likely GTPases, as indicated by the activity of CIITA and HET-E [E. V. Koonin, L. Aravind, *Trends Biochem. Sci.* **25**, 223 (2000)].
14. T. L. Beattie, W. Zhou, M. O. Robinson, L. Harrington, *Curr. Biol.* **8**, 177 (1998).
15. E. Diez, Z. Yaraghi, A. MacKenzie, P. Gros, *J. Immunol.* **164**, 1470 (2000).
16. A. M. Verhagen *et al.*, *Cell* **102**, 43 (2000).
17. L. Goyal, K. McCall, J. Agapite, E. Hartwieg, H. Steller, *EMBO J.* **19**, 589 (2000).
18. The eukaryotic crown group is the assemblage of relatively late-diverging, major eukaryotic taxa whose exact order of radiation is difficult to determine with confidence. The crown group includes the multicellular eukaryotes (animals, fungi, and plants) and some unicellular eukaryotic lineages such as slime molds and Acanthamoebae [A. H. Knoll, *Science* **256**, 622 (1992); S. Kumar, A. Rzhetsky, *J. Mol. Evol.* **42**, 183 (1996)].
19. The sister group of the classic animal caspase family of thiol proteases are the paracaspases that thus far have been identified only in animals and *Dictyostelium*; together, these two families constitute the sister group of the metacaspases that have been detected in plants, protists, and bacteria [A. G. Uren *et al.*, *Mol. Cell* **6**, 961 (2000)]. On the basis of conserved structural features, Uren *et al.* showed that the paracaspases and metacaspases are specifically related to the caspases, to the exclusion of other members of the caspase-gingipain fold [A. Eichinger *et al.*, *EMBO J.* **18**, 5453 (1999)].
20. The A20 protein is a regulator of apoptosis that appears to be involved in the NFκB pathway and interactions with the TRAFs [R. Beyaert, K. Heyninck, S. Van Huffel, *Biochem. Pharmacol.* **60**, 1143 (2000)]. A20 belongs to a distinct family of predicted thiol proteases that is conserved in all crown-group eukaryotes and many viruses. None of the members of this family has a known biochemical function, but they share two conserved motifs with the cysteine proteases of arteriviruses, which led to the prediction of the protease activity. A20 and another protein of this family, cezanne, contain a specialized finger module that is also found in some proteins of the ubiquitin pathway. Together with a fusion of an A20-like protease domain with a ubiquitin hydrolase that has been detected in *C. elegans*, this suggests a functional connection between these predicted proteases and the ubiquitin system [K. S. Makarova, L. Aravind, E. V. Koonin, *Trends Biochem. Sci.* **25**, 50 (2000)]. An additional connection between apoptosis and the ubiquitin system is indicated by the demonstration that, similar to other RING fingers, the one in TRAF6 is an E3-like ubiquitin ligase pathway [L. Deng *et al.*, *Cell* **103**, 351 (2000)].
21. The AP-GTPase is a previously undetected predicted GTPase typified by the COOH-terminal domain of the conserved apoptosis regulator, the DAP protein kinase [B. Inbal *et al.*, *Nature* **390**, 180 (1997)]. This predicted GTPase family appears to be the sister group of the RAS/ARF family GTPases, but differs from them in having a divergent P-loop motif and a THXD instead of the NKXD signature motif. Additional AP-GTPases are found in plants and animals as multidomain proteins that also contain ankyrin, Lrr, and kinase domains. This domain architecture suggests that AP-GTPases participate in GTP-dependent assembly of signaling complexes.
22. A. G. Uren *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10170 (1999).
23. The ZU5 domain is a previously undetected conserved domain that is present in receptors (such as netrin receptors and vertebrate zona pellucida proteins) and cytoskeletal proteins (such as ankyrins) and is predicted to be involved in anchoring receptors to the cytoskeleton.
24. S. L. Ackerman, B. B. Knowles, *Genomics* **52**, 205 (1998).
25. H. Sakahira, M. Enari, S. Nagata, *Nature* **391**, 96 (1998).
26. A. M. Aguinaldo *et al.*, *Nature* **387**, 489 (1997).
27. L. Aravind, H. Watanabe, D. J. Lipman, E. V. Koonin, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11319 (2000).
28. J. R. Brown, W. F. Doolittle, *Microbiol. Mol. Biol. Rev.* **61**, 456 (1997).
29. We thank E. Birney and A. Bateman (The Sanger Center, Hinxton, UK) for kindly providing the preliminary version of the Integrated Protein Index and A. Uren for critical reading of the manuscript and useful comments. The release of the unpublished WormPep data set by The Sanger Center is acknowledged and greatly appreciated.

25 October 2000; accepted 18 January 2001

Human DNA Repair Genes

Richard D. Wood,[1]* Michael Mitchell,[2] John Sgouros,[2] Tomas Lindahl[1]

Cellular DNA is subjected to continual attack, both by reactive species inside cells and by environmental agents. Toxic and mutagenic consequences are minimized by distinct pathways of repair, and 130 known human DNA repair genes are described here. Notable features presently include four enzymes that can remove uracil from DNA, seven recombination genes related to RAD51, and many recently discovered DNA polymerases that bypass damage, but only one system to remove the main DNA lesions induced by ultraviolet light. More human DNA repair genes will be found by comparison with model organisms and as common folds in three-dimensional protein structures are determined. Modulation of DNA repair should lead to clinical applications including improvement of radiotherapy and treatment with anticancer drugs and an advanced understanding of the cellular aging process.

The human genome, like other genomes, encodes information to protect its own integrity (*1*). DNA repair enzymes continuously monitor chromosomes to correct damaged nucleotide residues generated by exposure to carcinogens and cytotoxic compounds. The damage is partly a consequence of environmental agents such as ultraviolet (UV) light from the sun, inhaled cigarette smoke, or incompletely defined dietary factors. However, a large proportion of DNA alterations are caused unavoidably by endogenous weak mutagens including water, reactive oxygen species, and metabolites that can act as alkylating agents. Very slow turnover of DNA consequently occurs even in cells that do not proliferate. Genome instability caused by the great variety of DNA-damaging agents would be an overwhelming problem for cells and organisms if it were not for DNA repair.

On the basis of searches of the current draft of the human genome sequence (*2*), we compiled a comprehensive list of DNA repair genes (Table 1). This inventory focuses on genes whose products have been functionally linked to the recognition and repair of damaged DNA as well as those showing strong sequence homology to repair genes in other organisms. Readers desiring further information on specific genes should consult the primary references and links available through the accession numbers. Recent review articles on the evolutionary relationships of DNA repair genes (*3*) and common sequence motifs in DNA repair genes (*4*) may also be helpful.

The functions required for the three distinct forms of excision repair are described separately. These are base excision repair (BER), nucleotide excision repair (NER), and mismatch repair (MMR). Additional sections discuss direct reversal of DNA damage, recombination and rejoining pathways for repair of DNA strand breaks, and DNA polymerases that can bypass DNA damage.

The BER proteins excise and replace damaged DNA bases, mainly those arising from endogenous oxidative and hydrolytic decay of DNA (*1*). DNA glycosylases initiate this process by releasing the modified base. This is followed by cleavage of the sugar-phosphate chain, excision of the abasic residue, and local DNA synthesis and ligation. Cell nuclei and mitochondria contain several related but nonidentical DNA glycosylases obtained through alternative splicing of transcripts. Three different nuclear DNA glycosylases counteract oxidative damage, and a fourth mainly excises alkylated purines. Remarkably, four of the eight identified DNA glycosylases can remove uracil from DNA. Each of them has a specialized function, however. UNG, which is homologous to the *Escherichia coli* Ung enzyme, is associated with DNA replication forks and corrects uracil misincorporated opposite adenine. SMUG1, which is unique to higher eukaryotes, probably removes the uracil that arises in DNA by deamination of cytosine. MBD4 excises uracil and thymine specifical-

[1]Imperial Cancer Research Fund, Clare Hall Laboratories, Blanche Lane, South Mimms, Herts EN6 3LD, UK. [2]Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK.

*Present address: University of Pittsburgh Cancer Institute, S867 Scaife Hall, 3550 Terrace Street, Pittsburgh, PA 15261, USA.

ly at deaminated CpG and 5-methyl-CpG sequences, and TDG removes ethenoC, a product of lipid peroxidation, and also slowly removes uracil and thymine at G·U and G·T base pairs. The existence of multiple proteins with similar activities is a recurring theme in human DNA repair (*1*). Another illustration of this is the set of at least four adenosine triphosphate (ATP)–dependent DNA ligases encoded by three genes, with LIG3-XRCC1 providing the main nick-joining function for BER.

Until recently, only one endonuclease for abasic sites had been found encoded in the human genome, although there are two each in *E. coli* and the yeast *Saccharomyces cerevisiae* and three genes are predicted in the genome of the plant *Arabidopsis thaliana*. A second human gene, *APE2*, has recently appeared. Apparently this encodes a minor activity, as deletion of the major gene *APE1* causes early embryonic lethality in mice. Repair of the DNA replication–blocking lesion 3-methyladenine is another case where the human genome is frugal. In other organisms, several DNA glycosylases, unrelated at the primary sequence level, can remove 3-meA. Among them are Tag1 of *E. coli*, AlkA of *E. coli* (similar to MAG of *S. cerevisiae*), and MPG in higher eukaryotes. Only the MPG enzyme has been characterized so far in the human genome. This is in contrast to the at least two *alkA* and six *tag1* homologs found in *Arabidopsis* (*5*). However, like the genomes of other multicellular animals, the current human genome draft contains no obvious *tag1* and *alkA* homologs (*6*).

A few unusual enzymes reverse rather than excise DNA damage. The human MGMT removes methyl groups and other small alkyl groups from the O^6 position of guanine. There are two such proteins (Ada and Ogt) in *E. coli*, but no additional homologs have been detected in the human genome sequence. MGMT resembles the COOH-terminal half of Ada. The NH_2-terminal half of *E. coli* Ada can remove a methyl group from a DNA phosphate residue. We found no homologs of this region of Ada, and it remains unclear whether such backbone methylations are repaired in human cells.

Many organisms contain photolyases that can monomerize lesions induced by UV light such as cyclobutane pyrimidine dimers and (6-4) photoproducts. The human genome has two *CRY* genes with similarity to photolyase sequences. These encode blue light photoreceptors involved in setting of circadian rhythms but not in photoreactivation of DNA damage. We have not detected additional homologs of DNA repair photolyases in the human genome, confirming previous reports that photolyase activity is present in many vertebrates including fish, reptiles, and marsupials, but not in placental mammals.

NER mainly removes bulky adducts caused by environmental agents. In *E. coli*, the three polypeptides UvrA, UvrB, and UvrC can locate a lesion and incise on either side of it to remove a segment of nucleotides containing the damage. Eukaryotes, including yeast and human cells, do not have direct UvrABC homologs but use a more elaborate assembly of gene products to carry out NER (*1*). For example, *E. coli* UvrA can bind to sites of DNA damage, whereas at least four different human NER factors have this property (the XPC complex, DDB complex, XPA, and RPA). The formation of an unwound preincision intermediate in human cells requires two DNA helicases, XPB and XPD, instead of the single UvrB in *E. coli*, and there are dedicated human nucleases (XPG and ERCC1-XPF) for each of the two incisions, instead of the single UvrC in bacteria. *S. cerevisiae* encodes two additional gene products, Rad7 and Rad16, which are important for NER. No convincing homologs to these can be identified in the human genome, although Rad16 is a difficult case because it is a member of the amply represented Swi/Snf family of DNA-stimulated adenosine triphosphatases (ATPases).

Some organisms such as the fission yeast *Schizosaccharomyces pombe* have a second system for excision of pyrimidine dimers, initiated by a UVDE nuclease. The human genome apparently lacks a homolog of this nuclease and has no such backup system, consistent with the fact that cells from NER-defective xeroderma pigmentosum patients totally lack the ability to remove pyrimidine dimers from DNA.

The transcribed strand of active human genes is repaired faster than the nontranscribed strand in a transcription-coupled repair process known to involve the products of *CSA*, *CSB*, and *XAB2*. The mechanism of such transcription-coupled repair is not known, and future investigation is expected to reveal additional participants.

MMR corrects occasional errors of DNA replication as well as heterologies formed during recombination. The bacterial *mutS* and *mutL* genes encode proteins responsible for identifying mismatches, and there are numerous homologs of these genes in the human genome, of greater variety than those found in yeast, *Drosophila melanogaster*, or *Caenorhabditis elegans*. Some of these proteins are specialized for locating distinct types of mismatches in DNA, some are specialized for meiotic recombination, and some have functions yet to be determined. In *E. coli*, the newly synthesized DNA strand is identified with the aid of the MutH endonuclease, which has no human ortholog. Strand discrimination in human cells may be signaled instead by the orientation of components of the DNA replication complex such as PCNA or by other factors not yet identified.

DNA double-strand breaks may be rectified by either homologous or nonhomologous recombination pathways. Particularly notable in the human sequence is the presence of at least seven genes encoding proteins distantly related to the single Rad51 of *S. cerevisiae* and the single RecA of *E. coli*. The latter proteins function in strand pairing and exchange during recombination. By comparison, four members of the Rad51 family have been found in the *Drosophila* genome (*7*) and four in *Arabidopsis* (*5*). Homologous recombination in human cells is likely to involve branch migration enzymes and resolvases that are functionally analogous to the bacterial RuvABC system. Recent biochemical experiments have revealed human activities for such concerted branch migration/resolution reactions, but the responsible gene products have not yet been identified (*8*).

The nonhomologous end-joining pathway (NHEJ) involves the factors listed in Table 1, and additional components will most likely be discovered. For example, the DNA-dependent protein kinase is believed to phosphorylate key molecules involved in the repair process. These substrates have yet to be fully defined.

Single-strand interruptions in DNA can be rectified by enzymes from the BER pathway. Enzymes of the PARP family, as well as XRCC1, temporarily bind to single-strand interruptions in DNA and may act to recruit repair proteins. We have not listed the telomere-binding proteins protecting the ends of chromosomes, but one member of the PARP family, tankyrase, is present in this complex.

During the past year, the human genome sequence has revealed many previously unrecognized DNA polymerases (*1*). There are currently at least 15 DNA polymerases in humans, exceeding the number found in any other organism. For repair of nuclear DNA, the main form of BER uses Pol β, whereas Pol δ or Pol ε are the main enzymes employed for NER and MMR. Genetic and biochemical evidence has implicated many of the newly discovered polymerases in the DNA damage response, but others may have specialized roles such as sister chromatid cohesion. Table 1 includes the catalytic subunits of these DNA polymerases, but not other subunits and DNA polymerase cofactors.

REV3L, the catalytic subunit of DNA polymerase ζ, illustrates how DNA sequence homology searches can yield unexpected results. The DNA polymerase domain at the COOH-terminus of the human protein resembles *S. cerevisiae* Rev3, but most of the first 2000 amino acids are not present in the yeast protein. A second human gene highly homologous to 1200 residues in this region (outside the polymerase domains) is encoded on the X chromosome (accession number AL139395). It is premature to classify this as a DNA repair gene, but study of it is expected to shed light on the function of REV3L.

Table 1. Human DNA repair genes. A version of this table with active links to Gene Cards (bioinformatics.weizmann.ac.il/cards) and to the National Center for Biotechnology Information is available (*24*) on *Science* Online. A version with updates is available at www.cgal.icnet.uk/DNA_Repair_Genes.html. XP, xeroderma pigmentosum.

Gene name (synonyms)	Activity	Chromosome location	Accession number
	Base excision repair (BER)		
	DNA glycosylases: major altered base released		
UNG	U	12q23-q24.1	NM_003362
SMUG1	U	12q13.1-q14	NM_014311
MBD4	U or T opposite G at CpG sequences	3q21-q22	NM_003925
TDG	U, T or ethenoC opposite G	12q24.1	NM_003211
OGG1	8-oxoG opposite C	3p26.2	NM_002542
MYH	A opposite 8-oxoG	1p34.3-p32.1	NM_012222
NTH1	Ring-saturated or fragmented pyrimidines	16p13.3-p13.2	NM_002528
MPG	3-meA, ethenoA, hypoxanthine	16p13.3	NM_002434
	Other BER factors		
APE1 (HAP1, APEX, REF1)	AP endonuclease	14q12	NM_001641
APE2 (APEXL2)	AP endonuclease	X	NM_014481
LIG3	Main ligation function	17q11.2-q12	NM_013975
XRCC1	Main ligation function	19q13.2	NM_006297
	Poly(ADP-ribose) polymerase (PARP) enzymes		
ADPRT	Protects strand interruptions	1q42	NM_001618
ADPRTL2	PARP-like enzyme	14q11.2-q12	NM_005485
ADPRTL3	PARP-like enzyme	3p21.1-p22.2	AF085734
	Direct reversal of damage		
MGMT	O^6-meG alkyltransferase	10q26	NM_002412
	Mismatch excision repair (MMR)		
MSH2	Mismatch and loop recognition	2p22-p21	NM_000251
MSH3	Mismatch and loop recognition	5q11-q12	NM_002439
MSH6	Mismatch recognition	2p16	NM_000179
MSH4	MutS homolog specialized for meiosis	1p31	NM_002440
MSH5	MutS homolog specialized for meiosis	6p21.3	NM_002441
PMS1	Mitochondrial MutL homolog	2q31.1	NM_000534
MLH1	MutL homolog	3p21.3	NM_000249
PMS2	MutL homolog	7p22	NM_000535
MLH3	MutL homolog of unknown function	14q24.3	NM_014381
PMS2L3	MutL homolog of unknown function	7q11-q22	D38437
PMS2L4	MutL homolog of unknown function	7q11-q22	D38438
	Nucleotide excision repair (NER)		
XPC	Binds damaged DNA as complex	3p25	NM_004628
RAD23B (HR23B)	Binds damaged DNA as complex	3p25.1	NM_002874
CETN2	Binds damaged DNA as complex	Xq28	NM_004344
RAD23A (HR23A)	Substitutes for HR23B	19p13.2	NM_005053
XPA	Binds damaged DNA in preincision complex	9q22.3	NM_000380
RPA1	Binds DNA in preincision complex	17p13.3	NM_ 002945
RPA2	Binds DNA in preincision complex	1p35	NM_002946
RPA3	Binds DNA in preincision complex	7p22	NM_002947
TFIIH	Catalyzes unwinding in preincision complex		
XPB (ERCC3)	3′ to 5′ DNA helicase	2q21	NM_000122
XPD (ERCC2)	5′ to 3′ DNA helicase	19q13.2-q13.3	X52221
GTF2H1	Core TFIIH subunit p62	11p15.1-p14	NM_005316
GTF2H2	Core TFIIH subunit p44	5q12.2-q13.3	NM_001515
GTF2H3	Core TFIIH subunit p34	12q	NM_001516
GTF2H4	Core TFIIH subunit p52	6p21.3	NM_001517
CDK7	Kinase subunit of TFIIH	2p15-cen	NM_001799
CCNH	Kinase subunit of TFIIH	5q13.3-q14	NM_001239
MNAT1	Kinase subunit of TFIIH	14q23	NM_002431
XPG (ERCC5)	3′ incision	13q33	NM_000123
ERCC1	5′ incision subunit	19q13.2-q13.3	NM_001983
XPF (ERCC4)	5′ incision subunit	16p13.3-p13.13	NM_005236
LIG1	DNA joining	19q13.2-q13.3	NM_000234
NER-related			
CSA (CKN1)	Cockayne syndrome; needed for transcription-coupled NER	5q12-q31	NM_000082
CSB (ERCC6)	Cockayne syndrome; needed for transcription-coupled NER	10q11	NM_000124
XAB2 (HCNP)	Cockayne syndrome; needed for transcription-coupled NER	19	NM_020196
DDB1	Complex defective in XP group E	11q12-q13	NM_001923
DDB2	Mutated in XP group E	11p12-p11	NM_000107
MMS19	Transcription and NER	10q24.1	AW852889
	Homologous recombination		
RAD51	Homologous pairing	15q15.1	NM_002875
RAD51L1 (RAD51B)	Rad51 homolog	14q23-q24	U84138
RAD51C	Rad51 homolog	17q11-qter	NM_002876
RAD51L3 (RAD51D)	Rad51 homolog	17q11	NM_002878

Table 1. Continued.

Gene name (synonyms)	Activity	Chromosome location	Accession number
DMC1	Rad51 homolog, meiosis	22q13.1	NM_007068
XRCC2	DNA break and cross-link repair	7q36.1	NM_005431
XRCC3	DNA break and cross-link repair	14q32.3	NM_005432
RAD52	Accessory factor for recombination	12p13-p12.2	NM_002879
RAD54L	Accessory factor for recombination	1p32	NM_003579
RAD54B	Accessory factor for recombination	8q21.3-q22	NM_012415
BRCA1	Accessory factor for transcription and recombination	17q21	NM_007295
BRCA2	Cooperation with RAD51, essential function	13q12.3	NM_000059
RAD50	ATPase in complex with MRE11A, NBS1	5q31	NM_005732
MRE11A	3′ exonuclease	11q21	NM_005590
NBS1	Mutated in Nijmegen breakage syndrome	8q21-q24	NM_002485
	Nonhomologous end-joining		
Ku70 (G22P1)	DNA end binding	22q13.2-q13.31	NM_001469
Ku80 (XRCC5)	DNA end binding	2q35	M30938
PRKDC	DNA-dependent protein kinase catalytic subunit	8q11	NM_006904
LIG4	Nonhomologous end-joining	13q33-q34	NM_002312
XRCC4	Nonhomologous end-joining	5q13-q14	NM_003401
	Sanitization of nucleotide pools		
MTH1 (NUDT1)	8-oxoGTPase	7p22	NM_002452
DUT	dUTPase	15q15-q21.1	NM_001948
	DNA polymerases (catalytic subunits)		
POLB	BER in nuclear DNA	8p11.2	NM_002690
POLG	BER in mitochondrial DNA	15q25	NM_002693
POLD1	NER and MMR	19q13.3	NM_002691
POLE1	NER and MMR	12q24.3	NM_006231
PCNA	Sliding clamp for pol delta and pol epsilon	20p12	NM_002592
REV3L (POLZ)	DNA pol zeta catalytic subunit, essential function	6q21	NM_002912
REV7 (MAD2L2)	DNA pol zeta subunit	1p36	NM_006341
REV1	dCMP transferase	2q11.1-q11.2	NM_016316
POLH	XP variant	6p12.2-p21.1	NM_006502
POLI (RAD30B)	Lesion bypass	18q21.1	NM_007195
POLQ	DNA cross-link repair	3q13.31	NM_006596
DINB1 (POLK)	Lesion bypass	5q13	NM_016218
POLL	Meiotic function	10q23	NM_013274
POLM	Presumed specialized lymphoid function	7p13	NM_013284
TRF4-1	Sister-chromatid cohesion	5p15	AF089896
TRF4-2	Sister-chromatid cohesion	16p13.3	AF089897
	Editing and processing nucleases		
FEN1 (DNase IV)	5′ nuclease	11q12	NM_004111
TREX1 (DNase III)	3′ exonuclease	3p21.2-p21.3	NM_007248
TREX2	3′ exonuclease	Xq28	NM_007205
EX01 (HEX1)	5′ exonuclease	1q42-q43	NM_003686
SPO11	endonuclease	20q13.2-q13.3	NM_012444
	Rad6 pathway		
UBE2A (RAD6A)	Ubiquitin-conjugating enzyme	Xq24-q25	NM_003336
UBE2B (RAD6B)	Ubiquitin-conjugating enzyme	5q23-q31	NM_003337
RAD18	Assists repair or replication of damaged DNA	3p24-p25	AB035274
UBE2VE (MMS2)	Ubiquitin-conjugating complex	8p	AF049140
UBE2N (UBC13, BTG1)	Ubiquitin-conjugating complex	12	NM_003348
	Genes defective in diseases associated with sensitivity to DNA damaging agents		
BLM	Bloom syndrome helicase	15q26.1	NM_000057
WRN	Werner syndrome helicase/3′-exonuclease	8p12-p11.2	NM_000553
RECQL4	Rothmund-Thompson syndrome	8q24.3	NM_004260
ATM	Ataxia telangiectasia	11q22-q23	NM_000051
Fanconi anemia			
FANCA	Involved in tolerance or repair of DNA cross-links	16q24.3	NM_000135
FANCB	Involved in tolerance or repair of DNA cross-links	N/A	N/A
FANCC	Involved in tolerance or repair of DNA cross-links	9q22.3	NM_000136
FANCD	Involved in tolerance or repair of DNA cross-links	3p26-p22	N/A
FANCE	Involved in tolerance or repair of DNA cross-links	6p21-p22	NM_021922
FANCF	Involved in tolerance or repair of DNA cross-links	11p15	AF181994
FANCG (XRCC9)	Involved in tolerance or repair of DNA cross-links	9p13	NM_004629
	Other identified genes with a suspected DNA repair function		
SNM1 (PS02)	DNA cross-link repair	10q25	D42045
SNM1B	Related to SNM1	1p13.1-p13.3	AL137856
SNM1C	Related to SNM1	10	AA315885
RPA4	Similar to RPA2	Xq	NM_013347
ABH (ALKB)	Resistance to alkylation damage	14q24	X91992
PNKP	Converts some DNA breaks to ligatable ends	19q13.3-q13.4	NM_007254

Table 1. Continued.

Gene name (synonyms)	Activity	Chromosome location	Accession number
	Other conserved DNA damage response genes		
ATR	ATM- and PI-3K–like essential kinase	3q22-q24	NM_001184
RAD1 (S. pombe) homolog	PCNA-like DNA damage sensor	5p13.3-p13.2	NM_002853
RAD9 (S. pombe) homolog	PCNA-like DNA damage sensor	11q13.1-q13.2	NM_004584
HUS1 (S. pombe) homolog	PCNA-like DNA damage sensor	7p13-p12	NM_004507
RAD17 (RAD24)	RFC-like DNA damage sensor	5q13	NM_002873
TP53BP1	BRCT protein	15q15-q21	NM_005657
CHEK1	Effector kinase	11q22-q23	NM_001274
CHK2 (Rad53)	Effector kinase	22q12.1	NM_007194

The human genome sequence has already markedly influenced the field of DNA repair. Many of the genes listed were discovered as investigators searched the expanding database for sequence similarity to genes discovered in model organisms. This approach will no doubt continue, and new human genes will be identified as additional repair functions are identified in other systems. One source that is likely to be fruitful is the genome of *Deinococcus radiodurans* (*9*). This bacterium has an exceptionally high resistance to DNA-damaging agents, especially ionizing radiation, in comparison to other microorganisms. Some of the currently uncharacterized genes in *D. radiodurans* are expected to contribute to DNA repair, and it remains to be seen if there will be homologs of such functions in the human genome.

The sequence database also makes it increasingly straightforward to use mass spectrometry fingerprinting to identify new subunits of repair protein complexes (*10*). In this sensitive technique, isolated proteins are digested with an enzyme such as trypsin, and the exact molecular masses of the resulting fragments are measured. Comparison of these fragments with a computer-simulated tryptic digest of each human gene product can unambiguously identify the protein.

In addition, new genes will be found as novel biochemical assays are developed for various aspects of repair. For example, human cells can repair cross-links between the two DNA strands. Interstrand cross-links are generated by natural psoralen compounds and their chemotherapeutic derivatives, by other drugs used for cancer treatment such as nitrogen mustards, and to some extent by ionizing and ultraviolet radiation. Repair of such cross-links involves the NER genes and the XRCC2 and XRCC3 recombination genes and is predicted to involve the DNA polymerase POLQ. In addition, the sensitivity of cells from individuals with Fanconi anemia (FA) points to a role for the FANC group of genes in cross-link repair. However, the mechanism of interstrand DNA cross-link repair remains obscure, and further investigation may implicate even more gene products.

Several other classes of DNA damage exist for which repair has been relatively unexplored. New genes may be identified, for instance, involved in the repair of damage caused by lipid peroxidation (*1*). Other uncharacterized forms of DNA damage caused by reactive metabolites and catabolites may be found. For example, the genome is dynamic, and single-stranded regions are temporarily exposed during DNA replication and gene transcription. Positions that are normally protected by base-pairing within the double helical structure are then vulnerable to group-specific reagents, creating new classes of lesions. Alkylating agents can form the cytotoxic lesions 1-methyladenine and 3-methylcytosine in single-stranded DNA, and new repair strategies may be needed to remove such lesions.

DNA is assembled into several levels of ordered chromatin structure, and so DNA metabolic processes need a close connection with proteins that allow chromatin remodeling or disassembly. Several human chromatin remodeling complexes are known, for instance, that allow and control access to DNA during gene transcription (*11*). The great majority of enzymological DNA repair studies to date have worked with naked DNA, but chromatin presents a substantial barrier to recognition of DNA damage. It is expected that human protein complexes will be found that are dedicated to DNA repair and recombination, facilitating access of DNA repair enzymes to the genome.

The three-dimensional structures of DNA repair proteins are being determined at an ever-increasing pace (*12*). Structural biologists will soon turn their attention to open reading frames of unknown function, and new repair genes will become apparent in the process. As an example, the functionally related SMUG1, TDG, and UNG enzymes show little or no primary sequence homology yet have common structural folds and belong to a single protein superfamily (*13*). As the structures of new protein folds are documented, more members of DNA repair enzyme families are likely to be found with the aid of three-dimensional structure prediction models. In this way, the new field of structural genomics will help guide functional studies of presently uncharacterized open reading frames in the human genome.

For an impressive number of genes involved in human DNA repair, disruptions of the corresponding murine genes have been reported (*14*), are in progress, or have recently been constructed. The results are beginning to guide searches for additional DNA repair enzymes. Knockouts of DNA glycosylases in mice have unexpectedly mild consequences by comparison with budding yeast and *E. coli* models. This implies that more backup systems exist, probably because endogenous damage presents a more frequent problem for larger genomes.

As the genes from the human genome sequence continue to be cataloged, studying the activity of the protein products will become increasingly important. More effective methods for rapid expression of active proteins will be required to test for possible functions. An alternative approach is to selectively inactivate individual proteins in vivo. An efficient method for selective proteolytic destruction has been successful in budding yeast (*15*) and should be extendable to mammalian cells. Alternatively, systematic interference with gene expression with the use of inhibitory RNA molecules, as employed successfully in *C. elegans* (*16*), is proving to be a powerful way to dissect gene functions.

Intense activity is being devoted to understanding how DNA damage transmits signals to the cell-cycle checkpoint machinery and to the monitoring systems that control cellular apoptosis. There is recent progress on this complex extended network, which involves damage recognition factors, protein kinases, and transcription factors such as p53 (*17*). Attempts are already being made to obtain an integrated picture of DNA repair with regard to signaling (*18*). The subject is of great interest as some inherited human syndromes associated with sensitivity to DNA-damaging agents result from loss of functions such as ATM, which is involved in damage sensing.

New clinical applications relating to human DNA repair genes are certain to emerge. Tumor cells often acquire resistance to therapeutic drugs or radiation. Genomics approaches such as array technology will be used to define any DNA repair genes that may be overexpressed in this context. Furthermore, it will be important to find ways to

specifically inhibit DNA repair in these resistant cells by targeting the key enzymes. Genetic polymorphisms in relevant repair genes will be identified and efforts made to correlate them with effects on activity of the respective proteins, with response to particular therapies and with clinical outcomes. Although a number of polymorphisms in DNA repair genes are being reported, there is presently little functional information on the consequences of the attendant amino acid changes. It will be important to find out which polymorphisms actually affect protein function and then concentrate on these in epidemiological and clinical studies. For example, homozygosity for a particular polymorphism in the DNA ligase subunit XRCC1 is associated with higher sister chromatid exchange frequencies in smokers, suggesting an association of this allele with a higher risk for tobacco- and age-related DNA damage (*19*). Larger studies and comparison with other polymorphisms having known biochemical effects will be needed to further validate and extend these findings.

Furthermore, with the use of gene and protein array techniques, it should be possible to compare expression profiles of DNA repair genes in normal and tumor cells—information that could eventually lead to individually tailored therapies with chemicals and radiation. For example, tumors with low levels of NER should be more susceptible to treatment with cisplatin (*20*). In experimental systems, MMR-deficient cells are highly tolerant to alkylating chemotherapeutic drugs. MMR-defective tumors such as those found in hereditary nonpolyposis colon cancer may be resistant to treatment with such agents (*21*).

Some variation in DNA repair gene expression is epigenetic in origin and has been found for instance with MGMT and MSH6 (*22*). The MGMT gene promoter is often methylated in gliomas, resulting in suppressed expression that can be associated with an improved response after tumor treatment with an alkylating agent (*23*). The complete human genome sequence now allows the definition of promoter regions so that the DNA methylation status of relevant CpG islands can be investigated readily. Finally, DNA repair, especially repair of oxidative damage, has often been suggested as a relevant factor in counteracting aging. An examination of polymorphisms and gene expression levels in human DNA repair genes and a comparison with the equivalent genes in shorter lived mammalian species should help determine the importance of DNA repair in normal aging processes.

References and Notes

1. T. Lindahl, R. D. Wood, *Science* **286**, 1897 (1999).
2. The data in this paper were based on searches of the Ensembl sequence data freeze of 17 July 2000 and analysis updates as of 13 December 2000; see www.ensembl.org. Additional data were from the Golden Path server at the University of California at Santa Cruz (http://genome.ucsc.edu/). Expressed sequence tag (EST) searches used dbEST [M. S. Boguski, T. M. Lowe, C. M. Tolstoshev, *Nature Genet.* **4**, 332 (1993)].
3. J. A. Eisen, P. C. Hanawalt, *Mutat. Res. DNA Repair* **435**, 171 (1999).
4. L. Aravind, D. R. Walker, E. V. Koonin, *Nucleic Acids Res.* **27**, 1223 (1999).
5. The Arabidopsis Genome Initiative, *Nature* **408**, 796 (2000).
6. An exact match to much of *E. coli tag1* was found (AC010537.2), indicating that there is occasional contamination of the draft sequence that warrants caution.
7. J. J. Sekelsky, M. H. Brodsky, K. C. Burtis, *J. Cell Biol.* **150**, F31 (2000).
8. A. Constantinou, A. A. Davies, S. C. West, *Cell* **104**, 259 (2001).
9. O. White *et al.*, *Science* **286**, 1571 (1999).
10. R. E. Banks *et al.*, *Lancet* **356**, 1749 (2000).
11. M. Vignali, A. H. Hassan, K. E. Neely, J. L. Workman, *Mol. Cell. Biol.* **20**, 1899 (2000).
12. J. A. Tainer, E. C. Friedberg, *Mutat. Res.* **460**, 139 (2000). See also the other reviews on the structural biology of DNA repair in this August 2000 issue.
13. K. A. Haushalter, M. W. Todd-Stukenberg, M. W. Kirschner, G. L. Verdine, *Curr. Biol.* **9**, 174 (1999); L. Aravind, E. V. Koonin, *Genome Biol.* **1**, research0007.1 (2000) (available at genomebiology.com/2000/1/4/research/0007).
14. E. C. Friedberg, L. B. Meira, *Mutat. Res. DNA Repair* **459**, 243 (2000).
15. K. Labib, J. A. Tercero, J. F. X. Diffley, *Science* **288**, 1643 (2000).
16. P. Gonczy *et al.*, *Nature* **408**, 331 (2000); A. G. Fraser *et al.*, *Nature* **408**, 325 (2000).
17. B. B. S. Zhou, S. J. Elledge, *Nature* **408**, 433 (2000).
18. K. W. Kohn, *Mol. Biol. Cell* **10**, 2703 (1999).
19. E. J. Duell *et al.*, *Carcinogenesis* **21**, 965 (2000).
20. B. Köberle, J. R. W. Masters, J. A. Hartley, R. D. Wood, *Curr. Biol.* **9**, 273 (1999).
21. N. Claij, H. te Riele, *Exp. Cell Res.* **246**, 1 (1999).
22. A. Bearzatto, M. Szadkowski, P. Macpherson, J. Jiricny, P. Karran, *Cancer Res.* **60**, 3262 (2000).
23. M. Esteller *et al.*, *N. Engl. J. Med.* **343**, 1350 (2000).
24. Supplemental material is available on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1284/DC1.

27 September 2000; accepted 9 January 2001

The Human Transcriptome Map: Clustering of Highly Expressed Genes in Chromosomal Domains

Huib Caron,[1,2] Barbera van Schaik,[1,3] Merlijn van der Mee,[3] Frank Baas,[4] Gregory Riggins,[6] Peter van Sluis,[1] Marie-Christine Hermus,[1] Ronald van Asperen,[1] Kathy Boon,[1] P. A. Voûte,[2] Siem Heisterkamp,[5] Antoine van Kampen,[3] Rogier Versteeg[1]

The chromosomal position of human genes is rapidly being established. We integrated these mapping data with genome-wide messenger RNA expression profiles as provided by SAGE (serial analysis of gene expression). Over 2.45 million SAGE transcript tags, including 160,000 tags of neuroblastomas, are presently known for 12 tissue types. We developed algorithms to assign these tags to UniGene clusters and their chromosomal position. The resulting Human Transcriptome Map generates gene expression profiles for any chromosomal region in 12 normal and pathologic tissue types. The map reveals a clustering of highly expressed genes to specific chromosomal regions. It provides a tool to search for genes that are overexpressed or silenced in cancer.

GeneMap'99 (*1*) gives the chromosomal position of 45,049 human expressed sequence tags (ESTs) and genes belonging to 24,106 UniGene clusters. To obtain an expression profile of these genes, we made use of the SAGE technology and databases. SAGE can quantitatively identify all transcripts expressed in a tissue or cell line (*2*). It is based on the extraction of a 10–base pair (bp) tag from a fixed position in each transcript and the sequencing of thousands of these tags. Software programs and databases support the identification of the mRNAs corresponding to the tags in a SAGE library. However, this step is prone to errors, and tag assignment requires manual verification. The National Center for Biotechnology Information (NCBI) SAGEmap database has electronically extracted tags from mRNAs and ESTs in UniGene clusters. A manual

[1]Department of Human Genetics, [2]Department of Pediatric Oncology, Emma Children's Hospital, Academic Medical Center, University of Amsterdam, Post Office Box 22700, 1100 DE Amsterdam, Netherlands. [3]Bioinformatics Laboratory, [4]Neurozintuigen Laboratory, [5]Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands. [6]Department of Pathology and Department of Genetics, Duke University Medical Center, Durham, NC 27710, USA.