Phylogenomic Analysis of the Uracil-DNA Glycosylase Superfamily

J. Ignacio Lucas-Lledó,*1 Rohan Maddamsetti,2 and Michael Lynch 1

1Department of Biology, Indiana University Bloomington
2Program in Ecology, Evolutionary Biology & Behavior, Michigan State University

*Corresponding author: E-mail: joslucas@indiana.edu.
Associate editor: Manolo Gouy

Abstract

The spontaneous deamination of cytosine produces uracil mispaired with guanine in DNA, which will produce a mutation, unless repaired. In all domains of life, uracil-DNA glycosylases (UDGs) are responsible for the elimination of uracil from DNA. Thus, UDGs contribute to the integrity of the genetic information and their loss results in mutator phenotypes. We are interested in understanding the role of UDG genes in the evolutionary variation of the rate and the spectrum of spontaneous mutations. To this end, we determined the presence or absence of the five main UDG families in more than 1,000 completely sequenced genomes and analyzed their patterns of gene loss and gain in eu bacterial lineages. We observe nonindependent patterns of gene loss and gain between UDG families in Eubacteria, suggesting extensive functional overlap in an evolutionary timescale. Given that UDGs prevent transitions at G:C sites, we expected the loss of UDG genes to bias the mutational spectrum toward a lower equilibrium G + C content. To test this hypothesis, we used phylogenetically independent contrasts to compare the G + C content at intergenic and 4-fold redundant sites between lineages where UDG genes have been lost and their sister clades. None of the main UDG families present in Eubacteria was associated with a higher G + C content at intergenic or 4-fold redundant sites. We discuss the reasons of this negative result and report several features of the evolution of the UDG superfamily with implications for their functional study.

Key words: uracil-DNA glycosylase, mutation rate evolution, mutational bias, GC content, DNA repair, mutator gene.

Introduction

Thanks to advances in high-throughput sequencing, we can estimate the spontaneous mutation rate of many eukaryotes with startling precision. Whole-genome sequencing of mutation accumulation lines has allowed the mutational spectrum for several model organisms to be characterized and compared in great detail (fig. 1; Denver et al. 2009; Keightley et al. 2009; Ossowski et al. 2010). Knowledge of the mutation rate to different base substitutions is critical to the detection of positive selection, the quantification of selective coefficients, and the estimation of effective population sizes.

Three main factors have been proposed to explain the diversity of mutation rates among lineages: generation time, metabolic rate, and quality of DNA-repair machinery (Baer et al. 2007). Despite the fact that the organisms represented in figure 1 have different numbers of germ-line cell divisions per generation (8.5 in Caenorhabditis elegans, Wilkins 1992; 36 in Drosophila melanogaster, Drost and Lee 1995; and ~50 in Arabidopsis thaliana, Hoffman et al. 2004), the mutation rates per generation of most kinds of base substitutions are remarkably similar among them, although significant interspecific differences exist for the mutation spectrum. Therefore, in addition to generation time, other factors such as the presence and quality of specific pathways of DNA modification and repair must be playing an important role in the diversity of mutation rates and spectra. For example, the relatively high rate of transitions at G:C sites in A. thaliana can be largely, if not completely, explained by the levels of cytosine methylation in this species (Ossowski et al. 2010).

In parallel to the estimation of spontaneous mutation rates in several species, others and we are following a phylogenomic approach, pioneered by Eisen and Hanawalt (1999), to characterize the DNA-repair machinery present in as many species as possible, one pathway at a time (Denver et al. 2003; Lin et al. 2007; Lucas-Lledó and Lynch 2009). One aim of this effort is to gain insight into the mechanisms underlying differences in mutation rates among species. Here, we focus on the uracil excision repair pathway, which removes uracils from DNA. Uracils appear in DNA from two different sources. During DNA replication, polymerases introduce uracil in front of adenine with a frequency similar to the ratio of the concentrations of dUTP and dTTP (Tye et al. 1978). Properly paired with adenine, uracil is not mutagenic, although it can disrupt the binding sites of transcription factors and other DNA-binding proteins (Ivarie 1987). Uracils can also appear in DNA, mispaired with guanine, by deamination of cytosine, which is one of the most frequent spontaneous mutagenic reactions (Lindahl and Nyberg 1974). Half the progeny of a DNA molecule with an U:G mismatch will carry a mutation. The enzymes that specifically target uracil and start the uracil excision repair pathway are called uracil-DNA glycosylases.
(UDGs). They catalyze the excision of uracil, producing an abasic site and free uracil.

We hypothesize that the mutation rate, the transition–transversion ratio, and the mutational bias toward A + T (Lynch 2010) should be increased in lineages where UDG activity has been evolutionarily lost. To test this hypothesis, we first performed a presence/absence analysis of the five main families of UDGs in completely sequenced genomes of Archaea, Eubacteria, and Eukarya, and then we tested the correlation between the presence or absence of each family with the G + C content in intergenic and 4-fold redundant sites of 779 eubacterial genomes. The rationale behind this test is that intergenic and 4-fold redundant sites may reflect mutational biases more faithfully than more functionally constrained sites. We also study the evolutionary interaction between the four UDG families widely spread among Eubacteria and show that, in a large temporal scale, almost any of them can compensate for the absence of the other.

The five UDG families considered in this study are usually numbered in order of discovery (Pearl 2000). Family 1 is present in Eubacteria, Eukarya and in some eukaryotic viruses. Members of family 1 are, by far, the most efficient UDGs (Slupphaug et al. 1995). Family 2 UDGs are also found in Eubacteria and Eukarya, where they seem to have complementary functions to those of family 1. For example, family 2 UDGs target uracil (and also thymine in Eukaryotes) only when mismatched with guanine (Neddermann and Jiricny 1994; Gallinari and Jiricny 1996), also target ethenocytosine (Saparbaev and Laval 1998), are upregulated during stationary phase in Escherichia coli (Mokkapati et al. 2001) and prefer CpG sites in human cells (Abu and Waters 2003). Family 3 was originally described in animals and characterized by the ability to excise 5-hydroxymethyluracil and uracil from single-stranded DNA (Haushalter et al. 1999; Boorstein et al. 2001), although family 3 UDGs were later shown to be active on double-stranded DNA and present in a few eubacterial species (Pettersen et al. 2007). Family 4 is present in Eubacteria, Archaea, and some phages. It is thermostable and functionally similar to family 1, as far as it has been characterized (Koulis et al. 1996; Sandigursky and Franklin 1999, 2000; Sandigursky et al. 2001). Family 5 is also present in Eubacteria and Archaea but absent in Eukarya. It is also thermostable. A family 5 UDG from Mycobacterium tuberculosis complements a strain of E. coli deficient in its family 1 (Srinath et al. 2007). The five families have a common evolutionary origin, although their homology is hardly recognized at the sequence level, even in the two most conserved motifs of the catalytic pocket (Aravind and Koonin 2000). Other enzymes with UDG activity have evolved in some lineages but have been excluded from this study, such as the Archaea-specific UDGs from the helix-hairpin-helix family (Chung et al. 2003), and the MBD4 homologs found in mammals (Hendrich et al. 1999). We also excluded viral UDGs from our analysis.

Materials and Methods

Phylogenies

Using queries from the five UDG families identified in the literature, we performed PSI-Blast searches (Altschul et al. 1997) against the nonredundant protein database in the National Center for Biotechnology Information (NCBI) server until convergence. Results for families 2, 4, and 5 overlapped, and they were combined in a single file. An alignment using Muscle (Edgar 2004) and a preliminary neighbor joining tree was built for these three families, and family 2 was then defined to include a clade (family 2a) with all proteins previously annotated as family 2 UDGs, and a sister clade (family 2b) of only prokaryotic sequences that did not have any functional annotation and that were closer to family 2 proteins than to either family 4 or 5. All the remaining proteins in the alignment, including families 4 and 5, and other more distantly related sequences, were kept together and labeled “extended family 4”. The resulting four alignments (family 1, family 2, family 3, and extended family 4) were manually curated with BioEdit (Hall 1999). We removed from the alignments the less reliable sites with Gblocks (Talavera and Castresana 2007), and we used ProtTest (Abascal et al. 2005) to determine the most likely models of amino acid evolution, which were then used to infer four maximum likelihood (ML) phylogenies with RAxML (Stamatakis 2006). An approximation to bootstrap sampling was used to determine the branch support (Stamatakis et al. 2008).

To analyze the patterns of evolutionary loss of UDG genes in Eubacteria, we needed a phylogenetic tree of the
eubacterial species with completely sequenced genomes. To this end, we downloaded the aligned sequences of their small ribosomal subunit from the RDP II (Cole et al. 2007) or the Greengenes databases (DeSantis et al. 2006). To merge sequences from the two databases, we aligned their profiles with muscle, and then we manually checked the alignment with BioEdit (Hall 1999). We also trimmed the ends and used GBLOCKS to purge from the alignment the unreliably aligned regions. The final alignment consisted of 832 sequences and was 1,437 nucleotides long. We ran 240 replicates of the search of the ML tree with RAxML, using the GTRMIX model of molecular evolution. We kept the 100 most likely trees to represent the uncertainty about the true phylogeny. We used Archaea as outgroup.

Protein Domain Analysis
We searched for additional protein domains from the Pfam database (Finn et al. 2006) in all full-length UDGs found in the PSI-Blast searches, using HMMER3 (Eddy 1998). We disregarded all domain predictions with an E value higher than 0.001.

Presence–Absence Data
Given the sensitivity of the PSI-Blast searches, we consider unlikely that any homologous protein with UDG activity present in the databases was missed. However, some UDGs could be missing from the protein databases if their genes were not annotated. To fill this gap, we undertook exhaustive TBLASTN searches of all UDG families in completely sequenced genomes with or without a definitive assembly. We scanned over 1,000 genomes through either the microbial genomic database or the whole-genome shotgun database at NCBI (www.ncbi.nlm.nih.gov), Last accessed December 8, 2010), or through the species-specific databases at the Broad Institute (www.broadinstitute.org, Last accessed December 8, 2010), the Genome Center at Washington University (genome.wustl.edu, Last accessed December 8, 2010), or the DOE Joint Genome Institute (www.jgi.doe.gov, Last accessed December 8, 2010), depending on where each genome was available. To search for a UDG family in a genome, we used as queries all the sequences of that UDG family already present in our alignment that belonged to the lowest taxonomic rank of that genome’s classification in which the presence of that protein family was already known.

A presence was tentatively called whenever at least one of the queries retrieved a hit with an E value lower than 0.001. Best hits with E values higher than 10^{-5} were manually checked, and they were determined to be false positives if any of the following were true: Homology was limited to a portion of the protein lower than 70%; an additional domain in the query, other than UDG, was causing the hit; or the best reciprocal hit on the “nr” protein database was not a UDG. In addition, we excluded several hits suspected to be caused by bacterial contamination of eukaryotic contigs. Given that all UDG families analyzed share a common origin (Aravind and Koonin 2000), we also corrected a few cases of cross-detection between families. To avoid false negatives, we excluded from the analysis several eukaryotic genomes supposed to be in an assembly stage but with sequencing coverage depths lower than 2.

Test of Correlated Evolution among UDG Families
We tested whether the probabilities of loss and gain of a UDG family were independent of the presence of other UDG families in Eubacteria. We used the program BayesTraits (Pagel and Meade 2006), which implements the comparison of dependent and independent models of evolution of pairs of binary traits developed by Pagel (1994). We ran Bayesian analyses with reversible jumps between models to compare dependent and independent models of evolution between families 1 and 2, 1 and 4, and 2 and 4. Because the presence of family 3 in Eubacteria is incidental (see below), we did not include it in the analyses. The use of more than one phylogeny to represent its uncertainty was not feasible due to extremely low rates of acceptance of parameter value change proposals in the Markov chains, resulting in a very inefficient exploration of the parameter space. We therefore used the most likely Eubacterial phylogeny inferred before, with branch lengths in units of substitutions per site. Because we expected both gains and losses of UDG genes to happen at a slower rate than the fixation of base substitutions, we used exponential priors of mean 1.0 for all the parameters. The Markov chains were sampled every 1,000 iterations, during 65 million iterations after a burn in period of 500,000 iterations. Under these settings, no correlation between successive samples of parameter values was observed.

Test of Correlation between Presence–Absence of UDG Families and G + C Content
We used a python script (R.M.) to scan all GenBank contigs composing the complete genomes analyzed. In those contigs with gene predictions, the script determined the G + C content in coding, intergenic, and 4-fold redundant sites, as well as the overall G + C content. Plasmids were excluded from the analysis. We used the aotf routine of the Phylocom 4.1 package (Webb et al. 2008) to test whether the presence of families 1, 2, or extended family 4 was correlated with the G + C content at 4-fold redundant sites among 779 eubacterial species. The best of the phylogenies found in the ML searches was used, with untransformed branch lengths, and the logarithmic transformation of the G + C proportion was used because it removed any correlation between the absolute value of the standardized contrasts and their standard deviations (Garland et al. 1992) as tested with Mesquite (Maddison WP and Maddison DR 2010; Midford et al. 2010). A first run of aotf with presence–absence data set as a binary trait determined what nodes of the tree were informative about the relationship between G + C content and presence–absence of UDG. A second run, calculated the standardized contrasts for all the nodes as suggested in the Phylocom manual. A t-test was performed with the standardized contrasts of the informative nodes.

The presence of any UDG family is expected to increase the mutation equilibrium G + C content. However, the G + C content at 4-fold redundant sites in Eubacteria is also
Phylogenomics of Uracil-DNA Glycosylases

FIG. 2. ML phylogenies of the main UDG families, and the sequence logos of their most conserved homologous motif. The branch color indicates eubacterial (black), archaeal (green), or eukaryotic (red) sequences. A putative family 6, not functionally described in the literature, is indicated with a question mark.

Results

Evolution of the UDG Families

Because they are present in all domains of life (fig. 2), UDGs must have appeared very early in evolution. Family 1 is the most conserved one. We failed to recover the likely monophyly of all eukaryotic family 1 UDGs with our ML phylogenetic analysis (fig. 2). Specifically, some highly divergent protozoan UDGs are grouped together with eubacterial UDGs with high bootstrap values, probably due to long branch attraction. Most family 1 UDGs are encoded by single-copy genes. The high degree of conservation of the catalytic motifs in this family suggests that all its members have the same function. Family 1 UDG domains are very rarely associated with other domains in the same protein. The main exceptions to this are the Actin domain in at least two fungal genera of the order Onygenales and the ART domain in three genera of Actinobacteria and Planctomycetes (supplementary table S2, Supplementary Material online). None of these domains are known to participate in DNA repair.

All proteins functionally characterized as family 2 UDGs are located in figure 2 in the clade labeled “2a.” Another clade, “2b,” is composed of proteins clearly more similar to family 2 UDGs than to any other UDG families, although none of them has been functionally characterized, to the best of our knowledge. Family 2 has been previously described only in Eubacteria and Eukarya, and we identify for the first time two Archaeal orthologs of subfamily 2a (accession numbers EET89853 and EEZ92451) and several others in subfamily 2b. Archaeal and eukaryotic sequences in subfamily 2b are scattered in the clade in positions that do not match their taxonomy. Although some of them may be artificial or due to lateral gene transfer, we consider it very likely that the last universal common ancestor had at least one and probably two family 2 UDGs.

The ML phylogeny of family 3 also reveals the existence of two main clades: the canonical family 3 UDGs, almost exclusively animal, and a eubacterial clade with some representatives from Bacteroidetes, Chlorobi, Actinobacteria, and Firmicutes. Among the animal family 3 UDGs, we found some nonanimal proteins: three from the moss Physcomitrella patens, and some from nonpathogenic eubacterial species, that belong to Proteobacteria, Planctomycetes, or Verrucomicrobia.

UDGs from families 4 and 5 can be reliably aligned together with many other sequences of uncertain function. Figure 2 depicts the ML phylogeny of all these proteins.
Family 5 and another clade of UDG-like proteins (putative family 6) are supported by bootstrap values of 66% and 69%, respectively. However, family 4 UDGs form a highly polytomic ensemble together with unannotated proteins, many of which do not conserve the catalytic motifs. To simplify, we considered families 4 and 5, putative family 6, and all other related proteins as part of the “extended family 4” in all subsequent analyses.

Two family 4 UDGs have been independently acquired by the eukaryotes *Micromonas* sp. (Chlorophyta) and *Paulinella chromatophora* (Rhizaria) through their primary photosynthetic organelles. The annotation of many family 4 UDGs as “Phage SPO1 polymerase-related protein” is an unfortunate consequence of the existence of a polymerase in *Bacillus* phage SPO1 fused to a UDG domain related to family 4. Interestingly, plasmids in *Yersinia pestis* and in *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18, but not in other serovars of the same subspecies, also encode a polymerase III alpha subunit fused to a family 4 UDG domain (supplementary table S1, Supplementary Material online).

**Comparative Functional Review**

All UDG families share specificity for uracil in U:G mismatches and probably for uracil paired with adenine (see supplementary table S3, Supplementary Material online). All families, and especially families 2 and 3, have additional substrates, which may have become their primary targets, such as ethenocytosine or T:G mismatches in many family 2 UDGs.

A convenient measure of catalytic efficiency, to establish comparisons across species, would be the ratio between the maximum reaction rate, *V*$_{\text{max}}$, and the Michaelis constant, *K*$_{\text{M}}$ (substrate concentration at which the reaction rate is half the maximum). Unfortunately, many studies do not report *V*$_{\text{max}}$ but only *K*$_{\text{M}}$ and the turnover, *k*$_{\text{cat}}$ (maximum number of enzymatic reactions catalyzed per unit of time). Figure 3 represents the distributions of several measures of *K*$_{\text{M}}$ and *k*$_{\text{cat}}$ for family 1 UDGs of four different species. Most experiments attribute a lower *K*$_{\text{M}}$ toward U:A (higher affinity) to *E. coli*’s family 1 UDG than to its human ortholog (*t*-test *P* value = 0.012). Supplementary table S4, Supplementary Material online, displays the catalytic properties of the family 1 UDGs represented in figure 3 and also of some members of families 2, 3, and 4.

**Correlated Rates of Gain and Loss of UDG Families**

The patchy taxonomic distribution of some UDG families suggested that UDG genes have been evolutionarily lost and gained several times in Eubacterial lineages. Therefore, we used the patterns of gene gain and loss to study the functional relationships among UDG families along a large
timescale. First, we combined the presence–absence data with a ML phylogeny of Eubacteria, built from the alignment of the 16S ribosomal subunit of 779 species with completely sequenced genomes, in which Archaea was used as an out-group. With this data set, we performed Bayesian analyses to fit two kinds of models of the evolution of the genomic UDG composition: Independent models in which the probabilities of gain and loss of a family are unaffected by the presence of another family; and dependent models, where the presence of one family changes the probability of losing or gaining another one. Each specific model is defined by the number of parameters used to represent the rates of loss and gain of each family and by what rates are set to be equal. Due to computational limitations, only pairwise comparisons among families can be performed with currently available software. Specifically, we performed all pairwise comparisons among families 1, 2, and the extended family 4. The Markov chains were allowed to jump between dependent and independent models to estimate the posterior probability of dependent evolution between families. We found strong evidence for the dependent evolution of families 1 and 2 (Bayes factor, or ratio between posterior and prior odds of dependent evolution, 52.4) and of family 1 and extended family 4 (Bayes factor > 157). In contrast, family 2 and the extended family 4 were shown to be lost or gained in Eubacterial lineages independently (Bayes factor, 0.01). The use of alternative, less likely, phylogenies gave similar results (data not shown).

Dependent evolution between UDG families imply that the presence of one family in a genome affects the chance of another family to be lost or gained. To determine exactly what rates are modified by the presence or absence of each family, we also compared the posterior probabilities of specific models of dependent evolution, taken from the frequency with which the Markov chain visited them, with their prior probabilities. Following Pagel and Meade (2006), we determined that the prior odds of any specific dependence is 4.11 because among all the models randomly sampled by the Markov chain, there are this many more models where a specific rate depends on the presence of the alternative family than models where the rate is equal under the presence or absence of the alternative family. Using this information, we determine that only two rates showed strong evidence of dependent evolution: The rate of loss of family 1 is dependent on the presence of family 2 (Bayes factor, 12.2), and the rate of loss of the extended family 4 is dependent on the presence of family 1 (Bayes factor > 15,800).

Figure 4 represents the posterior distributions of the rates of loss of family 1 and of extended family 4 in the presence and the absence of families 2 and 1, respectively. It can be
seen that family 1 and extended family 4 are very rarely lost in the absence of either family 2 or family 1, whereas the presence of family 2 increases the rate of loss of family 1, and the presence of family 1 increases the rate of loss of the extended family 4. To put it another way, the loss of family 2 enhances the preservation of family 1 (but not of family 4), and the loss of family 1 enhances the preservation of family 4.

**Effect of UDG Families in the Evolution of G + C Content**

The genomic G + C content evolves by mutation, drift, and natural selection. The removal of uracil from DNA by UDGs prevents G:C → A:T transitions, and therefore the presence of UDGs is expected not only to reduce the total mutation rate (see table 1) but also to balance the mutational spectrum in favor of a higher G + C composition. We hypothesized that lineages where a UDG family has been lost, the mutational spectrum will have changed in favor of a lower G + C content. To test this hypothesis, we calculated the G + C composition in intergenic and 4-fold redundant sites of 779 completely sequenced and annotated eubacterial genomes, and we used their ML phylogeny to calculate independent contrasts of G + C composition in all nodes. We then used the contrasts at nodes whose daughter branches differed in the presence of a UDG family to test the effect of the absence of that UDG family in the G + C composition. In all, 40 nodes were informative for family 1, 83 nodes for family 2, 18 nodes for family 3, and 55 for the extended family 4. We also combined the presence of family 1 and extended family 4 to test whether the absence of both of them affected the G + C composition, but only 9 nodes were informative because most eubacterial genomes have either one or the other. Within each family, the standardized independent contrasts at informative nodes represent the difference in G + C content between a lineage with that family and a lineage without it. Overall, a positive average was expected. None of the t-tests reported a significant departure from zero at the 0.05 level for the G + C content in 4-fold redundant sites (table 2). Sign tests were also performed, with similar results.

To avoid the confounding effect of codon usage bias, intergenic G + C content was also analyzed, using a suboptimal phylogeny that allowed the assumption of a Brownian model of evolution (see Methods). The t-tests yielded nonsignificant results for all families, indicating that intergenic G + C content is not different between species with and without a UDG family (data not shown). To account for the effect of optimal growth temperature and other factors affecting the genome-wide G + C content, we repeated the analysis using the residuals of the regression of intergenic G + C content on G + C content at coding sites. Again, none of the UDG families was associated with a significantly higher intergenic G + C content relative to the G + C content at coding sites.

**Discussion**

We now have the ability to detect small differences in the genomic mutation rate and the mutational spectra among species, and we can aim to explain those differences. One of the types of base substitutions contributing to the higher mutation rate in *D. melanogaster* relative to *C. elegans* are G:C → A:T transitions (fig. 1). It is tempting to hypothesize that the lack of the main UDG family in the former (supplementary table S1, Supplementary Material online) could be at least part of the explanation, although there are other types of substitution with significantly different rates between the two species.

We showed that the presence of none of the UDG families was associated with a significantly higher G + C content at intergenic or 4-fold redundant sites (table 2) in contrast to what was expected from the known effect of UDG enzymes on the mutational spectrum. There are several reasons for this negative result. First, G + C content is the net outcome of mutation and selection effects. Selection for optimal codon usage (Sharp et al. 2005) and the optimal growth temperature (Musto et al. 2006) affect the G + C content at 4-fold redundant sites and limit our ability to detect the mutational effects. Second, a change in the mutational spectrum, such as the one expected after the loss of a UDG family, does not affect the G + C composition immediately, but through potentially long periods of time, whereas the new mutation selection–drift equilibrium is reached. Therefore, recent losses of UDG genes may have had a strong effect in the mutation rate and spectrum, but little effect in the G + C composition of intergenic and

**Table 2.** Average of Independent Contrasts of (logarithmic) G + C Content Between Lineages with or without a UDG Family, Their Standard Deviation (SD), Sample Size (N), and P value of a t-Test.

<table>
<thead>
<tr>
<th>Family</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
<th>t-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.0197</td>
<td>0.5180</td>
<td>40</td>
<td>0.8114</td>
</tr>
<tr>
<td>2</td>
<td>0.0993</td>
<td>0.9717</td>
<td>83</td>
<td>0.3547</td>
</tr>
<tr>
<td>3</td>
<td>0.0973</td>
<td>0.2826</td>
<td>18</td>
<td>0.1625</td>
</tr>
<tr>
<td>4</td>
<td>0.0454</td>
<td>0.4206</td>
<td>55</td>
<td>0.4265</td>
</tr>
<tr>
<td>1 or 4</td>
<td>0.3491</td>
<td>0.5000</td>
<td>9</td>
<td>0.5999</td>
</tr>
</tbody>
</table>
4-fold redundant sites. Indeed, all complete UDG losses have happened relatively recently in lineages that have not diversified yet (supplementary table S1, Supplementary Material online) and may never diversify, unless they recover the ability to excise uracil. And third, given the scarcity of species where all UDG families have been lost, it is difficult to rule out the possibility of the loss of one UDG family being rapidly compensated by other UDG families still present in the lineage.

Despite the failure to detect a relationship between the absence of UDG families and the mutational spectrum of eubacterial lineages, we set the bases for more detailed studies in the future. We gathered important information about the functions and the evolution of the UDG superfamily. Our phylogenomic analysis underlines the essentiality of the UDG activity in all domains of life and suggests that the UDG superfamily may have been the main provider of UDG activity since the last universal common ancestor. The only obvious exception to this hypothesis are some families of Euryarchaeota, whose genomes are devoid of all UDG families. A different family of enzymes with UDG activity seems to have taken over the ancestral function of the UDG superfamily in those archaea (Chung et al. 2003). In addition, archaeal DNA polymerases have the ability to stall replication 4–6 nucleotides before the presence of a uracil residue (Greagg et al. 1999), providing them with another line of defense against uracil.

Overall, either family 1 or families 4 or 5 seem to be essential, given the few examples of species without them (supplementary table S1, Supplementary Material online). The vast majority of thermophiles have either family 4 or 5, rather than (or sometimes in addition to) family 1, although families 4 and 5 are not restricted by any means to thermophiles. Our analysis of rates of gene loss and gain also supports the idea of a functionally equivalent role of families 1 and 4 or 5 (fig. 4). Results from UDG-deficiency experiments where more than one family was sequentially knocked out (table 1) show that either family 1 (E. coli, and Mus musculus) or family 4 (Thermus thermophilus) act as the main uracil remover, whereas UDGs from families 2, 3, or 5 have a secondary role in mutation avoidance in the presence of the former. It is likely that the difference between the main and the secondary UDGs is the ability of the former to interact with the replication machinery and to function during DNA synthesis in a highly processive fashion (Kavli et al. 2002; Dionne and Bell 2005; Kiyonari et al. 2008). The fusion of UDG domains and polymerase domains in phages and plasmids (supplementary table S2, Supplementary Material online) also underlines the tendency of some UDGs to become associated with DNA replication.

Although all UDG families share a common origin (Aravind and Koonin 2000) and uracil as a common substrate (supplementary table S3, Supplementary Material online), it has been argued that the main biological role of some members of families 2 and 3 is different from the excision of uracil from DNA in species where family 1 is also present (Saparbaev and Laval 1998; Lutsenko and Bhagwat 1999; Masaoka et al. 2003). Although that might well be the case, our finding that a few species lost all UDGs but those from families 2 or 3 strongly suggests that these UDG families can become the main providers of UDG activity, at least over an evolutionary timescale. The finding that the presence of family 2 significantly increases the chance of evolutionarily loss of family 1 in Eubacteria (fig. 4) strongly supports the idea that family 2 UDGs can complement the absence of family 1.

It is sometimes assumed that the ability of an enzyme to repair a damaged base must be adaptive, even when that damage is very rare (O’Brien 2006). Thus, it was thought that the main biological role of family 3 UDGs was the removal 5-hydroxymethyluracil (Boorstein et al. 1989), before its UDG activity was discovered (Boorstein et al. 2001); or that ethenocytosine was the main substrate of the family 2 UDG MUG from E. coli (Saparbaev and Laval 1998), even though ethenocytosine is not detected in E. coli DNA (Mokkapati et al. 2001). Given the large number of substrates of some UDG families (supplementary table S3, Supplementary Material online), we consider it unlikely that the specificity for all of them has been actively acquired by positive selection. Instead, we consider that a broad substrate specificity may be the result of relaxed selection, on the bases of two likely assumptions: First, that most neutral mutations expand the substrate specificity, rather than narrowing it, and second, that the lower the frequency of a substrate, the less efficient natural selection is to promote the specificity for that substrate. Under this scenario, natural selection can promote and keep a narrow substrate specificity of a DNA repair enzyme only if the substrate is frequent enough in DNA. When more than one enzyme repairs the same damage, as in the case of the UDG superfamily, and one of them ends up repairing most of the damage, as in the case of families 1 (An et al. 2005) and 4 (Sakai et al. 2008), the rest of the enzymes may evolve a broader substrate specificity by random drift. Along this line of thinking, it is commonly thought that the reason why UDG enzymes able to recognize thymine can only act on double-stranded substrates is the strong negative selection against the acquisition of thymine-DNA glycosylase activity in single-stranded DNA or in properly paired contexts. In addition, it has been shown that single substitutions on the human family 1 UDG gene can produce cytosine- and thymine-DNA glycosylases with mutator phenotypes (Kavli et al. 1996).

Our functional review is consistent with the idea that species with larger effective population sizes, where natural selection is more efficient, should have more efficient DNA-repair enzymes, and lower spontaneous mutation rates, in consequence. Thus, family 1 UDGs seem to have higher affinity for its most frequent substrate, U/A (fig. 3), and to prevent more mutations (table 1) in E. coli than in mammals. However, the data available up to now are not fully conclusive due to difficulties in the comparison of catalytic efficiency among species. First, data are scarce for most species and very noisy in some of them. Experimental conditions, such as the saline concentration (Slupphaug et al. 1995; Masaoka et al. 2003) or the presence of Mg^{2+} and other cofactors (Kavli et al. 2002), affect the catalytic
properties of these enzymes and are not homogeneous across the reports (supplementary table S4, Supplementary Material online). UDGs from all families, including highly processive members of family 1, have been reported to be inhibited, to different extents, by their products, the abasic site, and the free uracil (Domena et al. 1988; Waters and Swann 1998; Sandigursky and Franklin 2000; Sung and Mosbaugh 2000; Sandigursky et al. 2001; Srinath et al. 2007; Nilsen et al. 2001; Kavli et al. 2002; Hardeland et al. 2003; O’Neill et al. 2003; Peña Díaz et al. 2004) (although, see as well Kavli et al. 2002; Sartori et al. 2002). This implies that measures of $K_M$ or $k_{cat}$ could be biased if the reaction times chosen to determine the amount of products and the enzyme and substrate concentrations were not chosen carefully (Waters and Swann 1998; Abu and Waters 2003).

Another problem with this comparative approach is that differences in the ability to process a substrate could be attributed to divergence in the preferred substrate, rather than to the efficacy of selection. This is likely to have haploid origin with diverse fates. Genome Biol. 1. tresearch0007.1–research0007.8.


