# An Evolutionary Analysis of the Helix-Hairpin-Helix Superfamily of DNA Repair Glycosylases

*Dee R. Denver, Stephanie L. Swenson, and Michael Lynch*

Department of Biology, Indiana University

The helix-hairpin-helix (HhH) superfamily of base excision repair DNA glycosylases is composed of multiple phylogenetically diverse enzymes that are capable of excising varying spectra of oxidatively and methyl-damaged bases. Although these DNA repair glycosylases have been widely studied through genetic, biochemical, and biophysical approaches, the evolutionary relationships of different HhH homologs and the extent to which they are conserved across phylogeny remain enigmatic. We provide an evolutionary framework for this pervasive and versatile superfamily of DNA glycosylases. Six HhH gene families (named AlkA: alkyladenine glycosylase; MpgII: N-methylpurine glycosylase II; MutY/Mig: A/G-specific adenine glycosylase/mismatch glycosylase; Nth: endonuclease III; OggI: 8-oxoguanine glycosylase I; and OggII: 8-oxoguanine glycosylase II) are identified through phylogenetic analysis of 234 homologs found in 94 genomes (16 archaea, 64 bacteria, and 14 eukaryotes). The number of homologs in each gene family varies from 117 in the Nth family (nearly every genome surveyed harbors at least one Nth homolog) to only five in the divergent OggII family (all from archaeal genomes). Sequences from all three domains of life are included in four of the six gene families, suggesting that the HhH superfamily diversified very early in evolution. The phylogeny provides evidence for multiple lineage-specific gene duplication events, most of which involve eukaryotic homologs in the Nth and AlkA gene families. We observe extensive variation in the number of HhH superfamily glycosylase genes present in different genomes, possibly reflecting major differences among species in the mechanisms and pathways by which damaged bases are repaired and/or disparities in the basic rates and spectra of mutation experienced by different genomes.

## Introduction

Genomic stability in all forms of life is under constant threat by numerous mutagenic sources that include replication errors, environmental factors such as UV radiation, and endogenous mutagens such as oxygen free radicals (Lindahl 1993). Damaged DNA structures can have profound biological consequences that include the disruption of basic cellular and metabolic processes, tumorigenesis, and mutation (Lindahl 1993; Demple and Harrison 1994; Backlund et al. 2001; Epel 2003). Multiple DNA repair systems have evolved in response to the pressures imposed by DNA damage and drastically reduce the overall damage load and spectrum of mutations suffered by genomes over time (Eisen and Hanawalt 1999). Most DNA repair pathways have been detected in all three domains of life (Eisen and Hanawalt 1999; Weller et al. 2002), suggesting ancient origins.

The base excision DNA repair (BER) pathway involves the initial recognition of specific types of base damage (such as 8-oxoguanine and 3-methyladenine) and/or damage-associated base-pairing mismatches by a BER DNA glycosylase. After damage recognition, the glycosylase cleaves the damaged or misincorporated base from the DNA backbone, leaving an apurinic/apyrimidinic (AP) site (Lu et al. 2001). After base excision, the AP site is then recognized by an AP endonuclease that cleaves the DNA backbone at the AP site (in some cases the backbone is cleaved by the actual glycosylase, such as with *Escherichia coli* endonuclease III (Nth)) (Demple and Linn 1980). A patch on the nicked strand is then exonucleolytically digested followed by resynthesis and ligation (Lu et al. 2001).

The helix-hairpin-helix (HhH) superfamily of BER glycosylases includes a diverse assortment of enzymes capable of specifically recognizing and excising varying spectra of damaged bases and base pairing mismatches (Demple and Harrison 1994; Alseth et al. 1999; Eisen and Hanawalt 1999; Yang et al. 2000). HhH glycosylases such as Nth and 8-oxoguanine glycosylase (Ogg) remove oxidatively damaged bases from the DNA backbone (Demple and Harrison 1994; Rosenquist, Zharkov, and Grollman 1996). The archaeal mismatch glycosylase (Mig) has been shown to remove thymines from T/G mispairs that presumably arise from cytosine deamination (Yang et al. 2000) whereas A/G-specific adenine glycosylases (MutY) remove adenines mispaired with 8-oxoguanine bases and some oxidatively damaged purines (Hashiguchi et al. 2002). In vitro mutagenesis studies have shown that only two amino acid replacements can change the specificity of Mig glycosylases from T/G mispairs to A/G mispairs that are generally recognized by MutY glycosylases (Fondufe-Mittendorf et al. 2002). Alkyladenine glycosylase (AlkA) and N-methylpurine glycosylase II (MpgII) are glycosylases that excise methyl-damaged bases such as 3-methyladenine and 7-methylguanine (Begley et al. 1999). All of the BER glycosylases mentioned share the canonical HhH DNA binding domain, whose amino acid sequences are thought to impart specificity in damage recognition and provide the basis for their inclusion in the HhH superfamily of DNA glycosylases (Michaels et al. 1990; Roldan-Arjona, Anselmino, and Lindahl 1996; Begley et al. 1999; Gogos et al. 2000). Despite the widespread conservation and obvious importance of HhH glycosylases in maintaining genomic stability, limited phylogenetic information is available about this prevalent and highly adaptable superfamily of DNA repair enzymes.

Most eukaryotes harbor DNA in multiple subcellular compartments (such as nuclei, mitochondria, and

Key words: base excision repair, gene duplication, genome, glycosylase, helix-hairpin-helix, phylogenetic analysis.

E-mail: ddenver@bio.indiana.edu.

*Mol. Biol. Evol.* 20(10):1603–1611. 2003
DOI: 10.1093/molbev/msg177
*Molecular Biology and Evolution*, Vol. 20, No. 10,

chloroplasts), thereby providing an additional challenge to the maintenance of genome stability. Many eukaryotic BER glycosylases such as Nth, MutY, and uracil-DNA glycosylase have been shown to be present in both the cell nucleus and the mitochondria through subcellular localization studies (Nilsen et al. 1997; Takao et al. 1998; Alseth et al. 1999). In some instances differential glycosylase targeting is achieved through subfunctionalization of duplicated genes (Alseth et al. 1999) and in other cases by alternative splicing of a single gene (Nilsen et al. 1997; Takao et al. 1998). These studies, however, are limited to *Saccharomyces cerevisiae* and mammals; it is unclear whether most eukaryotic species target glycosylases to both the nucleus and the mitochondria and, if so, whether the duplicate gene or the alternative splicing approach is more commonly employed. Most data on DNA repair in chloroplasts is limited to studies on recombination repair and photoreactivation; relatively little is known about BER activities on chloroplast DNA (Allen and Raven 1996; Peterson and Small 2001).

Current progress in the sequencing of entire genomes from multiple species has enabled the study of DNA repair genes from a wide, "phylogenomic" vantage point (Eisen and Hanawalt 1999). Genes involved in mismatch repair and recombination repair have been the focus of many broad-based phylogenetic analyses of DNA repair (Eisen 1995, 1996; Jiricny 1998; Culligan et al. 2000), and additional studies have considered the evolution of DNA ligases and photolyases (Kanai et al. 1997; Martin and MacNeill 2002). A previous phylogenetic analysis of 43 HhH superfamily sequences has provided some insights into the evolution of these DNA glycosylases (Yang et al. 2000). However, a large-scale analysis that encompasses many genomes across the three domains of life, necessary for a broad-based understanding of the evolution of this important DNA repair gene superfamily, is lacking.

This study provides an evolutionary analysis of 234 HhH glycosylase homologs found across 94 genome sequences. We first evaluate the overall phylogenetic structure of the HhH superfamily and identify major gene families. Each of the gene families is then analyzed in terms of the abundance and phylogenetic distribution of sequences placed into each of the families. We also characterize the occurrence of putative lineage-specific gene duplication events in the different gene families. The large-scale phylogenetic analysis of the HhH superfamily presented here provides insights into the evolution of this widespread and highly adaptable group of DNA repair enzymes across the three domains of life.

## Materials and Methods

### Identification and Alignment of HhH Sequences

We initially searched for HhH superfamily glycosylases in 94 complete or nearly complete genome sequences by performing BlastP and TBlastN queries (Altschul et al. 1990) against genome databases using *Homo sapiens* (Nth, MutY, and Ogg), *S. cerevisiae* (Ntg1, Ntg2, Ogg, AlkA), and *E. coli* (Nth, MutY, AlkA) HhH superfamily sequences that were retrieved from GenBank. Sequences scoring significant hits (e scores < 0.05) were stored in a database repository. After preliminary phylogenetic analyses (described below), additional Blast searches were done using *Pyrobaculum aerophilum* Mig, *Methanobacterium thermoautotrophicum* MTHE746, and *Clostridium perfringens* Ogg protein sequences to further search for any HhH glycosylase genes missed by our initial queries. We also performed PSI-Blast searches against the complete National Center for Biotechnology (NCBI) database for a less stringent approach to finding more divergent HhH homologs (Altschul et al. 1997). PSI-Blast searches were done with the same search sequences used in conventional Blast searches (listed above), in addition to the *Methanocaldococcus jannaschi* MJ0724 protein sequence. Sequences scoring significant PSI-Blast hits from the 94 genomes surveyed that were not overlapping with conventional Blast hits were added to the database. GenBank accession numbers were provided for most sequences, except for six cases where homologs were found in unsubmitted or unannotated sequences or contigs (see Supplementary Material online, table A).

The HhH DNA binding domain and flanking sequences were identified in the 234 retrieved sequences through Blast search alignments and comparisons to published alignments (Eide et al. 1996; Roldan-Arjona, Anselmino, and Lindahl 1996). Approximately 50 (ranging from 48 to 52) conserved amino acids from the HhH domain and flanking sequence were then batch-aligned using ClustalW (Higgins, Thompson, and Gibson 1996). No other portions of the sequences were sufficiently conserved for reliable alignment across the 234 sequences retrieved. Minor alignment adjustments were then made manually using the ESEE alignment program (Eyeball Sequence Editor [ESEE] v3.01). Final aligned amino acid sequences were then converted to Mega format for subsequent phylogenetic analyses.

### Phylogenetic Analysis

Aligned amino acid sequences from the HhH domain and flanking residues were subjected to phylogenetic analysis with the Neighbor-Joining (NJ) method, using Mega version 2.1 (Kumar et al. 2001). The Poisson correction model for amino acid evolution that corrects for multiple substitutions at the same site was used, because the analysis included a large number of sequences from multiple species across the three domains of life. Statistical evaluation of the reliability of phylogenetic reconstructions was carried out in Mega using both the conventional bootstrap test and the less conservative interior branch test. Bootstrap tests were performed with 1,000 replicates. Interior branch tests were also performed: this normal deviate (Z) test operates under the null hypothesis that the interior branch under consideration has a length equal to zero (Nei and Kumar 2000). Unlike the bootstrap test, the interior branch test has the same statistical properties irrespective of the number of sequences analyzed (Sitnikova, Rzhetsky, and Nei 1995). The tree was then examined to identify major HhH gene families and putative cases of gene duplication, and to characterize the phylogenetic diversity of sequences in each gene family. All sequences analyzed were assigned Tree ID numbers to provide a simple identifier for all sequences analyzed in this study (see Supplementary Material online, table A).

## Results

### HhH Superfamily Phylogenetic Framework

We subjected 234 HhH partial protein sequences (see Supplementary Materials online, table A) from 94 bacterial, eukaryotic, and archaeal genomes to phylogenetic analysis, and the midpoint-rooted results were displayed (fig. 1). Six distinct HhH gene families were identified and named according to functionally characterized glycosylases included in the respective families (Nth, OggI, MutY/Mig, AlkA, MpgII, and OggII). The phylogeny grouped the Nth, OggI, and MutY/Mig gene families in one higher-order clade and the AlkA and MpgII families into a second clade (fig. 1). The OggII family was placed as an outgroup relative to the other five gene families. Each of the six HhH gene families was then characterized in terms of the number and phylogenetic distribution of sequences placed in each family, and to identify putative cases of gene duplication.

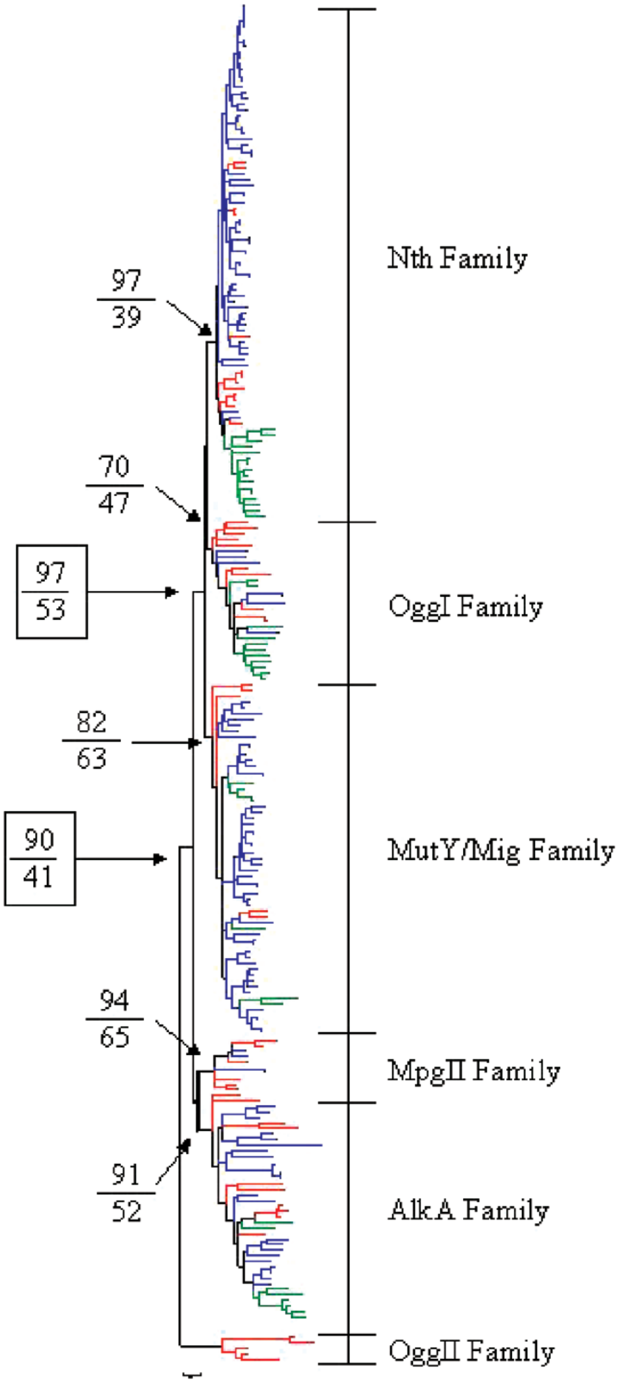

FIG. 1.—Evolution of HhH superfamily glycosylases in the three domains of life. This NJ tree was constructed using 234 sequences from 94 complete or nearly complete genome sequences. Branches leading to bacterial sequences were labeled in blue, eukaryotic sequences in green, and archaeal sequences in red. Six distinct gene families were identified and named on the right. Arrows indicate the statistical support for nodes grouping the six gene families (top number indicates interior branch test support, bottom number indicates bootstrap support). Boxed scores refer to statistical support for higher-order groupings of HhH gene families. The horizontal line at the bottom indicates 0.2 substitutions per site.

### The Nth Family

HhH glycosylases placed in the Nth gene family include all of the multiple biochemically characterized proteins related to the *E. coli* enzyme endonuclease III (removes oxidized pyrimidines), such as the human and yeast homologs (Demple and Linn 1980; Demple and Harrison 1994; Alseth et al. 1999). Every genome surveyed contained at least one glycosylase gene in the Nth family, with the exception of five archaeal genomes and the five bacterial genomes (tables 1–3). Among the six families, the Nth family contained the largest number of HhH homologs: of the 234 total sequences analyzed, 89 were placed in the Nth family. The genome of the bacterial species *Aquifex aeolicus* was found to encode two Nth homologs (Tree IDs = B57, B60) that were placed in disparate positions in the Nth family gene tree (see Supplementary Materials online, fig. 1). The bacterial genomes surveyed from *Mollicutes* (includes the four *Mycoplasma* genomes and the *Ureaplasma urealyticum* genome), a group of bacteria with extraordinarily reduced genome sizes (Rocha and Blanchard 2002), were all devoid of Nth homologs. All other bacterial genomes contained a single Nth homolog. All eukaryotes contained at least one Nth family homolog and the genomes of *Arabidopsis thaliana* and *S. cerevisiae* each had two (table 2). In both instances the two Nth homologs (*A. thaliana* Tree IDs = E8, E9; *S. cerevisiae* Tree IDs = E2, E3) were grouped together, indicating lineage-specific gene duplication events (see Supplementary Material online, fig. 1). In archaea, Nth homologs were missing from all four of the genomes surveyed from *Crenarchaeota*, as well as the genome of *Methanopyrus kandleri* from *Euryarchaeota* (table 3; see also Supplementary Material online, fig. 1). Two Nth homologs were found in each of the *Methanosarcina activorans* (Tree IDs = A3, A10) and *Methanosarcina mazei* (Tree IDs = A4, A11) genomes. For each of the *Methanosarcina* genomes surveyed, one homolog grouped with the majority of other archaeal Nth homologs

**Table 1**
**HhH Superfamily Glycosylases in Bacteria**

| | Nth | OggI | MutY | MpgII | AlkA | OggII |
|---|---|---|---|---|---|---|
| Actinobacteria | | | | | | |
| *Corynebacterium glutamicum* | + | − | + | − | − | − |
| *Mycobacterium leprae* | + | − | + | − | − | − |
| *Mycobacterium tuberculosis H37Rv* | + | − | + | − | − | − |
| *Streptomyces coelicolor A3(2)* | + | − | + | − | − | − |
| Chlamydiales | | | | | | |
| *Chlamydia muridarum* | + | − | + | − | − | − |
| *Chlamydia trachomatis* | + | − | + | − | − | − |
| *Chlamydophila pneumoniae* | + | − | + | − | − | − |
| Cyanobacteria | | | | | | |
| *Nostoc sp. PCC 7120* | + | − | − | − | − | − |
| *Thermosynechococcus elongatus BP-1* | + | − | + | − | − | |
| *Synechocystis sp. PCC 6803* | + | − | − | − | − | − |
| Firmicutes | | | | | | |
| *Bacillus anthracis str. A2012* | + | − | + | − | ++ | − |
| *Bacillus halodurans* | + | − | + | − | + | − |
| *Bacillus subtilis* | + | − | + | − | ++ | − |
| *Clostridium acetobutylicum* | + | + | + | − | − | − |
| *Clostridium perfringens* | + | + | − | − | − | − |
| *Lactococcus lactis subsp. lactis* | + | − | + | − | − | − |
| *Listeria innocua* | + | − | + | − | − | − |
| *Listeria monocytogenes EGD-e* | + | − | + | − | − | − |
| *Staphylococcus aureus Mu50* | + | − | + | + | − | − |
| *Mycoplasma genitalium* | − | − | − | − | − | − |
| *Mycoplasma penetrans* | − | − | − | − | − | − |
| *Mycoplasma pneumoniae* | − | − | − | − | − | − |
| *Mycoplasma pulmonis* | − | − | − | − | + | − |
| *Staphylococcus aureus MW2* | + | − | + | + | − | − |
| *Streptococcus pneumoniae R6* | + | − | + | − | − | − |
| *Streptococcus pneumoniae TIGR4* | + | − | + | − | − | − |
| *Streptococcus pyogenes MGAS315* | + | − | + | − | − | − |
| *Streptococcus pyogenes MGAS8232* | + | − | + | − | − | − |
| *Thermoanaerobacter tengcongensis* | + | + | − | − | − | − |
| *Ureaplasma urealyticum* | − | − | − | − | − | − |
| Proteobacteria | | | | | | |
| *Agrobacterium tumefaciens str. C58* | + | − | + | − | + | − |
| *Brucella melitensis* | + | − | + | − | − | − |
| *Buchnera aphidicola str.* Sg | + | − | + | − | − | − |
| *Buchnera sp. APS* | + | − | + | − | − | − |
| *Campylobacter jejuni* | + | − | + | − | − | − |
| *Caulobacter crescentus CB15* | + | + | + | − | + | − |
| *Escherichia coli K12* | + | − | + | − | + | − |
| *Escherichia coli O157:H7* | + | − | + | − | + | − |
| *Haemophilus influenzae Rd* | + | − | + | − | − | − |
| *Helicobacter pylori 26695* | + | − | + | − | − | − |
| *Helicobacter pylori J99* | + | − | + | − | − | − |
| *Mesorhizobium loti* | + | − | + | − | + | − |
| *Neisseria meningitidis MC58* | + | − | + | − | − | − |
| *Neisseria meningitidis Z2491* | + | − | + | − | − | − |
| *Pasteurella multocida* | + | − | + | − | − | − |
| *Pseudomonas aeruginosa* | + | − | + | − | + | − |
| *Ralstonia solanacearum* | + | − | + | − | +++ | − |
| *Rickettsia conorii* | + | − | − | − | − | − |
| *Rickettsia prowazekii* | + | − | − | − | − | − |
| *Salmonella enterica subsp. Enterica* | + | − | + | − | + | − |
| *Salmonella typhimurium LT2* | + | − | + | − | + | − |
| *Sinorhizobium meliloti* | + | − | + | − | + | − |
| *Vibrio cholerae* | + | − | + | − | − | − |
| *Xanthomonas axonopodis str. 306* | + | − | + | − | + | − |
| *Xanthomonas campestris* | + | − | + | − | + | − |
| *Xylella fastidiosa 9a5c* | + | − | + | − | + | − |
| *Yersinia pestis* | + | − | + | − | + | − |
| Spirochaetales | | | | | | |
| *Borrelia burgdorferi* | + | − | − | − | − | − |
| *Treponema pallidum* | + | − | + | − | − | − |

**Table 1**
**Continued**

| | Nth | OggI | MutY | MpgII | AlkA | OggII |
|---|---|---|---|---|---|---|
| Others | | | | | | |
| *Aquifex aeolicus* | ++ | − | − | + | − | − |
| *Chlorobium tepidum TLS* | + | + | + | − | − | − |
| *Deinococcus radiodurans* | + | ++ | + | − | + | − |
| *Fusobacterium nucleatum ATCC 25586* | + | − | − | − | − | − |
| *Thermotoga maritima* | + | − | − | + | − | − |

NOTE.—The distribution of bacterial HhH glycosylase genes into the six HhH families is shown. (+++) indicates the presence of three sequences, (++) the presence of two sequences, (+) the presence of one sequence, and (–) the presence of zero sequences.

(A10, A11), whereas the other was placed in an Nth family subclade that was otherwise largely composed of bacterial sequences (A3, A4) (see Supplementary Material online, fig. 1).

## The OggI Family

Among the HhH sequences placed in the OggI gene family were the widely studied yeast and mammalian Ogg glycosylases that excise 8-oxoguanine residues from DNA (Bruner et al. 1996; Rosenquist, Zharkov, and Grollman 1996; Lu et al. 2001). Compared to the Nth gene family, a much smaller fraction of the genomes assayed contained sequences that were placed in the OggI family (tables 1–3; see also Supplementary Material online, fig. 2). The majority of eukaryotic genomes (11/14) contained a member of the OggI family; however, the occurrence of OggI family homologs was somewhat reduced in archaea (9/16) and greatly reduced in bacteria (6/64). All eukaryotic genomes that harbored OggI family sequences contained only a single homolog, whereas two OggI homologs were found in one bacterial genome (*Deinococcus radiodurans*—Tree IDs = B62, B63) and one archaeal genome (*Pyrobaculum aerophilum*—Tree IDs = A16, A20). The two *P. aerophilum* OggI sequences were placed in disparate positions in the OggI gene tree: A16 was grouped in a clade with four other archaeal sequences (indicated by a star in figure 2 of the Supplementary Material online), and A20 was placed in a second subclade containing the majority (20/28 total) of OggI sequences. All of the glycosylases in the OggI subclade (indicated by the star in figure 2 of the Supplementary Material online) came from the five unique genomes that did not contain a homolog in the Nth family (table 3). The two *D. radiodurans* OggI sequences were placed between the two aforementioned OggI family subclades along with a sequence from the *Caulobacter crescentus* genome.

## The MutY/Mig Family

The HhH homologs placed in the MutY/Mig gene family included the well-studied *E. coli* and human MutY glycosylases (remove adenines mispaired opposite 8-oxoguanines—Demple and Harrison 1994) and the more recently characterized archaeal Mig glycosylases (remove thymines when paired opposite guanines—Yang et al. 2000). No genomes were found to harbor >1 HhH glycosylase genes in the MutY/Mig family. Most of the bacterial genomes surveyed (49/64) harbored a single HhH homolog in this gene family. A much smaller proportion of eukaryotic (7/14) and archaeal (5/16) genomes contained MutY/Mig family glycosylases. The two well-characterized Mig glycosylases from the genomes of *Aeropyrum pernix* (Tree ID = A24) and *P. aerophilum* (Tree ID = A25) were grouped in an exclusive basal subclade in the MutY/Mig gene family (see Supplementary Material online, fig. 3).

## The AlkA and MpgII Families

The AlkA and MpgII gene families each contained glycosylases that excise methyl-damaged bases (Begley et al. 1999). AlkA gene family homologs were found in roughly one-third of the bacteria (18/64) and eukaryotes (5/14), whereas over half of the archaeal genomes (9/16) contained at least one AlkA family gene. Three eukaryotic genomes (*A. thaliana*, *Oryza sativa*, and *Schizosaccharomyces pombe*) were each found to harbor two AlkA homologs (table 2). The phylogenetic analysis suggested that two gene duplication events, one specific to *S. pombe* and the other along the *A. thaliana*/*O. sativa* branch, were responsible for the three sets of duplicated genes (see Supplementary Material online, fig. 4). The single *S. cerevisiae* (Tree ID = E35) and *Candida albicans* (Tree ID = E36) AlkA sequences were not placed in the clade with the duplicated *A. thaliana* (Tree IDs = E39, E41), *O. sativa* (Tree IDs = E40, E42), and *S. pombe* (Tree IDs = E37, E38) AlkA sequences. Two AlkA homologs were found in each of the *Bacillus anthracis* (Tree IDs = B124, B135) and *Bacillus subtilis* (Tree IDs = B125, B136) genomes. The two identified as B124 and B125 were grouped together, and B135 and B136 were also placed together (see Supplementary Material online, fig. 4). The genome of *Ralstonia solanacearum* had three AlkA homologs (Tree IDs = B128, B137, B143) that were scattered across the gene tree (see Supplementary Material online, fig. 4). An AlkA homolog was detected in the genome of *Mycoplasma pulmonis* (Tree ID = B126) and was the only HhH homolog found in any of the five genomes surveyed from *Mollicutes* (table 1). Only a few bacterial (4/64) and about one-third of the archaeal (6/16) genomes contained genes in the MpgII family, and no eukaryotic members of this gene family were found. All MpgII homologs were from thermophilic archaeal or bacterial species, with the exception of the two *Staphylococcus aureus* strains (see Supplementary Material online, fig. 4).

**Table 2**
**HhH Superfamily Glycosylases in Eukaryotes**

| | Nth | OggI | MutY | MpgII | AlkA | OggII |
|---|---|---|---|---|---|---|
| *Anopheles gambiae* | + | + | − | − | − | − |
| *Arabidopsis thaliana* | ++ | + | + | − | ++ | − |
| *Caenorhabditis elegans* | + | − | − | − | − | − |
| *Caenorhabditis briggsae* | + | − | − | − | − | − |
| *Candida albicans* | + | + | − | − | + | − |
| *Drosophila melanogaster* | + | + | − | − | − | − |
| *Encephalitozoon cuniculi* | + | + | − | − | − | − |
| *Fugu rubripes* | + | + | + | − | − | − |
| *Homo sapiens* | + | + | + | − | − | − |
| *Mus musculus* | + | + | + | − | − | − |
| *Oryza sativa* | + | + | + | − | ++ | − |
| *Plasmodium falciparum* | + | + | + | − | − | − |
| *Saccharomyces cerevisiae* | ++ | + | − | − | + | − |
| *Schizosaccharomyces pombe* | + | − | + | − | ++ | − |

NOTE.—The distribution of eukaryotic HhH glycosylase genes into the six HhH families is shown. (++) indicates the presence of two sequences, (+) the presence of one sequence, and (–) the presence of zero sequences.

**Table 3**
**HhH Superfamily Glycosylases in Archaea**

| | Nth | OggI | MutY | MpgII | AlkA | OggII |
|---|---|---|---|---|---|---|
| Crenarchaeota | | | | | | |
| *Aeropyrum pernix* | − | + | + | − | + | − |
| *Pyrobaculum aerophilum* | − | ++ | + | − | − | − |
| *Sulfolobus solfataricus* | − | + | − | − | + | + |
| *Sulfolobus tokodaii* | − | + | − | − | ++ | + |
| Euryarchaeota | | | | | | |
| *Archaeoglobus fulgidus* | + | − | − | − | + | + |
| *Halobacterium sp.* | + | + | + | − | + | − |
| *Methanobacterium thermoautotrophicum* | + | + | + | + | − | − |
| *Methanocaldococcus jannaschii* | + | − | − | + | − | + |
| *Methanopyrus kandleri* | − | + | − | − | − | + |
| *Methanosarcina activorans* | ++ | + | − | + | + | − |
| *Methanosarcina mazei* | ++ | + | + | + | − | − |
| *Pyrococcus abyssi* | + | − | − | − | + | − |
| *Pyrococcus furiosus* | + | − | − | − | + | − |
| *Pyrococcus horikoshii* | + | − | − | − | + | − |
| *Thermoplasma acidophilum* | + | − | − | + | − | − |
| *Thermoplasma volcanium* | + | − | − | + | − | − |

NOTE.—The distribution of archaeal HhH glycosylase genes into the six HhH families shown. (++) indicates the presence of two sequences, (+) the presence of one sequence, and (–) the presence of zero sequences.

## The OggII Family

Only five HhH glycoyslase sequences were placed in the divergent OggII gene family, and all were from archaeal genomes (table 3; see also Supplementary Material online, fig. 5). The OggII glycosylase from *M. jannaschii* (Tree ID = A48), the only member of this gene family that has been biochemically characterized (Gogos et al. 2000), was shown to excise 8-oxoguanine residues, as do members of the OggI gene family. Principal component analyses of the *Archaeoglobus fulgidus* OggII homolog (Tree ID = A45) suggest that it also likely excises 8-oxoguanine residues (Gogos et al. 2000). The sequences in this family were only found when using PSI-Blast searches, whereas glycosylase sequences from the other five HhH gene families were all found in our initial BlastP and TBlastN searches, consistent with the outgroup placement of the OggII gene family relative to the other five HhH gene families (fig. 1).

## Discussion

This study provides a broad-based phylogenetic framework for the HhH superfamily of BER DNA glycosylases. The phylogeny presented here is in general agreement with two previous smaller-scale analyses of HhH superfamily evolution (Gogos et al. 2000; Yang et al. 2000), although neither of the previous studies included sequences from all of the gene families identified in this study. Six major HhH gene families (Nth, OggI, MutY/Mig, AlkA, MpgII, and OggII) (fig. 1) are identified, and sequences from all three domains of life are represented in four of the families. This observation suggests that the HhH superfamily diversified very early in evolution, prior to the divergence of the three domains of life. Three HhH gene families (MutY/Mig, OggI, and OggII) contain glycosylases that either directly excise 8-oxoguanines or correct base pairing mismatches associated with this type of damage. The AlkA and MpgII gene families contain enzymes that are involved in removing methyl-damaged bases, and the Nth family glycosylases repair oxidatively damaged pyrimidines. On average, archaeal and eukaryotic genomes contain more HhH glycosylase genes (archaea average 3.1 homologs/genome; eukaryotes, 3.0 homologs/genome) compared to bacteria (2.2 homologs/genome). However, the HhH glycosylase gene content of eukaryotic and archaeal genomes differs substantially, as homologs from all six of the gene families are found in archaeal genomes whereas eukaryotic homologs are limited to the Nth, OggI, MutY/Mig, and AlkA families (tables 2 and 3). The elevated number of HhH glycosylase genes in eukaryotic genomes is due primarily to lineage-specific gene duplication events in the Nth and AlkA gene families.

Extensive variation in the total number of HhH superfamily genes encoded by different genomes is observed (tables 1–3). In eukaryotes, for instance, each of the two nematode genomes surveyed from the *Caenorhabditis* genus harbors only one HhH superfamily homolog (both in the Nth family). Alternatively, the genome of *A. thaliana* contains six HhH glycosylase genes distributed in all four of the gene families where eukaryotic sequences are found (table 2). With the exception of *M. pulmonis*, all of the bacterial genomes surveyed from *Mollicutes* are completely devoid of HhH homologs, whereas the genome of *D. radiodurans* has five (table 1). Archaeal genomes contain from two to five HhH glycosylase genes (table 3).

The presence of "extra" glycosylase genes may reflect fundamental differences in DNA damage tolerance levels, particularly in the cases such as the radiation-resistant bacterial species *D. radiodurans* (Saffary et al. 2002). The genome of *D. radiodurans* contains a much greater number and variety of DNA repair genes than other bacterial genomes, including the five HhH homologs analyzed here (White et al. 1999; Makarova et al. 2002). The *D. radiodurans* genome contains two HhH glycosylase genes in the OggI family, whereas virtually all of the other bacterial genomes surveyed lack OggI homologs (table 1). This observation is in stark contrast to recent suggestions that *D. radiodurans* encodes a typical bacterial repertoire of DNA repair enzymes (Levin-Zaidman et al. 2003). The genome of *Ralstonia solanacearum*, a bacterial plant pathogen, harbors three distinct genes in the AlkA family, the only instance of any genome containing >2 sequences in any single gene family (table 1; see also Supplementary Material online, fig. 4). Some of these AlkA homologs may have been acquired through horizontal transfer, as the genome of *R. solanacearum* has a highly mosaic structure indicative of horizontal transfer playing a major role in its evolution (Salanoubat et al. 2002). No previous studies suggest that *R. solanacearum* is exceptionally resistant to DNA alkylating agents (such as ethanemethylsulfonate); nor do they provide any other biological explanations for the presence of three AlkA homologs in its genome.

In eukaryotes, the presence of additional HhH glycosylase genes is due in large part to multiple instances of lineage-specific gene duplications in the Nth and AlkA gene families. Differential sorting of Nth paralogs to the nucleus and mitochondria in eukaryotes has been documented to occur by both subfunctionalization of duplicate genes in yeast (Alseth et al. 1999) and by alternative splicing in mammals (Takao et al. 1998). The *S. cerevisiae* and *A. thaliana* genomes each contain lineage-specific gene duplicates in the Nth gene family (see Supplementary Material online, fig. 1). For *S. cerevisiae*, the protein product of one paralog (Ntg2, TREE ID = E2) is sorted exclusively to the nucleus, whereas the other (Ntg1, TREE ID = E3) is targeted to both the nucleus and the mitochondrion (Alseth et al. 1999). Knowledge of the subcellular sorting of the two *A. thaliana* paralogs would be particularly exciting, as plants have three subcellular compartments that contain DNA (nucleus, mitochondria, and chloroplasts) to which the protein products of different glycosylase paralogs may be targeted. Analyses of the *A. thaliana* Nth family paralog leader sequences using PSort, a subcellular targeting prediction computer program (Nakai and Kanehisa 1992), suggest that one duplicate is likely targeted to the nucleus, whereas the other is likely sorted to mitochondria. Eukaryotic lineage-specific duplications are also observed in the AlkA family. In *S. pombe*, PSort suggests that one duplicate is likely targeted to the nucleus, whereas the other is predicted to be cytosolic, suggesting that the latter paralog may no longer be involved in BER. Alternatively, one of the *A. thaliana* AlkA family paralogs is predicted by PSort to be mitochondrial; the other sorted to the nucleus. Although these computer-based predictions offer exciting possibilities, empirical immunolocalization studies of these *A. thaliana* and *S. pombe* glycosylases are required to gain an accurate understanding of their subcellular targeting patterns.

In contrast to the above instances, four bacterial genomes, all from *Mollicutes*, were found to harbor no HhH homologs; however, these bacterial species have undergone radical genome contraction (sizes ranging from ~0.5 to 1.4 MB) associated with their obligate intracellular parasitic lifestyles (Rocha and Blanchard 2002). The single HhH homolog found in the *M. pulmonis* genome is predicted to encode a protein with an N-terminus that contains a HhH DNA binding domain (placed in the AlkA family) and a C-terminal region that is homologous to $O^6$-methylguanine methyltransferases, suggesting that this is either a hybrid protein or that this single predicted open reading frame was incorrectly annotated and is actually two overlapping genes (frequently observed in mycoplasmal genomes [Fukuda, Washio, and Tomita 1999]). The genomes of other bacterial and eukaryotic species (such as *Rickettsia conorii* and *Caenorhabditis elegans*) are found to harbor only a single homolog in the entire HhH superfamily (all in the Nth family). How are these species able to deal with the diverse spectrum of base damage that assaults their genomes? One possibility is that a greater fraction of damaged bases simply goes unrepaired in species with fewer glycosylase genes. A second possibility is that the single HhH glycosylases of these species are able to recognize a broader spectrum of damage than the orthologous glycosylases found in species with genomes that encode multiple HhH homologs. Third, damaged bases in the genomes of these species may be repaired by BER glycosylases outside of the HhH superfamily and/or completely separate DNA repair pathways.

Although the first and second possibilities cannot be ruled out, there is extensive empirical support for the third scenario. Glycosylases outside of the HhH superfamily such as formamidopyrimidine glycosylase (Fpg) and alkyladenine glycosylase (Aag, also called MpgI) have been shown to overlap in the spectra of damage recognized by Ogg and AlkA glycosylases, respectively (Demple and Harrison 1994; Lau et al. 2000). Homologs of Fpg and Aag, however, are not found in the *C. elegans* or *C. briggsae* genomes with Blast searches (data not shown), suggesting that completely separate DNA repair pathways may play a major role in the repair of oxidatively damaged purines and methyl-damaged bases in these nematode species. This possibility is supported by studies in yeast that show overlap in the types of damage recognized by BER and nucleotide excision repair pathways, such as 8-oxoguanine (Swanson et al. 1999). Mismatch repair is another likely candidate for overlap with BER, as many common base-pairing mismatches are repaired by mismatch repair systems and BER glycosylases such as MutY and Mig (Jiricny 1998; Yang et al. 2000). Repair of alkyl-damaged bases can also occur by BER-independent mechanisms, such as with direct repair by $O^6$-methylguanine methytransferases (Dolan, Moschel, and Pegg 1990).

Multiple thermophile-specific clades of HhH glycosylases have been detected in this study and add to the existing knowledge of DNA repair proteins and pathways unique to thermophilic species (Begley et al. 1999; Makarova et al. 2002). The divergent OggII gene family contains only five sequences that are all from the genomes of thermophilic species of archaea (table 3; see also Supplementary Material online, fig. 5). The MpgII gene family contains archaeal and bacterial sequences that are almost exclusively found in the genomes of thermophiles (the two *S. aureus* genomes surveyed being the exceptions). Another distinctive group of glycosylases unique to thermophilic archaea is the OggI family subclade (indicated by a star in figure 2 of the Supplementary Material online). Curiously, the five archaeal genomes that contain sequences in this OggI subclade (which includes all four genomes from *Crenarchaeota*) also happen to be the only genomes found to lack a gene in the Nth family (excluding genomes from *Mollicutes*). This observation suggests that these OggI homologs may compensate for the lack of Nth homologs in their genomes. Biochemical analyses of the substrate specificities of these OggI glycosylases are required to provide insight into this possibility.

The phylogenetic analysis of the HhH superfamily of BER DNA glycosylases presented here provides an evolutionary framework for an important group of DNA repair enzymes whose diversity and adaptability have long been appreciated. We observed tremendous diversity in the numbers and types of HhH superfamily glycosylases encoded by different genomes across the three domains of life. This observation may reflect major disparities in the mechanisms and pathways by which different species repair oxidatively damaged and methyl-damaged bases. This widespread variation may also suggest significant differences in the underlying spontaneous mutation spectra experienced by different genomes.

## Acknowledgments

Funding for this work was provided by National Institutes of Health grant R01 GM-36827 to M.L. and grant F32 GM-66652 to D.D. We thank Jonathan A. Eisen at The Institute for Genomic Research, Jeffrey D. Palmer at Indiana University, and two anonymous reviewers for helpful comments.

## Literature Cited

Allen, J. F., and J. A. Raven. 1996. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. J. Mol. Evol. **42**:482–492.

Alseth, I., L. Eide, M. Pirovano, T. Rognes, E. Seeberg, and M. Bjørås. 1999. The *Saccharomyces cerevisiae* homologues of endonuclease III from *Escherichia coli*, and Ntg2, are both required for efficient repair of spontaneous and induced oxidative DNA damage in yeast. Mol. Cell Biol. **19**:3779–3787.

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

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Backlund, M. G., S. L. Trasti, D. C. Backlund, V. L. Cressman, V. Godfrey, and B. H. Koller. 2001. Impact of ionizing radiation and genetic background on mammary tumorigenesis in p53-deficient mice. Cancer Res. **61**:6577–6582.

Begley, T. J., B. J. Haas, J. Noel, A. Shekhtman, W. A. Williams, and R. P. Cunningham. 1999. A new member of the endonuclease III family of DNA repair enzymes that removes methylated purines from DNA. Curr. Biol. **9**:653–656.

Bruner, S. D., H. W. Nash, W. S. Lane, and G. L. Verdine. 1996. Repair of oxidatively damaged guanine in *Saccharomyces cerevisiae* by an alternative pathway. Curr. Biol. **8**: 393–403.

Culligan, K. M., G. Meyer-Gauen, J. Lyons-Weiler, and J. B. Hays, 2000. Evolutionary origin, diversification and specialization of eukaryotic MutS homolog mismatch repair proteins. Nucleic Acids Res. **28**:463–471.

Demple, B., and L. Harrison. 1994. Repair of oxidative damage to DNA: enzymology and biology. Annu. Rev. Biochem. **63**:915–948.

Demple, B., and S. Linn. 1980. DNA N-glycosylases and UV repair. Nature **287**:203–208.

Dolan, M. E., R. C. Moschel, and A. E. Pegg. 1990. Depletion of mammalian $O^6$-alkylguanine-DNA alkyltransferase activity by $O^6$-benzylguanine provides a means to evaluate the role of this protein in protection against carcinogenic and therapeutic alkylating agents. Proc. Natl. Acad. Sci. USA **87**: 5368–5372.

Eide, L., M. Bjørås, M. Pirovano, I. Alseth, K. G. Berdal, and E. Seeberg. 1996. Base excision of oxidative purine and pyrimidine DNA damage in *Saccharomyces cerevisiae* by a DNA glycosylase with sequence similarity to endonculease III from *Escherichia coli*. Proc. Natl. Acad. Sci. USA **93**:10735–10740.

Eisen, J. A. 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J. Mol. Evol. **41**:1105–1123.

———. 1996. A phylogenomic study of the MutS family of proteins. Nucleic Acids Res. **26**:4291–4300.

Eisen, J. A., and P. C. Hanawalt. 1999. A phylogenomic study of DNA repair genes, proteins, and processes. Mutat. Res. **435**:171–213.

Epel, D. 2003. Protection of DNA during early development: adaptations and evolutionary consequences. Evol. Dev. **5**: 83–88.

Fondufe-Mittendorf, Y. N., C. Harer, W. Kramer, and H. J. Fritz. 2002. Two amino acid replacements change the substrate preference of DNA mismatch glycosylase Mig/MthI from T/G to A/G. Nucleic Acids. Res. **30**:614–621.

Fukuda, Y., T. Washio, and M. Tomita. 1999. Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. Nucleic Acids Res. **27**:1847–1853.

Gogos, A., D. Jantz, S. Senturker, D. Richardson, M. Dizdaroglu, and N. D. Clarke. 2000. Assignment of enzyme substrate specificity by principal component analysis of aligned protein sequences: an experimental test using DNA glycosylase homologs. Proteins **40**:98–105.

Hashiguchi, K., Q. M. Zhang, H. Sugiyama, S. Ikeda, and S. Yonei. 2002. Characterization of 2-hydroxyadenine DNA glycosylase activity of *Escherichia coli* MutY protein. Int. J. Radiat. Biol. **78**:585–592.

Higgins, D. G., J. D. Thompson, and T. J. Gibson, 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266**:383–402.

Jiricny, J. 1998. Eukaryotic mismatch repair: an update. Mutat. Res. **409**:107–121.

Kanai, S., R. Kikuno, H. Toh, H. Ryo, and T. Todo. 1997. Molecular evolution of the photolyase-blue-light photoreceptor family. J. Mol. Evol. **45**:535–548.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: Molecular Evolutionary Genetics Analysis Software. Bioinformatics **17**:1244–1245.

Lau, A. Y., M. D. Wyatt, B. J. Glassner, L. D. Samson, and T. Ellenberger. 2000. Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG. Proc. Natl. Acad. Sci. USA **97**:13573–13578.

Levin-Zaidman, S., J. Englander, E. Shimoni, A. K. Sharma, K. W. Minton, and A. Minsky. 2003. Ringlike structure of the Deinococcus radiodurans genome: a key to radioresistance? Science **299**:254–256.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA. Nature **262**:709–715.

Lu, A. L., X. Li, Y. Gu, P. M. Wright, and D. Y. Chang. 2001. Repair of oxidative DNA damage: mechanisms and functions. Cell Biochem. Biophys. **35**:141–170.

Makarova, K. S., L. Aravind, N. V. Grishin, I. B. Rogozin, and E. V. Koonin. 2002. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Res. **30**:482–496.

Martin, I. V., and S. A. MacNeill. 2002. ATP-dependent DNA ligases. Genome Biol. **3**:3005.1–3005.7.

Michaels, M. L., L. Pham, Y. Nghiem, C. Cruz, and J. H. Miller. 1990. MutY, an adenine glycosylase active on G-A mispairs, has homology to endonuclease III. Nucleic Acids Res. **18**:3841–3845.

Nakai, K., and M. Kanehisa. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics **14**:897–911.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, New York.

Nilsen, H., M. Otterlei, T. Haug, K. Solum, T. A. Nagelhus, F. Skorpen, and H. E. Krokan. 1997. Nuclear and mitochondrial uracil-DNA glycosylases are generated by alternative splicing and transcription from different positions in the UNG gene. Nucleic Acids Res. **25**:750–755.

Peterson, J. L., and G. D. Small. 2001. A gene required for the novel activation of a class II DNA photolyase in *Chlamydomonas*. Nucleic Acids Res. **29**:4472–4481.

Rocha, E. P., and A. Blanchard. 2002. Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. Nucleic Acids Res. **30**:2031–2042.

Roldan-Arjona, T., C. Anselmino, and T. Lindahl. 1996. Molecular cloning and functional analysis of a *Schizosaccharomyces pombe* homologue of *Escherichia coli* endonuclease III. Nucleic Acids Res. **24**:3307–3312.

Rosenquist, T. A., D. O. Zharkov, and A. P. Grollman. 1996. Cloning and characterization of a mammalian 8-oxoguanine DNA glycosylase. Proc. Natl. Acad. Sci. USA **94**:7429–7434.

Saffary, R., R. Nandakumar, D. Spencer, F. T. Robb, J. M. Davila, M. Swartz, L. Ofman, R. J. Thomas, and J. DiRuggiero. 2002. Microbial survival of space vacuum and extreme ultraviolet irradiation: strain isolation and analysis during a rocket flight. FEMS Microbiol. Lett. **215**:163.

Salanoubat, M., S. Genin, F. Artiguenave et al. (28 co-authors). 2002. Genome sequence of the plant pathogen *Ralstonia solanacearum*. Nature **415**:497–450.

Sitnikova, T., A. Rzhetsky, and M. Nei. 1995. Interior branch and bootstrap tests of phylogenetic trees. Mol. Biol. Evol. **12**: 319–333.

Swanson, R. L., N. J. Morey, P. W. Doetsch, and S. Jinks-Robertson. 1999. Overlapping specificities of base excision repair, nucleotide excision repair, recombination, and translesion synthesis pathways for DNA base damage in *Saccharomyces cerevisiae*. Mol. Cell Biol. **19**: 2929–2935.

Takao, M., H. Aburatani, K. Kobayashi, and A. Yasui. 1998. Mitochondrial targeting of human DNA glycosylases for repair of oxidative DNA damage. Nucleic Acids Res. **26**:2917–2922.

Weller, G. R., B. Kysela, R. Roy et al. (14 co-authors). 2002. Identification of a DNA nonhomologous end-joining complex in bacteria. Science **297**:1686–1689.

White, O., J. A. Eisen, J. F. Heidelberg et al. (25 co-authors). 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science **286**:1571–1577.

Yang, H., S. Fitz-Gibbon, E. M. Marcotte, J. H. Tai, E. C. Hyman, and J. H. Miller. 2000. Characterization of a thermostable DNA glycosylase specific for U/G and T/G mismatches from the hyperthermophilic archaeon *Pyrobaculum aerophilum*. J. Bacteriol. **182**:1272–1279.

Kenneth Wolfe, Associate Editor

Accepted May 19, 2003