NER-deficient complementation groups for which no gene has been identified yet failed to induce correction, indicating that a *HHR23A* mutant is not existing in the class of known NER syndromes (P.J.van der Spek, unpublished observations).

Possible function of the XPCC – HHR23B complex

The function of the XPCC complex must be accommodated in a step unique to the genome-overall NER subpathway. The purification procedure indicates that the complex has a high affinity for ssDNA. At present we do not know which of the components (or both) is responsible for this property. Previously putative DNA binding motifs have been postulated for the RAD4 protein (Gietz and Prakash, 1988), however, comparison with the XPCC amino acid sequence reveals that these are not conserved. No obvious DNA binding domains are apparent from the sequence. Also no enzymatic activity was detected for the purified complex (see Results). The only striking domain recognizable using sequence comparison is the ubiquitin-like N-terminus of the RAD23 homologues. Ubiquitin itself is a highly conserved 76 amino acid polypeptide found in all eukaryotes. One or multiple ubiquitin moieties are covalently attached post-translationally to acceptor proteins. This reversible conjugation reaction appears to play an important role in a surprisingly diverse set of regulatory processes, such as selective protein degradation, DNA repair, protein translocation and cell cycle control (reviewed by Jentsch, 1992). Ubiquitin conjugation may also serve as a molecular chaperone.

A number of naturally occurring ubiquitin fusion proteins has been identified. From the alignment shown in Figure 7B, it is apparent that within this functionally diverse family, specific amino acid residues are conserved. Figure 8 shows the position of the conserved amino acids of the ubiquitin-like family, when projected into the known tertiary structure of ubiquitin itself (Vijay-Kumar *et al.*, 1987). It is clear that most residues are clustered in the inner part of the molecule, whereas the periphery appears more prone to divergence. Particularly, the inner half of the α-helix displays a striking conservation. These observations suggest that the core of the molecule is important for the function of this domain. An additional notable feature is the strict conservation of lysine residue K^{48} in all RAD23 derivatives (Figure 8, arrow). This amino acid is involved in multi-ubiquitination since it can serve as point for attachment for ubiquitin conjugation (Jentsch, 1992). The alignment in Figure 7B shows also that the C-terminal glycine doublet is absent in all RAD23 derivatives, suggesting that the ubiquitin moiety can not be cleaved off from the remainder. The function of the ubiquitin(-like) domain in different hybrid proteins is not known. Genetic studies in yeast indicated that the ubiquitin moiety of a ribosomal fusion protein might function as a chaperone, facilitating ribosome assembly (Finley *et al.*, 1989). In analogy with this idea, the ubiquitin-like motif in RAD23 may perform a similar role in assembly of the XPCC–HHR23B complex. If so the intrapolation of Figure 8 suggests that the core rather than the outside of the molecule is important for this function. During the preparation of this manuscript, it was demonstrated that the ubiquitin-like domain is required for RAD23 function in *S.cerevisiae* (Watkins *et al.*, 1993). No other functional clues are yielded up by the primary amino acid sequence of either XPCC or HHR23B.

The identification of the XPCC–HHR23B complex adds to the recent discovery of several multi-protein complexes in mammalian NER. The recently described ERCC1 complex consists of a minimum of three proteins: ERCC1, ERCC4, ERCC11 and XPFC (when this protein is not identical to ERCC4 or ERCC11) (Biggerstaff *et al.*, 1993; van Vuuren *et al.*, 1993). In analogy with the yeast RAD1/RAD10 counterpart, this complex may simultaneously be implicated in a mitotic recombination pathway (Schiestl and Prakash, 1990; Bailly *et al.*, 1992; Bardwell *et al.*, 1992). In addition, the *ERCC3* gene product, responsible for the rare XP complementation group B, was recently uncovered as one of the components of the multisubunit transcription initiation factor BTF2 (TFIIH) (Schaeffer *et al.*, 1993). This finding disclosed an unexpected functional overlap between basal transcription and NER. It is possible that the entire BTF2 transcription complex is involved in NER. The ERCC1 and ERCC3 complexes play a role in both transcription-coupled as well as genome-overall repair and are thus implicated in the core of the NER reaction mechanism (Hoeijmakers, 1993b). The XPCC–HHR23B complex is the first to be described which appears to be specific for the genome-overall subpathway. In view of the tight link between transcription and NER, the function of this complex could be to uncouple the NER machinery from the basal transcription process, enabling it to scan the non-transcribed bulk of the genome for the presence of lesions. The availability of the protein complex and *in vitro* NER systems provide the necessary tools to investigate the function(s) of this NER component.

Materials and methods

Cells and cell culture

Five SV40-transformed fibroblast lines XP2OSSV (group A), XP4PASV (group C), XP6BESV (group D), XP2YOSV (group F) and XP3BRSV (group G), three non-transformed fibroblast lines CRL1199 (group B), XP3KA (group C) and XP2RO (group E), and repair-proficient lines 293 cells were cultured at 37°C in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum. HeLa cells were grown in spinner flasks at 37°C in RPMI 1640 medium supplemented with 5% calf serum and harvested at a density of 10^6 cells/ml.

Preparation of whole cell extracts

The 293 cell line was grown at 37°C in 150 mm tissue culture plates (Falcon), treated with phosphate-buffered saline containing 0.05% Na_3EDTA and collected by gentle pipetting. XP cells were grown in 850 cm^2 roller bottles (Corning) and collected by scraping. The harvested cells were washed with phosphate-buffered saline, and whole cell extracts were prepared as described previously (Manley *et al.*, 1983; Wood *et al.*, 1988). Protein concentration was determined by the method of Bradford (1976) with bovine serum albumin as standard. Extracts contained 10–20 mg of protein/ml.

Preparation of SV40 minichromosomes and plasmid DNA

SV40 virions were prepared as described previously (Sugasawa *et al.*, 1993). Minichromosomes were obtained by alkali disruption of the SV40 virions as described (Christiansen *et al.*, 1977) and irradiated with 200 J/m^2 of UV light (254 nm) as described previously (Sugasawa *et al.*, 1993).

Plasmid pUC19 DNA was propagated in *E.coli* strain HB101. Closed circular DNA was prepared by the alkali lysis method and CsCl–ethidium bromide equilibrium density gradient centrifugation as described (Sambrook *et al.*, 1989). In our previous studies, we used a plasmid DNA sample prepared with a single CsCl centrifugation step (Masutani *et al.*, 1993; Sugasawa *et al.*, 1993). In these previous studies we observed a significant level of DNA synthesis with unirradiated pUC19 DNA. We found that the DNA preparations contained detectable amounts of nicked molecules. These molecules were likely to be used as a template for DNA synthesis, because on repeating the CsCl centrifugation one or two more times, the UV-independent DNA synthesis on pUC19 DNA decreased in proportion to reduction in the amount of nicked molecules. Therefore, in the present study we repeated CsCl centrifugation several times.

Cell-free DNA repair assay

The standard reaction mixture (20 μl) contained 40 mM creatine phosphate–Tris (pH 7.7), 1 mM dithiothreitol, 10 mM $MgCl_2$, 2 mM ATP, 50 μM each of dATP, dGTP and dTTP, 10 μM [α-^{32}P]dCTP (37–74 kBq), phosphocreatine kinase (Sigma, Type I; 0.5 μg), bovine serum albumin (6.4 μg), whole cell extracts (80 μg of protein), unirradiated pUC19 RFI DNA (0.3 μg) and UV-irradiated (200 J/m^2) or unirradiated SV40 mini-chromosomes (0.3 μg). The reaction was performed at 30°C for 3 h. The products were purified from the reaction mixtures, linearized with *Eco*RI and electrophoresed in a 1% agarose gel as described previously (Sugasawa *et al.*, 1993). Autoradiography was performed at −80°C with Fuji New RX X-ray film. The incorporation of radioactive materials into UV-irradiated or unirradiated SV40 minichromosomes was quantified with a Fujix BAS2000 Bio-Imaging Analyzer.

Purification of XP-C correcting protein from HeLa cells

All procedures were carried out at 0–4°C. The purification is summarized in Figure 1 and Table I. A frozen stock of 5×10^{10} HeLa cells (176 ml of packed cell volume) was thawed, washed once with hypotonic buffer

[10 mM Tris–HCl (pH 7.5), 1 mM Na_3EDTA, 2 mM $MgCl_2$, 5 mM dithiothreitol, 0.25 mM PMSF, 0.2 μg/ml aprotinin, 0.2 μg/ml leupeptin, 0.1 μg/ml antipain, and 50 μM EGTA], suspended in 700 ml of hypotonic buffer and homogenized in an all-glass Dounce homogenizer by 15 strokes with a pestle A. The nuclei were obtained by low speed centrifugation, washed twice with nuclei wash buffer [10 mM potassium phosphate (pH 7.5), 1 mM Na_3EDTA, 2 mM dithiothreitol, 0.25 mM PMSF, 0.2 μg/ml aprotinin, 0.2 μg/ml leupeptin, 0.1 μg/ml antipain and 50 μM EGTA] and then suspended in 380 ml of buffer 1 [20 mM potassium phosphate (pH 7.5), 1 mM Na_3EDTA, 5 mM dithiothreitol, 0.25 mM PMSF, 0.2 μg/ml aprotinin, 0.2 μg/ml leupeptin, 0.1 μg/ml antipain and 50 μM EGTA]. A suspension was made in 0.3 M KCl by the addition of 0.1 vol of buffer 1 containing 3.3 M KCl. An extract was obtained by gentle stirring for 30 min followed by centrifugation for 1 h at 100 000 *g*. The supernatant was dialysed against buffer 2 [20 mM potassium phosphate (pH 7.5), 1 mM Na_3EDTA, 10% glycerol, 1 mM dithiothreitol, 0.01% Triton X-100, 0.25 mM PMSF, 0.2 μg/ml aprotinin, 0.2 μg/ml leupeptin, 0.1 μg/ml antipain and 50 μM EGTA] containing 0.15 M KCl and centrifuged for 1 h at 100 000 *g*. The supernatant (nuclear extract) was loaded onto a phosphocellulose column (Whatman P11; 90 ml) equilibrated with buffer 2 containing 0.15 M KCl. The column was washed with three column volumes of the same buffer and the adsorbed proteins were eluted with buffer 2 containing 1 M KCl. The eluate was loaded onto a single-stranded DNA–cellulose column (Sigma; 4.3 mg DNA/g cellulose; 6 ml) equilibrated with buffer 2 containing 0.6 M KCl. The column was washed with three column volumes of the same buffer and the adsorbed proteins were eluted with buffer 2 containing 1.5 M KCl. The eluate was dialysed against buffer 2 containing 0.3 M KCl and adjusted to 0.3 M KCl by dilution with buffer 2. The following two steps were performed with an FPLC system. The dialysate was loaded onto a column of CM cosmogel (Nakalai tesque; 8 mm ID×75 mm) equilibrated with buffer 2 containing 0.3 M KCl. The column was washed with 10 ml of the same buffer and then proteins were eluted with buffer 2 containing 0.6 M KCl. The eluate was adjusted to 0.15 M KCl by diluting with buffer 2 and promptly loaded onto a column of Mono Q HR5/5 (Pharmacia) equilibrated with buffer 2 containing 0.15 M KCl. The column was washed with 10 ml of the same buffer and then proteins were eluted with 25 ml of a linear gradient of 0.15 to 0.45 M KCl in buffer 2. XP-C correcting activity was eluted with ~0.29 M KCl. The active fractions were pooled and stored at −80°C. A portion of the active fraction was dialysed against buffer 1 containing 0.2 M KCl and 50% glycerol, and stored at −20°C. In both pools the XP-C correcting activity was stable for at least 3 months. The XP-C protein could be obtained by another purification procedure in which Tris–HCl (pH 7.5) and NaCl were used instead of potassium phosphate (pH 7.5) and KCl, respectively (data not shown).

XP-C correcting activity was assayed with XP4PASV cell extract in standard conditions. One unit of XP-C correcting activity was defined as the amount of protein required to increase the XP4PASV cell extract-mediated incorporation of 1 pmol of dCMP into UV-irradiated SV40 minichromosomes. As the incorporation of dCMP reached a maximum at 100–150 fmol in standard conditions, units of activity were determined at the order of 10^{-2} by titration.

Gel filtration of XP-C correcting protein

A portion (80 μl) of the Mono Q fraction was loaded onto a Sephacryl S-300 column (6 mm×82 cm) equilibrated with buffer 2 containing 0.3 M KCl and run at 3 ml/h. Fractions (250 μl) were collected and used for assay of XP-C correcting activity and SDS–PAGE. Marker proteins were loaded in identical conditions and detected by SDS–PAGE followed by staining with Coomassie brilliant blue.

Glycerol density gradient centrifugation of the XP-C correcting protein

A portion (60 μl) of the Mono Q fraction was layered on 4.8 ml of a 15–35% (v/v) glycerol gradient in buffer 1 containing 0.3 M KCl and centrifuged in a Hitachi RPS65T rotor at 260 000 *g* for 22 h at 2°C. Fractions (200 μl) were collected from the top of the gradient and assayed for XP-C correcting activity. An identical gradient containing marker proteins was run at the same time. The markers were detected by SDS–PAGE followed by staining with Coomassie brilliant blue.

Assays of enzyme activities

DNA polymerase activity was assayed with activated DNA as template as described previously (Suzuki *et al.*, 1989). The Mono Q fraction of the XP-C correcting factor (60 ng) was incubated at 37°C for 2 h in 30 μl of a solution of 40 mM Tris–HCl (pH 8.0), 1 mM dithiothreitol, 10 mM $MgCl_2$, 2 mM ATP, 50 μM each of dATP, dGTP and dTTP, 10 μM [α-^{32}P]dCTP (74 kBq), 0.32 mg/ml bovine serum albumin and 0.5 mg/ml of activated DNA. The reaction was terminated by chilling on ice and the radioactivity incorporated into acid-insoluble materials was measured.

DNA helicase activity was assayed as oligomer displacing activity. The Mono Q fraction (60 ng) was incubated at 37°C for 1 h in 20 μl of a solution of 50 mM Tris–HCl (pH 7.5), 20 mM 2-mercaptoethanol, 5 mM $MgCl_2$, 5 mM ATP, 0.5 mg/ml bovine serum albumin and 0.017 pmol of 5′ ^{32}P-labelled 21mer annealed to M13 DNA. After termination of the reaction, products were analysed by polyacrylamide (12%) gel electrophoresis followed by autoradiography as described previously (Yanagisawa *et al.*, 1992).

Exonuclease activities were detected in the DNA helicase assay by monitoring the amounts of labelled oligomers and their sizes.

DNA ligase activity was assayed indirectly with bacterial alkaline phosphatase. For this, 60 ng of the Mono Q fraction were incubated at 37°C for 2 h in 30 μl of a solution of 40 mM Tris–HCl (pH 7.5), 1 mM dithiothreitol, 10 mM $MgCl_2$, 2 mM ATP, 0.32 mg/ml bovine serum albumin and 50 ng of 5′ [^{32}P]oligo(dT)$_{12-18}$-poly(dA)$_{400}$ (1:5). Then 0.4 unit of bacterial alkaline phosphatase (Takara) was added and after incubation at 65°C for 1 h, the radioactivity remaining in the acid insoluble material was measured.

Endonuclease activities were measured as nicking activities with UV-irradiated or unirradiated closed circular form I pUC19. The Mono Q fraction (60 ng) was incubated at 37°C for 2 h in 20 μl of solution containing 40 mM Tris–HCl (pH 8.0), 1 mM dithiothreitol, 10 mM $MgCl_2$, 2 mM ATP, 0.32 mg/ml bovine serum albumin and 0.1 μg of UV irradiated (500 J/m^2) or unirradiated closed circular form I pUC19. After the reaction, the plasmids were subjected to 1% agarose gel electrophoresis and detected by ethidium bromide staining.

SDS–PAGE

SDS–PAGE was performed by the method of Laemmli (1970).

Determination of partial amino acid sequences

The Mono Q fractions of the purified XP-C correcting protein were adjusted at 6 M guanidine–HCl and 10 mM sodium phosphate (pH 6.0) and subjected to gel filtration using tandemly joined TSK G3000SW$_{XL}$ and TSK G4000SW$_{XL}$ columns (Tosoh; 7.8×300 mm ea.) and a Gilson HPLC system at a flow rate of 0.5 ml/min. Protein peaks corresponding to the 125 and 58 kDa polypeptides were collected separately and digested with CNBr after removal of salts. The digests were applied to an Aquapore RP300 column (Applied Biosystems; 2.1×100 mm) and eluted with a linear gradient of 0.09% TFA to 80% acetonitrile–0.075% TFA in 40 min at a flow rate of 0.2 ml/min. Materials in clearly isolated peptide peaks were collected and applied to a protein sequencer (Applied Biosystems; model 477A/120A).

Screening of cDNA libraries

For isolation of cDNA clones encoding p125, a cDNA library with relatively long inserts was constructed. Complementary DNAs were synthesized from 5 μg of HeLa cell poly(A)$^+$ RNA using a cDNA synthesis kit (Pharmacia). After addition of *Eco*RI–*Not*I adaptors and size-fractionation by agarose gel electrophoresis, double-stranded cDNAs of >2.5 kb were eluted from the gel and ligated to an *Eco*RI-digested λgt10 vector. Some of the recombinant DNAs were packaged *in vitro* into bacteriophage particles, then amplified in *E.coli* strain, C600 hflA. The resulting cDNA library contained 8.8×10^5 independent clones.

To obtain a probe for screening the cDNA library, RT-PCR was carried out using synthetic oligonucleotide mixtures and first-strand cDNA synthesized from HeLa cell poly(A)$^+$ RNA. The sequences of the oligonucleotides used were 5′-GCI(C/A)GIAA(A/G)(C/A)GIGCIGCIG-GIGGIGA-3′ and 5′-(T/C)TT(T/C)TTIGGIGG(T/C)TT(T/C)TC(A/G)-TC(T/C)TC(A/G)AA-3′, where I indicates inosine. PAGE revealed amplification of 132 bp DNA fragments, which were then purified from the gel and cloned into pUC19 DNA for sequencing. Since the sequence of the 132 bp fragment was consistent with the determined amino acid sequence, this fragment was reamplified from the plasmid, gel-purified and used for screening the cDNA library.

About one million recombinant bacteriophage plaques were transferred to Hybond-N membranes (Amersham) in duplicate. Prehybridization was carried out at 68°C for 4 h in 6×SSC (1×SSC: 0.15 M NaCl, 15 mM sodium citrate), 4×Denhardt's solution (1×Denhardt's solution: 0.02% Ficoll 400, 0.02% bovine serum albumin, 0.02% polyvinylpyrrolidone) and 50 mg/ml heat-denatured salmon sperm DNA. Hybridization was performed at 42°C overnight in 30% formamide, 4×SSC, 4×Denhardt's solution, 50 μg/ml heat-denatured salmon sperm DNA and the DNA probe radiolabelled with [α-^{32}P]dCTP and a multiprime DNA labelling system (Amersham). The membranes were successively washed at room temperature for 10 min and at 55°C for 10 min with 2×blot wash buffer (1×blot wash buffer: 1×SSC, 10 mM sodium phosphate, 0.025% SDS), at 55°C for 10 min with 1×blot wash buffer, at 55°C for 30 min with 0.5×blot wash buffer, at 55°C for 30 min with 0.2×blot wash buffer and twice at 65°C for 30 min with 0.1×blot wash buffer. Then the membranes were air-dried and exposed at −80°C to Kodak X-OMAT film with intensifying screens. A positive plaque was picked up and purified by another round of plaque hybridization.

The 3.6 kb insert of the positive clone was obtained by *Not*I digestion and subcloned into the *Not*I site of pBluescript II KS$^+$. Deletion mutants were constructed by use of exonuclease III and mung bean nuclease (a deletion kit for kilo-sequencing; Takara Shuzo), and sequenced with a Taq Dye Deoxy Terminator cycle sequencing kit and an automated DNA sequencer (Applied Biosystems, model 373A).

For isolation of cDNA clones encoding the p58, an oligonucleotide, 5′-CCICCICCICC(C/T)TGICCICCIGC(C/T)TC(C/T)TGIACIGG(C/T)TC(A/G)TT-3′, was used for screening a λgt10 cDNA library from HeLa cells. Screening was performed as described above. The 2.9 kb insert of the positive clone was obtained by *Eco*RI digestion and subcloned into the *Eco*RI site of pUC19. Deletion mutants were constructed and sequenced as described above.

Cloning and nucleotide sequence analysis of HHR23A

Total RNA (10 μg) was used for preparing cDNA with *HHR23A*-specific primers (see below). RNA was dissolved in 9 μl of annealing buffer [250 mM KCl, 10 mM Tris−HCl (pH 8.3), 1 mM EDTA]. Following the addition of 1 μl (100 pmol/μl) of primers, the samples were first heated for 3 min at 80°C and transferred to a 37°C water bath for 1 h. Fifteen microlitres of cDNA buffer (24 mM Tris−HCl [pH 8.3], 16 mM $MgCl_2$, 8 mM DTT, 0.4 mM of dGTP, dATP, dTTP, dCTP) and 5 U of Moloney leukaemia virus reverse transcriptase (Promega) were added and the tube was incubated at 37°C for 1 h. To 5 μl cDNA, 10 μl of Taq buffer [100 mM Tris−HCl (pH 8.3), 15 mM $MgCl_2$, 500 mM KCl, 2 mg/ml bovine serum albumin], 4 μl dNTPs (2.5 mM), 75 μl water, 1 μl of each primer (100 pmol/μl) and 2 U of Taq polymerase (Cetus) were added.

Oligonucleotide primers for cDNA, DNA amplification and DNA sequencing were synthesized in an Applied Biosystems DNA synthesizer. The PCR primers used for this purpose are: 5′-ATCCAGATGC-TGAACGAGCC-3′ and 5′-CGGCAGGTGATTCAGCAGAAC-3′.

A PCR probe was used to screen a pre-B cell library and clones hybridizing with the PCR probe were picked up and examined by restriction enzyme analysis. Hybridization of human probes to human DNA was at 65°C in a hybridization mixture containing 10×Denhardt's solution, 10% dextran sulfate, 0.1% SDS, 3×SSC, 50 mg of sonicated salmon sperm DNA per litre. Washings were performed twice for 20 min each in 3×SSC, twice for 20 min each in 1×SSC and twice for 20 min each in 0.3×SSC at 65°C. Hybridization was detected by autoradiography on Fuji medical X-ray film RX with intensifying screens at −80°C.

Lambda zap phages (Short *et al.*, 1988) were after two rounds of rescreens converted into Bluescript vectors and transformed to competent DH5αF′ cells. Sequence analysis on double-stranded DNA was done by the T7 DNA polymerase modification (Pharmacia) of the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) using sequence-derived oligonucleotides prepared for sequencing both strands. For separation of the fragments, Hydrolink (AT Biochem, Malvern, PA) sequencing gels were used.

Acknowledgements

We are grateful to Drs M.Sekiguchi, A.Sarai, Christine Troelstra and Marcel Koken for their helpful suggestions, Sigrid Swagemakers for help with some of the experiments and to Satya Prakash (Galveston) who kindly provided us with the yeast *RAD23* gene sequence. Furthermore, we thank Dr Jack Leunissen from the Nijmegen University CAOS/CAMM center for his assistance with the molecular modelling packages. This work was supported by grants from the Ministry of Education, Science and Culture of Japan, the Biodesign Research Program of the Institute of Physical and Chemical Research (RIKEN), and the Cosmetology Research Foundation of Japan. The work at the Department of Cell Biology of the Erasmus University was financially supported by the Medical Genetics Centre South-West Netherlands.

References

Adams,M.D., Dubnick,M., Kerlavage,A.R., Moreno,R., Kelly,J.M., Utterback,T.R., Nagle,J.W., Fields,C. and Venter,J.C. (1992) *Nature*, **355**, 632−634.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403−410.

Bailly,V., Sommers,C.H., Sung,P., Prakash,L. and Prakash,S. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 8273−8277.

Banerji,J., Sands,J., Strominger,J.L. and Spies,T. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 2374−2378.

Bardwell,L., Cooper,A.J. and Friedberg,E.C. (1992) *Mol. Cell. Biol.*, **12**, 3041−3049.

Biggerstaff,M., Szymkowski,D.E. and Wood,R.D. (1993) *EMBO J.*, **12**, 3685−3692.

Bradford,M.M. (1976) *Anal. Biochem.*, **72**, 248−254.

Christiansen,G., Landers,T., Griffith,J. and Berg,P. (1977) *J. Virol.*, **21**, 1079−1084.

Cleaver,J.E. (1968) *Nature*, **218**, 652−656.

Cleaver,J.E. and Kraemer,K.H. (1989) In Scriver,C.R., Beaudet,A.L., Sly,W.A. and Valle,D. (eds), *Xeroderma Pigmentosum. The Metabolic Basis of Inherited Disease.* Vol. II, McGraw-Hill Book Co., New York, pp. 2949−2971.

Collins,A.R. (1993) *Mutat. Res.*, **293**, 99−118.

Finley,D., Bartel,B. and Varshavsky,A. (1989) *Nature*, **338**, 394−401.

Fleer,R., Nicolet,C.M., Pure,G.A. and Friedberg,E.C. (1987) *Mol. Cell. Biol.*, **7**, 1180−1192.

Flejter,W.L., McDaniel,L.D., Johns,D., Friedberg,E.C. and Schultz,R.A. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 261−265.

Friedberg,E.C. (1985) *DNA Repair.* W.H.Freeman and Co., New York.

Friedberg,E.C. (1988) *Microbiol. Rev.*, **52**, 70−102.

Gietz,R.D. and Prakash,S. (1988) *Gene*, **74**, 535−541.

Hanawalt,P. and Mellon,I. (1993) *Curr. Biol.*, **3**, 67−69.

Hoeijmakers,J.H.J. (1993a) *Trends Genet.*, **9**, 173−177.

Hoeijmakers,J.H.J. (1993b) *Trends Genet.*, **9**, 211−217.

Huang,J.-C., Svoboda,D.L., Reardon,J.T. and Sancar,A. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 3664−3668.

Jentsch,S. (1992) *Annu. Rev. Genet.*, **26**, 179−207.

Kantor,G.J., Barsalou,L.S. and Hanawalt,P.C. (1990) *Mutat. Res.*, **235**, 171−180.

Kas,K., Michiels,L. and Merregaert,J. (1992) *Biochem. Biophys. Res. Commun.*, **187**, 927−933.

Koken,M.H.M., Reynolds,P., Jaspers-Dekker,I., Prakash,L., Prakash,S., Bootsma,D. and Hoeijmakers,J.H.J. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 8865−8869.

Kozak,M. (1991) *J. Biol. Chem.*, **266**, 19867−19870.

Kraemer,K.H., Lee,M.M. and Scotto,J. (1987) *Arch. Dermatol.*, **123**, 241−250.

Kumar,S., Tomooka,Y. and Noda,M. (1992) *Biochem. Biophys. Res. Commun.*, **185**, 1155−1161.

Laemmli,U.K. (1970) *Nature*, **227**, 680−685.

Legerski,R. and Peterson,C. (1992) *Nature*, **359**, 70−73.

Li,L., Bales,E.S., Peterson,C.A. and Legerski,R.J. (1993) *Nature Genet.*, **5**, 413−417.

Linnen,J.M., Bailey,C.P. and Weeks,D.L. (1993) *Gene*, **128**, 181−188.

Madura,K. and Prakash,S. (1990) *Nucleic Acids Res.*, **18**, 4737−4742.

Manley,J.L., Fire,A., Samuels,M. and Sharp,P.A. (1983) *Methods Enzymol.*, **101**, 568−582.

Masutani,C., Sugasawa,K., Asahina,H., Tanaka,K. and Hanaoka,F. (1993) *J. Biol. Chem.*, **268**, 9105−9109.

Melnick,L. and Sherman,F. (1993) *J. Mol. Biol.*, **233**, 372−388.

O'Donovan,A. and Wood,R.D. (1993) *Nature*, **363**, 185−188.

Okubo,K., Hori,N., Matoba,R., Niiyama,T., Fukushima,A., Kojima,Y. and Matsubara,K. (1992) *Nature Genet.*, **2**, 173−179.

Perozzi,G. and Prakash,S. (1986) *Mol. Cell. Biol.*, **6**, 1497−1507.

Riboni,R., Botta,E., Stefanini,M., Numata,M. and Yasui,A. (1992) *Cancer Res.*, **52**, 6690−6691.

Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning, A Laboratory Manual.* 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sancar,A. and Hearst,J.E. (1993) *Science*, **259**, 1415−1420.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl Acad. Sci. USA*, **74**, 5463−5467.

Schaeffer,L., Roy,R., Humbert,S., Moncollin,V., Vermeulen,W.,

Hoeijmakers,J.H.J., Chambon,P. and Egly,J.-M. (1993) *Science*, **260**, 58–63.
Schiestl,R.H. and Prakash,S. (1990) *Mol. Cell. Biol.*, **10**, 2485–2491.
Short,J.M., Fernandez,M., Sorge,J.A. and Huse,W.D. (1988) *Nucleic Acids Res.*, **16**, 7583–7600.
Sibghat-Ullah, Husain,I., Carlton,W. and Sancar,A. (1989) *Nucleic Acids Res.*, **17**, 4471–4484.
Sugasawa,K., Masutani,C. and Hanaoka,F. (1993) *J. Biol. Chem.*, **268**, 9098–9104.
Suzuki,M., Enomoto,T., Masutani,C., Hanaoka,F., Yamada,M. and Ui,M. (1989) *J. Biol. Chem.*, **264**, 10065–10071.
Tanaka,K., Miura,N., Satokata,I., Miyamoto,I., Yoshida,M.C., Satoh,Y., Kondo,S., Yasui,A., Okayama,H. and Okada,Y. (1990) *Nature*, **348**, 73–76.
Toniolo,D., Persico,M. and Alcalay,M. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 851–855.
Troelstra,C., van Gool,A., de Wit,J., Vermeulen,W., Bootsma,D. and Hoeijmakers,J.H.J. (1992) *Cell*, **71**, 939–953.
van Duin,M., Vredeveldt,G., Mayne,L.V., Odijk,H., Vermeulen,W., Klein,B., Weeda,G., Hoeijmakers,J.H.J., Bootsma,D. and Westerveld,A. (1989) *Mutat. Res.*, **217**, 83–92.
Van Houten,B. (1990) *Microbiol. Rev.*, **54**, 18–51.
van Vuuren,A.J., Appeldoorn,E., Odijk,H., Yasui,A., Jaspers,N.G.J., Bootsma,D. and Hoeijmakers,J.H.J. (1993) *EMBO J.*, **12**, 3693–3701.
Venema,J., van Hoffen,A., Natarajan,A.T., van Zeeland,A.A. and Mullenders,L.H.F. (1990) *Nucleic Acids Res.*, **18**, 443–448.
Venema,J., van Hoffen,A., Karcagi,V., Natarajan,A.T., van Zeeland,A.A. and Mullenders,L.H.F. (1991) *Mol. Cell. Biol.*, **11**, 4128–4134.
Vijay-Kumar,S., Bugg,C.E., Wilkinson,K.D., Vierstra,R.D., Hatfield,P.M. and Cook,W.J. (1987) *J. Biol. Chem.*, **262**, 6396–6399.
Watkins,J.F., Sung,P., Prakash,L. and Prakash,S. (1993) *Mol. Cell. Biol.*, **13**, 7757–7765.
Weeda,G., van Ham,R.C.A., Vermeulen,W., Bootsma,D., van der Eb,A.J. and Hoeijmakers,J.H.J. (1990) *Cell*, **62**, 777–791.
Wood,R.D., Robins,P. and Lindahl,T. (1988) *Cell*, **53**, 97–106.
Yanagisawa,J., Seki,M., Ui,M. and Enomoto,T. (1992) *J. Biol. Chem.*, **267**, 3585–3588.

Received on January 4, 1994; revised on February 9, 1994