likely GTPases, as indicated by the activity of CITTA and HET-E [E. V. Koonin, L. Aravind, Trends Biochem. Sci. 25, 223 (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)].

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 Dicyostelium; 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 metacaspases and paracaspases 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 NF-kB pathway and interactions with the TRAFs [B. Beyaert, K. Heynck, S. Van Hulstel, Biochem. Pharmacol. 60, 1143 (2000)]. A20 belongs to a distinct family of predicted thiol proteases that is conserved in all 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 to other RING fingers, then in TRAF6 is an E3-like ubiquitin ligase pathway [L. Deng et al., Cell 103, 351 (2000)].
21. The AP-GTPass is a previously undetected predicted GTPass typhified by the COOH-terminal domain of the conserved apoptosis regulator, the DAP protein kinase [B. Inbal et al., Nature 390, 180 (1997)]. This predicted GTPass family appears to be the sister group of the RAS/ARF family GTPasses, but differs from them in having a divergent P-loop motif and a THXN instead of the NXXD signature motif. Additional AP-GTPasses are found in plants and animals, as multidomain proteins that also contain ankyrin, Lr, and kinase domains. This domain architecture suggests that AP-GTPasses participate in GTPass-dependent assembly of signaling complexes.
23. The ZUS 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.
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.

Human DNA Repair Genes
Richard D. Wood,1* Michael Mitchell,2 John Sgunoua,2 Tomas Lindahl1

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.

1*Imperial Cancer Research Fund, Clare Hall Laboratories, Blanche Lane, South Mimms, Herts EN6 3LD, UK.
2Imperial Cancer Research Fund, 44 Lincoln’s Inn Fields, London WC2A 3PX, UK.
3Present address: University of Pittsburgh Cancer Institute, 5867 Scaife Hall, 3550 Terrace Street, Pittsburgh, PA 15261, USA.

*Present address: University of Pittsburgh Cancer Institute, 5867 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 O6 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 NH2-terminalmost 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 photolases 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 photolases in the human genome, confirming previous reports that photolyase activity is present in many vertebrates including fish, reptiles, and marsupials, but not in placent 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 UvABC 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 precission intermediate in human cells requires two DNA helicases, XPB and XPD, instead of the single UvrC 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 (ATPas).

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 by 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 singleRecA 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 phosphor-ylate 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 AL139358). 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 on Science Online. A version with updates is available at www.cgal.icnet.uk/DNA_Repair_Genes.html. XP, xeroderma pigmentosum.

<table>
<thead>
<tr>
<th>Gene name (synonyms)</th>
<th>Activity</th>
<th>Chromosome location</th>
<th>Accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base excision repair (BER)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNG</td>
<td>DNA glycosylases: major altered base released</td>
<td>12q23-q24.1</td>
<td>NM_003362</td>
</tr>
<tr>
<td>SMUG1</td>
<td>U</td>
<td>12q13.1-q14</td>
<td>NM_014311</td>
</tr>
<tr>
<td>MBD4</td>
<td>U or T opposite G at CpG sequences</td>
<td>3q21-q22</td>
<td>NM_003925</td>
</tr>
<tr>
<td>TDG</td>
<td>U, T or ethenoC opposite G</td>
<td>12q24.1</td>
<td>NM_003211</td>
</tr>
<tr>
<td>OGG1</td>
<td>8-oxoG opposite C</td>
<td>3p26.2</td>
<td>NM_002542</td>
</tr>
<tr>
<td>MTH</td>
<td>A opposite 8-oxoG</td>
<td>1p34.3-p32.1</td>
<td>NM_012222</td>
</tr>
<tr>
<td>NTH1</td>
<td>Ring-saturated or fragmented pyrimidines</td>
<td>16p13.3-p13.2</td>
<td>NM_002528</td>
</tr>
<tr>
<td>MPG</td>
<td>3-meA, ethenoA, hypoxanthine</td>
<td>16p13.3</td>
<td>NM_002434</td>
</tr>
<tr>
<td><strong>Other BER factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APE1 (HAP1, APEX, REF1)</td>
<td>AP endonuclease</td>
<td>1q42</td>
<td>NM_001641</td>
</tr>
<tr>
<td>APE2 (APEXL2)</td>
<td>AP endonuclease</td>
<td>X</td>
<td>NM_014481</td>
</tr>
<tr>
<td>LIG3</td>
<td>Main ligation function</td>
<td>17q11.2-q12</td>
<td>NM_013975</td>
</tr>
<tr>
<td>XRCC1</td>
<td>Main ligation function</td>
<td>19q13.2</td>
<td>NM_006297</td>
</tr>
<tr>
<td><strong>Poly(ADP-ribose) polymerase (PARP) enzymes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADPRT</td>
<td>Protects strand interruptions</td>
<td>1q42</td>
<td>NM_001618</td>
</tr>
<tr>
<td>ADPRTL2</td>
<td>PARP-like enzyme</td>
<td>1q14.12-q12</td>
<td>NM_005485</td>
</tr>
<tr>
<td>ADPRTL3</td>
<td>PARP-like enzyme</td>
<td>3p21.1-p22.2</td>
<td>AF085734</td>
</tr>
<tr>
<td><strong>Mismatch excision repair (MMR)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSH2</td>
<td>Mismatch and loop recognition</td>
<td>2p22-p21</td>
<td>NM_002051</td>
</tr>
<tr>
<td>MSH3</td>
<td>Mismatch and loop recognition</td>
<td>5q11-q12</td>
<td>NM_002439</td>
</tr>
<tr>
<td>MSH6</td>
<td>Mismatch recognition</td>
<td>2p16</td>
<td>NM_000179</td>
</tr>
<tr>
<td>MSH4</td>
<td>MutS homolog specialized for meiosis</td>
<td>1p31</td>
<td>NM_002440</td>
</tr>
<tr>
<td>MSH5</td>
<td>MutS homolog specialized for meiosis</td>
<td>6q21.3</td>
<td>NM_002441</td>
</tr>
<tr>
<td>PMS1</td>
<td>Mitochondrial MutL homolog</td>
<td>2q31.1</td>
<td>NM_000534</td>
</tr>
<tr>
<td>MLH1</td>
<td>MutL homolog</td>
<td>3p21.3</td>
<td>NM_000249</td>
</tr>
<tr>
<td>PMS2</td>
<td>MutL homolog</td>
<td>7p22</td>
<td>NM_000535</td>
</tr>
<tr>
<td>MLH3</td>
<td>MutL homolog of unknown function</td>
<td>1q42.3</td>
<td>NM_014381</td>
</tr>
<tr>
<td>PMS2L3</td>
<td>MutL homolog of unknown function</td>
<td>7q11-q22</td>
<td>D38437</td>
</tr>
<tr>
<td>PMS2L4</td>
<td>MutL homolog of unknown function</td>
<td>7q11-q22</td>
<td>D38438</td>
</tr>
<tr>
<td><strong>Nucleotide excision repair (NER)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XPC</td>
<td>Binds damaged DNA as complex</td>
<td>3p25</td>
<td>NM_004628</td>
</tr>
<tr>
<td>RAD23B (HR23B)</td>
<td>Binds damaged DNA as complex</td>
<td>3p25.1</td>
<td>NM_002874</td>
</tr>
<tr>
<td>CETN2</td>
<td>Binds damaged DNA as complex</td>
<td>Xq28</td>
<td>NM_004344</td>
</tr>
<tr>
<td>RAD23A (HR23A)</td>
<td>Substitutes for HR23B</td>
<td>1p31.2</td>
<td>NM_000503</td>
</tr>
<tr>
<td>XPA</td>
<td>Binds damaged DNA in preincision complex</td>
<td>9q22.3</td>
<td>NM_000380</td>
</tr>
<tr>
<td>RAD51</td>
<td>DNA in preincision complex</td>
<td>17p13.3</td>
<td>D002954</td>
</tr>
<tr>
<td>RAD52</td>
<td>DNA in preincision complex</td>
<td>1p35</td>
<td>NM_002946</td>
</tr>
<tr>
<td>RPA3</td>
<td>DNA in preincision complex</td>
<td>7p22</td>
<td>NM_002947</td>
</tr>
<tr>
<td>TFIIH</td>
<td>Catalyzes unwinding in preincision complex</td>
<td>2q21</td>
<td>NM_000122</td>
</tr>
<tr>
<td><strong>Homologous recombination</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XPB (ERCC3)</td>
<td>3' to 5' DNA helicase</td>
<td>1q33.2-q13.3</td>
<td>X52221</td>
</tr>
<tr>
<td>XPD (ERCC2)</td>
<td>5' to 3' DNA helicase</td>
<td>1p11.1-p14</td>
<td>NM_005316</td>
</tr>
<tr>
<td>GTF2H1</td>
<td>Core TFIIH subunit p62</td>
<td>1p11.1-p14</td>
<td>NM_005316</td>
</tr>
<tr>
<td>GTF2H2</td>
<td>Core TFIIH subunit p44</td>
<td>5q12.2-q13.3</td>
<td>NM_001515</td>
</tr>
<tr>
<td>GTF2H3</td>
<td>Core TFIIH subunit p34</td>
<td>2q12</td>
<td>NM_001516</td>
</tr>
<tr>
<td>GTF2H4</td>
<td>Core TFIIH subunit p52</td>
<td>5p21.3</td>
<td>NM_001517</td>
</tr>
<tr>
<td>CDK7</td>
<td>Kinase subunit of TFIIH</td>
<td>2q21.5-cen</td>
<td>NM_001799</td>
</tr>
<tr>
<td>CCNH</td>
<td>Kinase subunit of TFIIH</td>
<td>5q13.3-q14</td>
<td>NM_001239</td>
</tr>
<tr>
<td>MAND1</td>
<td>Kinase subunit of TFIIH</td>
<td>1q42.3</td>
<td>NM_002431</td>
</tr>
<tr>
<td>XPC</td>
<td>3' incision</td>
<td>1q42.3</td>
<td>NM_000123</td>
</tr>
<tr>
<td>ERCC1</td>
<td>5' incision subunit</td>
<td>19q13.2-q13.3</td>
<td>NM_001983</td>
</tr>
<tr>
<td>XPF (ERCC4)</td>
<td>5' incision subunit</td>
<td>16p13.3-p13.13</td>
<td>NM_005236</td>
</tr>
<tr>
<td>LIG1</td>
<td>DNA joining</td>
<td>19q13.2-q13.3</td>
<td>NM_000234</td>
</tr>
<tr>
<td><strong>NER-related</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSA (CKN1)</td>
<td>Cockayne syndrome; needed for transcription-coupled NER</td>
<td>5q12-q31</td>
<td>NM_000082</td>
</tr>
<tr>
<td>CSB (ERCC6)</td>
<td>Cockayne syndrome; needed for transcription-coupled NER</td>
<td>1q11</td>
<td>NM_000124</td>
</tr>
<tr>
<td>XAB2 (HCNP)</td>
<td>Cockayne syndrome; needed for transcription-coupled NER</td>
<td>19</td>
<td>NM_020196</td>
</tr>
<tr>
<td>DDB1</td>
<td>Complex defective in XP group E</td>
<td>11q12-q13</td>
<td>NM_001923</td>
</tr>
<tr>
<td>DDB2</td>
<td>Mutated in XP group E</td>
<td>11p12-p11</td>
<td>NM_000107</td>
</tr>
<tr>
<td>MMS19</td>
<td>Transcription and NER</td>
<td>1q42.4</td>
<td>AW852889</td>
</tr>
<tr>
<td><strong>Homologous recombination</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RADS1</td>
<td>Homologous pairing</td>
<td>1q5.1</td>
<td>NM_002875</td>
</tr>
<tr>
<td>RADS1L1 (RADS1B)</td>
<td>Rad51 homolog</td>
<td>1q42-q24</td>
<td>U84138</td>
</tr>
<tr>
<td>RADS1L1 (RADS1D)</td>
<td>Rad51 homolog</td>
<td>17q11-qter</td>
<td>NM_000287</td>
</tr>
</tbody>
</table>

**XPANALYSIS OF GENOMIC INFORMATION**
### Table 1. Continued.

<table>
<thead>
<tr>
<th>Gene name (synonyms)</th>
<th>Activity</th>
<th>Chromosome location</th>
<th>Accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMC1</td>
<td>Rad51 homolog, meiosis</td>
<td>22q13.1</td>
<td>NM_007068</td>
</tr>
<tr>
<td>XRCC2</td>
<td>DNA break and cross-link repair</td>
<td>7q36.1</td>
<td>NM_005431</td>
</tr>
<tr>
<td>XRCC3</td>
<td>DNA break and cross-link repair</td>
<td>14q22.3</td>
<td>NM_005432</td>
</tr>
<tr>
<td>RAD5</td>
<td>Accessory factor for recombination</td>
<td>12p13-p12.2</td>
<td>NM_0002879</td>
</tr>
<tr>
<td>RAD54L</td>
<td>Accessory factor for recombination</td>
<td>1p32</td>
<td>NM_003579</td>
</tr>
<tr>
<td>RAD54B</td>
<td>Accessory factor for recombination</td>
<td>8q13.3-q22</td>
<td>NM_012415</td>
</tr>
<tr>
<td>BRCA1</td>
<td>Accessory factor for transcription and recombination</td>
<td>17q21</td>
<td>NM_007295</td>
</tr>
<tr>
<td>BRCA2</td>
<td>Cooperation with RAD51, essential function</td>
<td>13q12.3</td>
<td>NM_000059</td>
</tr>
<tr>
<td>RAD50</td>
<td>ATPase in complex with MRE11A, NBS1</td>
<td>5q31</td>
<td>NM_005732</td>
</tr>
<tr>
<td>MRE11A</td>
<td>3' exonuclease</td>
<td>11q21</td>
<td>NM_005590</td>
</tr>
<tr>
<td>NBS1</td>
<td>Mutated in Nijmegen breakage syndrome</td>
<td>8q21-q24</td>
<td>NM_002485</td>
</tr>
<tr>
<td><strong>Nonhomologous end-joining</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ku70 (G22P1)</td>
<td>DNA end binding</td>
<td>22q13.2-q13.31</td>
<td>NM_001469</td>
</tr>
<tr>
<td>Ku80 (XRCC5)</td>
<td>DNA end binding</td>
<td>2q35</td>
<td>M30938</td>
</tr>
<tr>
<td>PRKDC</td>
<td>DNA-dependent protein kinase catalytic subunit</td>
<td>8q11</td>
<td>NM_006904</td>
</tr>
<tr>
<td>LIG4</td>
<td>Nonhomologous end-joining</td>
<td>13q33-q34</td>
<td>NM_002312</td>
</tr>
<tr>
<td>XRCC4</td>
<td>Nonhomologous end-joining</td>
<td>13q33-q34</td>
<td>NM_002312</td>
</tr>
<tr>
<td><strong>Sanitization of nucleotide pools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ku70 (G22P1)</td>
<td>DNA end binding</td>
<td>22q13.2-q13.31</td>
<td>NM_001469</td>
</tr>
<tr>
<td>Ku80 (XRCC5)</td>
<td>DNA end binding</td>
<td>2q35</td>
<td>M30938</td>
</tr>
<tr>
<td>PRKDC</td>
<td>DNA-dependent protein kinase catalytic subunit</td>
<td>8q11</td>
<td>NM_006904</td>
</tr>
<tr>
<td>LIG4</td>
<td>Nonhomologous end-joining</td>
<td>13q33-q34</td>
<td>NM_002312</td>
</tr>
<tr>
<td>XRCC4</td>
<td>Nonhomologous end-joining</td>
<td>13q33-q34</td>
<td>NM_002312</td>
</tr>
<tr>
<td><strong>DNA polymerases (catalytic subunits)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POLG</td>
<td>BER in nuclear DNA</td>
<td>8p11.2</td>
<td>NM_002690</td>
</tr>
<tr>
<td>POLG</td>
<td>BER in mitochondrial DNA</td>
<td>15q25</td>
<td>NM_002693</td>
</tr>
<tr>
<td>POLD1</td>
<td>NER and MMR</td>
<td>19q13.3</td>
<td>NM_002691</td>
</tr>
<tr>
<td>POLE1</td>
<td>NER and MMR</td>
<td>12q24.3</td>
<td>NM_002631</td>
</tr>
<tr>
<td>PCNA</td>
<td>Sliding clamp for pol delta and pol epsilon</td>
<td>20p12</td>
<td>NM_002592</td>
</tr>
<tr>
<td>REV3L (POLZ)</td>
<td>DNA pol zeta catalytic subunit, essential function</td>
<td>6q21</td>
<td>NM_002912</td>
</tr>
<tr>
<td>REV7 (RAD2L2)</td>
<td>DNA pol zeta subunit</td>
<td>1p36</td>
<td>NM_006341</td>
</tr>
<tr>
<td>REV1</td>
<td>dCMP transferase</td>
<td>2q11.1-q11.2</td>
<td>NM_006341</td>
</tr>
<tr>
<td>POLH</td>
<td>XP variant</td>
<td>6p22.2-p22.1</td>
<td>NM_006502</td>
</tr>
<tr>
<td>POLI (RAD30B)</td>
<td>Lesion bypass</td>
<td>18q21.1</td>
<td>NM_007195</td>
</tr>
<tr>
<td>POLQ</td>
<td>DNA cross-link repair</td>
<td>3q13.3</td>
<td>NM_006596</td>
</tr>
<tr>
<td>DNB1 (POLK)</td>
<td>Lesion bypass</td>
<td>5q13</td>
<td>NM_016218</td>
</tr>
<tr>
<td>POLL</td>
<td>Meiotic function</td>
<td>10q23</td>
<td>NM_013274</td>
</tr>
<tr>
<td>POLM</td>
<td>Presumed specialized lymphoid function</td>
<td>7p13</td>
<td>NM_013284</td>
</tr>
<tr>
<td>TRF4-1</td>
<td>Sister-chromatid cohesion</td>
<td>5p15</td>
<td>AF089896</td>
</tr>
<tr>
<td>TRF4-2</td>
<td>Sister-chromatid cohesion</td>
<td>16p13.3</td>
<td>AF089897</td>
</tr>
<tr>
<td><strong>Editing and processing nucleases</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEN1 (DNase IV)</td>
<td>5' nuclease</td>
<td>11q12</td>
<td>NM_004111</td>
</tr>
<tr>
<td>TREX1 (DNase III)</td>
<td>3' exonuclease</td>
<td>3p21.2-p21.3</td>
<td>NM_002748</td>
</tr>
<tr>
<td>TREX2</td>
<td>3' exonuclease</td>
<td>Xq28</td>
<td>NM_002705</td>
</tr>
<tr>
<td>EXO1 (HEX1)</td>
<td>5' exonuclease</td>
<td>1q42-q43</td>
<td>NM_003686</td>
</tr>
<tr>
<td>SPO11</td>
<td>Endonuclease</td>
<td>20q13.2-q13.3</td>
<td>NM_012444</td>
</tr>
<tr>
<td><strong>Rad6 pathway</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBE2A (RAD6A)</td>
<td>Ubiquitin-conjugating enzyme</td>
<td>Xq24-q25</td>
<td>NM_003336</td>
</tr>
<tr>
<td>UBE2B (RAD6B)</td>
<td>Ubiquitin-conjugating enzyme</td>
<td>5q33-q31</td>
<td>NM_003337</td>
</tr>
<tr>
<td>RAD18</td>
<td>Assists repair or replication of damaged DNA</td>
<td>3p24-q25</td>
<td>AB035274</td>
</tr>
<tr>
<td>UBE2VE (MMS2)</td>
<td>Ubiquitin-conjugating complex</td>
<td>8p</td>
<td>AF049140</td>
</tr>
<tr>
<td>UBE2N (UBC13, BTG1)</td>
<td>Ubiquitin-conjugating complex</td>
<td>12</td>
<td>NM_003348</td>
</tr>
<tr>
<td><strong>Genes defective in diseases associated with sensitivity to DNA damaging agents</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLM</td>
<td>Bloom syndrome helicase</td>
<td>15q26.1</td>
<td>NM_000057</td>
</tr>
<tr>
<td>WRN</td>
<td>Werner syndrome helicase/3'-exonuclease</td>
<td>8p12-p11.2</td>
<td>NM_000553</td>
</tr>
<tr>
<td>RECQL4</td>
<td>Rothmund-Thompson syndrome</td>
<td>8q24.3</td>
<td>NM_004260</td>
</tr>
<tr>
<td>ATM</td>
<td>Ataxia telangiectasia</td>
<td>11q22-q23</td>
<td>NM_000051</td>
</tr>
<tr>
<td>Fanconi anemia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FANCA</td>
<td>Involved in tolerance or repair of DNA cross-links</td>
<td>16q24.3</td>
<td>NM_000135</td>
</tr>
<tr>
<td>FANCB</td>
<td>Involved in tolerance or repair of DNA cross-links</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>FANCC</td>
<td>Involved in tolerance or repair of DNA cross-links</td>
<td>9q22.3</td>
<td>NM_000136</td>
</tr>
<tr>
<td>FANCD</td>
<td>Involved in tolerance or repair of DNA cross-links</td>
<td>3p26-p22</td>
<td>N/A</td>
</tr>
<tr>
<td>FANCE</td>
<td>Involved in tolerance or repair of DNA cross-links</td>
<td>6p21-p22</td>
<td>NM_021922</td>
</tr>
<tr>
<td>FANC1</td>
<td>Involved in tolerance or repair of DNA cross-links</td>
<td>11q13</td>
<td>AF181994</td>
</tr>
<tr>
<td>FANC1 (XRCC9)</td>
<td>Involved in tolerance or repair of DNA cross-links</td>
<td>9p13</td>
<td>NM_004629</td>
</tr>
<tr>
<td><strong>Other identified genes with a suspected DNA repair function</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNM1 (PS02)</td>
<td>DNA cross-link repair</td>
<td>10q25</td>
<td>D42045</td>
</tr>
<tr>
<td>SNM1B</td>
<td>Related to SNM1</td>
<td>1p13.1-p13.3</td>
<td>AL137856</td>
</tr>
<tr>
<td>SNM1C</td>
<td>Related to SNM1</td>
<td>10</td>
<td>AA315885</td>
</tr>
<tr>
<td>RPA4</td>
<td>Similar to RPA2</td>
<td>Xq</td>
<td>NM_013347</td>
</tr>
<tr>
<td>ABH (ALKB)</td>
<td>Resistance to alkylation damage</td>
<td>14q24</td>
<td>X91992</td>
</tr>
<tr>
<td>PNKP</td>
<td>Converts some DNA breaks to ligatable ends</td>
<td>19q13.3-q13.4</td>
<td>NM_007254</td>
</tr>
</tbody>
</table>

**Analysis of Genomic Information**

www.sciencemag.org SCIENCE VOL 291 16 FEBRUARY 2001
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 characterized 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 trypptic 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 (I). 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 remodelling or disassembly. Several human chromatin remodelling complexes are known, for instance, that allow and control access to DNA during gene transcription (1I). 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 through 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, though not exclusively, found 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
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]).
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.
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 Sluijs,1 Marie-Christine Hermus,1 Ronald van Asperen,1 Kathy Boon,1 P. A. Voûte,2 Siem Heisterkamp,5 Antoine van Kampen,3 Rogier Versteeg1

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.

References and Notes
1. Department of Human Genetics, 2Department of Pediatric Oncology, Emma Children’s Hospital, Academic Medical Center, University of Amsterdam, Post Office Box 22700, 1100 DE Amsterdam, Netherlands. 3Bioinformatics Laboratory, 4Neurointuïtein Laboratory, 5Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands. 6Department of Pathology and Department of Genetics, Duke University Medical Center, Durham, NC 27710, USA.
Human DNA Repair Genes
Richard D. Wood, Michael Mitchell, John Sgouros and Tomas Lindahl

Science 291 (5507), 1284-1289.
DOI: 10.1126/science.1056154