*Published online 26 October 2007* *Nucleic Acids Research, 2007, Vol. 35, No. 22* ***7591–7603***
*doi:10.1093/nar/gkm921*

# The origins and early evolution of DNA mismatch repair genes—multiple horizontal gene transfers and co-evolution

**Zhenguo Lin[1,2], Masatoshi Nei[1] and Hong Ma[1,2,*]**

[1]Department of Biology and Institute of Molecular Evolutionary Genetics and [2]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA

Received September 6, 2007; Revised October 3, 2007; Accepted October 8, 2007

## ABSTRACT

**To understand the evolutionary process of the DNA mismatch repair system, we conducted systematic phylogenetic analysis of its key components, the bacterial *MutS* and *MutL* genes and their eukaryotic homologs. Based on genome-wide homolog searches, we identified three new *MutS* subfamilies (*MutS3-5*) in addition to the previously studied *MutS1* and *MutS2* subfamilies. Detailed evolutionary analysis strongly suggests that frequent ancient horizontal gene transfer (HGT) occurred with both *MutS* and *MutL* genes from bacteria to eukaryotes and/or archaea. Our results further imply that the origins of mismatch repair system in eukaryotes and archaea are largely attributed to ancient HGT from bacteria instead of vertical evolution. Specifically, the eukaryotic *MutS* and *MutL* homologs likely originated from endosymbiotic ancestors of mitochondria or chloroplasts, indicating that not only archaea, but also bacteria are important sources of eukaryotic DNA metabolic genes. The archaeal *MutS1* and *MutL* homologs were also acquired from bacteria simultaneously through HGT. Moreover, the distribution and evolution profiles of the *MutS1* and *MutL* genes suggest that they have undergone long-term coevolution. Our work presents an overall portrait of the evolution of these important genes in DNA metabolism and also provides further understanding about the early evolution of cellular organisms.**

## INTRODUCTION

Mismatched nucleotides are regularly introduced by DNA polymerase during cell division and uncorrected nucleotides will result in mutations. In most cellular organisms, such replication errors are repaired mainly by the DNA mismatch repair (MMR) system that enhances replication fidelity 50- to 1000-folds by repairing mismatched nucleotides, and small insertions and deletions (1–3). The MMR system also prevents recombination between divergent sequences and repairs mismatches on heteroduplex DNA that arise during homologous recombination (4). Therefore, defects in the MMR could lead to highly elevated mutation rates, meiotic defects and infertility (5,6). The function of the MMR system has been thoroughly studied in some model organisms. In *Escherichia coli*, MMR is initiated when the MutS homodimer proteins bind to mismatched nucleotides on the daughter strand and forms a MutS–DNA complex (1,3). The MutS–DNA complex then interacts with the MutL homodimer proteins in an ATP-dependent manner. The interaction between the MutS and MutL complexes activates the endonuclease MutH to cleave the newly synthesized strand and initiates subsequent DNA repair events, including excision of the incorrect nucleotides and incorporation of the correct nucleotides (1,3).

Homologs of the *E. coli MutS* have been identified in many bacterial species (7,8). To avoid confusion with other *MutS*-like genes, here we call them the *MutS1* genes. A second *MutS* homolog, *MutS2*, is also present in many bacterial species (7), but they are functionally different from *MutS1* genes (9–12). In eukaryotes, up to seven different *MutS* homologs have been identified and designated as *MSH1 (MutS Homolog 1)* to *MSH7* (Table 1). These *MSH* genes play different roles in MMR as well as meiotic recombination (13–17). In contrast, only limited information is available about *MutS* homologs in archaea (18,19). Like the *MutS* gene family, the *MutL* homologs are also present in most bacterial species and all eukaryotes examined (Table 1) (1,3).

MMR is crucial for maintaining replication fidelity and genome stability in both eukaryotes and prokaryotes. Therefore, it is of great interest to study the evolutionary history of the genes involved in this cellular process. Previous phylogenetic analyses of the *MutS* gene family

*To whom correspondence should be addressed. Tel: 1 814 863 6414; Fax: 1 814 863 1357; Email: hxm16@psu.edu

**Table 1.** Known and newly identified *MutS* and *MutL* homologs

| | Bacteria | Archaea | Eukaryotes | | |
|---|---|---|---|---|---|
| | | | Animals | Plants | Fungi |
| MutS family genes | MutS1 (MutS) | MutS1 (MutS) | | | MSH1 |
| | | | MSH2 | MSH2 | MSH2 |
| | | | MSH3 | MSH3 | MSH3 |
| | | | MSH4 | MSH4 | MSH4 |
| | | | MSH5 | MSH5 | MSH5 |
| | | | MSH6 | MSH6 | MSH6 |
| | | | | MSH7 | |
| | MutS2 | | | MutS2 | |
| | MutS3 | | | | |
| | MutS4 | MutS4 | | | |
| | | MutS5 | | | |
| MutL family genes | MutL | MutL (MutL?) | MLH1 | MLH1 | MLH1 |
| | | | MLH2 (PMS1) | | MLH2 |
| | | | MLH3 | MLH3 | MLH3 |
| | | | MLH4 (PMS2) | MLH4 (PMS1) | MLH4 (PMS1) |

The gene names without parentheses are used in this study. Their corresponding old names, if different, are shown in parentheses. The newly identified genes are depicted with an underline.

indicated that the *MutS* family can be divided into two subfamilies: *MutS1* and *MutS2* (7,8). However, it was not certain whether the eukaryotic *MSH4* and *MSH5* genes are members of the *MutS1* or *MutS2* subfamilies (7,8). Therefore, the evolutionary relationships of *MutS* homologs are still unclear. In addition, *MutS* homologs in archaea and their evolutionary relationships with eukaryotic and bacterial counterparts have not been systematically studied. With regard to the *MutL* genes, although several preliminary phylogenetic trees have been presented (20–22), a detailed evolutionary analysis has not been reported. Therefore, it is necessary to conduct systematic analyses of these MMR genes. Taking advantage of the rapid expansion of sequence data, we searched for homologs of the two gene families from a much broader spectrum of species and systematically investigated their origins and evolutionary history in this study.

## MATERIALS AND METHODS

### Data mining

The *Bacillus subtilis* MutS (NP_389586) and MutL (NP_389587) protein sequences were used as queries to search for homologs against complete genome sequences of 461 bacterial and 39 archaeal species (25 May 2007 data) from the National Center for Biotechnology Information (NCBI) databases (23) by TBLASTN. All significant hits with an *e*-value $<e^{-10}$ were considered as potential MutS and MutL homologs. Domain structures of these potential MutS and MutL homologs were analyzed by searching the Pfam (24) and SMART (25) protein domain databases. Multiple sequence alignments of the MutS or MutL homologs, respectively, were generated for sequence comparison and identification of conserved regions by using MUSCLE version 3.52 (26) (see subsequently). Preliminary neighbor joining (NJ) trees were constructed using MEGA 4.0 (27) to identify major subgroups in these two gene families. Protein sequences from each subgroup were used as queries for the second round search of homologous sequences against protein and genome database of NCBI and JGI (Joint Genome Institute) by using BLASTP and TBLASTN with an *e*-value $<e^{-10}$ as cutoff. Human MutS and MutL protein sequences were used as queries for searching eukaryotic MutS and MutL homologs from representative eukaryotes by using BLASTP or TBLASTN against the NCBI databases with an *e*-value $<e^{-10}$ as cutoff. For the following species, common names are shown in figures: Arabidopsis, *Arabidopsis thaliana*; beetle, *Tribolium castaneum*; budding yeast, *Saccharomyces cerevisiae*; chicken, *Gallus gallus*; frog, *Xenopus laevis*; fission yeast, *Schizosaccharomyces pombe*; fruitfly, *Drosophila melanogaster*; humans, *Homo sapiens*; mosquito, *Anopheles gambiae*; rice, *Oryza sativa japonica*; sea urchin, *Strongylocentrotus purpuratus*; and zebrafish, *Danio rerio*.

### Sequence alignment

Preliminary multiple sequence alignments of all putative *MutS* and *MutL* homologs were carried out using MUSCLE version 3.52 with default parameter settings (26). According to NJ trees based on preliminary alignments, we divided *MutS* and *MutL* gene families into several major subgroups. A second round of multiple sequence alignments on each subgroup was performed using MUSCLE. These alignments were subsequently inspected and corrected manually by using GeneDoc version 2.6.002 (28).The improved alignments were then combined by using the profile alignment mode of CLUSTALX 1.81 (29).

### Phylogenetic analysis

Because the *MutS* and *MutL* homologs from all living organisms have diverged over vast evolutionary distances, synonymous nucleotide substitutions are likely saturated

**Table 2.** Distribution of *MutS* and *MutL* homologs in bacteria

| Group | Representative species | Number[a] | MutL | MutS1 | MutS2 | MutS3 | MutS4 |
|---|---|---|---|---|---|---|---|
| Firmicutes | *Bacillus subtilis* | 83 | + | + | + | | |
| | *Staphylococcus aureus* | 8 | + | + | + | + | |
| | *Thermoanaerobacter tengcongensis* | 1 | + | + | + | | + |
| | *Mycoplasma hyopneumoniae* | 16 | | | | | |
| Proteobacteria | *Escherichia coli* | 223 | + | + | | | |
| | *Geobacter metallireducens* | 7 | + | + | + | | |
| | *Myxococcus xanthus* | 3 | + | + | + | + | |
| | *Helicobacter pylori* | 11 | | | + | | |
| Cyanobacteria | *Synechococcus sp.* | 23 | + | + | + | | |
| | *Gloeobacter violaceus* | 1 | + | + | + | + | |
| Others | *Cytophaga hutchinsonii* | 18 | + | + | | | |
| | *Chlorobium tepidum* | 19 | + | + | + | | |
| | *Mycobacterium sp.* | 37 | | | | | |
| | *Polaribacter irgensii* | 6 | + | + | + | + | |
| | *Rhodopirellula baltica* | 1 | + | + | | + | |
| | *Chloroflexus aurantiacus* | 4 | + | + | + | + | + |

'+' indicates the gene is present and blank means that the gene is not present.
[a]The number of species that has the pattern of gene distribution.

and DNA sequence analysis would be quite noisy in constructing phylogenetic trees (30). Therefore, we used protein sequences rather than nucleotide sequences in this study. NJ trees were constructed by using MEGA 4.0 (27) and maximum likelihood (ML) trees were constructed by using PHYML version 2.4 (31). The reliability of internal branches for NJ trees was assessed with 1000 bootstrap pseudoreplicates using 'pairwise deletion option' of amino acid sequences with Poisson correction (unless indicated otherwise). ML trees were generated in PHYML with 100 replicates of nonparametric bootstrap analysis. The discrete gamma model was used in ML analysis and Gamma shape parameters alpha and proportion of invariable sites were estimated from the data. The JTT (Jones, Taylor & Thornton) amino acid substitution model was used in ML analysis. The ML trees were also inferred by quartet puzzling method for reference (trees are available upon request) (32). Only NJ trees are presented and bootstrap values from both NJ and ML methods are shown on the NJ trees because the two methods yielded very similar tree topologies.

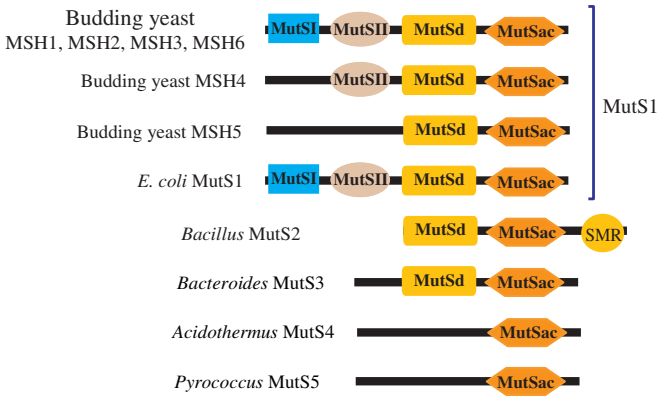


**Figure 1.** A schematic diagram of domain structures of representative MutS homologs. Different domains are depicted by different shapes and colors as indicated, not to scale. Domain names are indicated on each domain.

## RESULTS AND DISCUSSIONS

### Presence of four *MutS* subfamilies in bacterial species

Only two *MutS* subfamilies, *MutS1* and *MutS2*, were identified in previous studies (7,8). In contrast, we found at least four different *MutS* subfamilies in bacterial genomes (Table 2, Figure 2 and Supplementary Table S1). The two newly identified bacterial *MutS* subfamilies were designated as *MutS3* and *MutS4*. The *MutS1* homologs are present in 86% of bacterial species examined in this study, suggesting that the MMR system is widespread in bacteria (Table 2 and Supplementary Table S1). MutS1 proteins contain four conserved domains that are designated as MutS-I, MutS-II, MutSd and MutSac (Figure 1) (33–36). The MutSac domain is the most conserved domain and plays crucial roles in MMR, including dimerization, ATPase and DNA-binding activities (34). The *MutS2* homologs were also found in many bacterial species (36% of bacterial species examined, Table 2 and Supplementary Table S1). Interestingly, *MutS2* homologs are usually present in these *MutS1*-containing species except the ε-Proteobacteria (Table 2). MutS2 proteins lack the MutS-I and MutS-II domains, but share significant similarity with MutS1 proteins in the MutSd and MutSac domains (Figure 1). Furthermore, MutS2 proteins contain an extra ~250 amino acid C-terminal region, which contains a ~90 amino acid-conserved domain called SMR (Small MutS Related) (37).

The newly identified *MutS3* genes were found only in a limited number of distantly related bacterial species. Many of these species contain two copies of *MutS3*, denoted as *MutS3A* and *MutS3B* (Table 2 and Supplementary Table S1). The *MutS3A* and *MutS3B* genes from various bacterial species form two separate clades, suggesting that

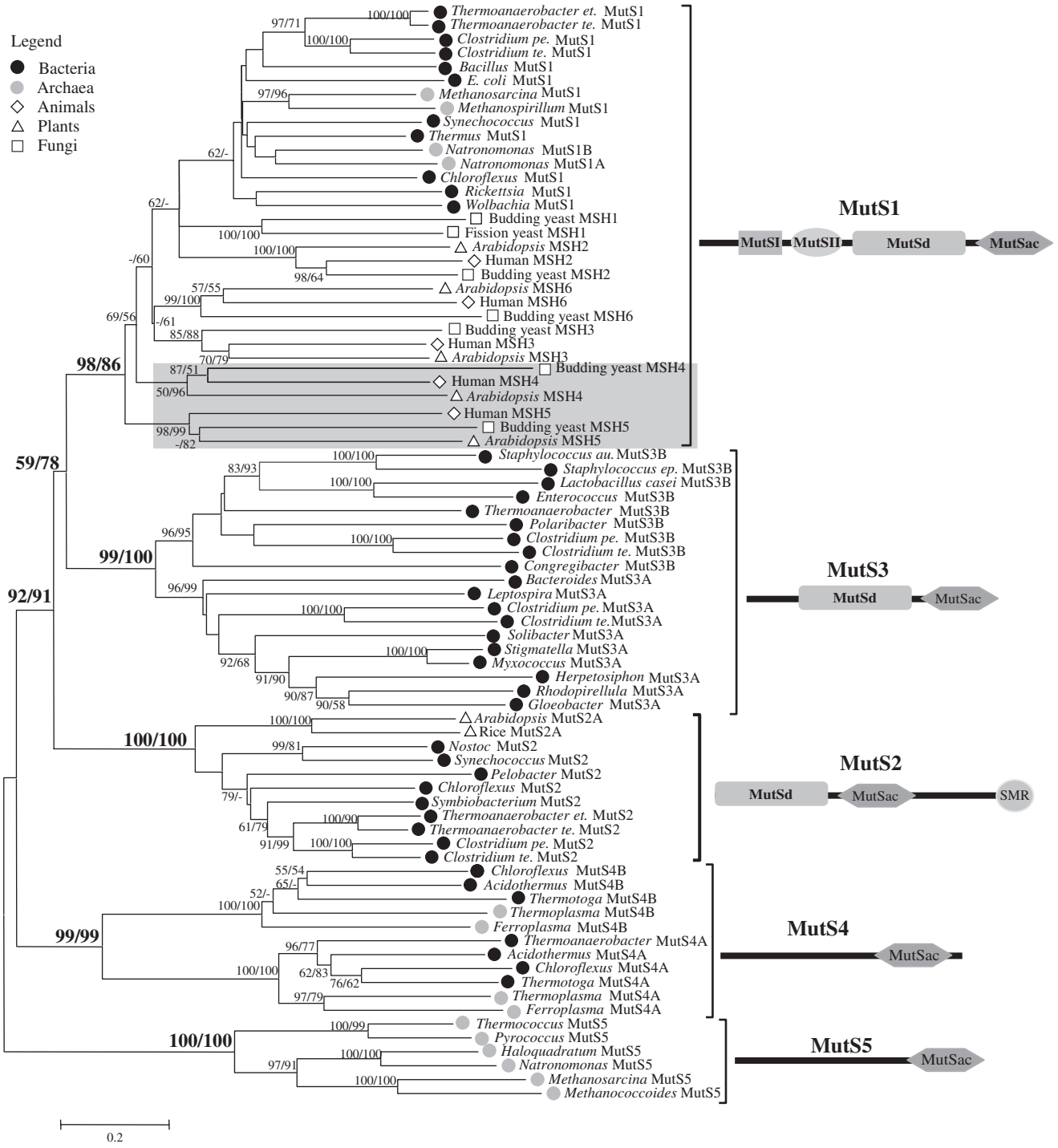


**Figure 2.** A phylogenetic tree of the *MutS* gene family. The evolutionary history was inferred using the NJ and ML methods using the MutSac domain region. The representative protein structures of each subfamily are shown. NJ and ML consensus trees were topologically congruent except for some internal branches that were not statistically significant. Only NJ tree is shown and the NJ tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Bootstrap values from NJ and ML are presented for each clade with >50%. The hyphen '-' is used if bootstrap supports <50% or inconsistent topology between NJ and ML. The *MSH4* and *MSH5* genes are highlighted with gray shading showing their grouping with other *MutS1* genes. The representative protein domain structures of each subfamily are shown next to each subfamily.

they were produced by duplication before the divergence of major bacterial groups, and have been lost in most of the bacterial species subsequently (Figure 2). Given the fact that *MutS3* genes are present in limited bacterial species, it is also possible that *MutS3* might have originated in a specific bacterial lineage and then spread to distantly related bacteria by horizontal gene transfer (HGT, defined as transfer between different species).

Because *MutS3A* and *MutS3B* are not closely linked, the HGT hypothesis would involve separate transfers of *MutS3A* and *MutSB* to multiple different lineages and more evidence is needed to support the latter scenario. All deduced MutS3 proteins contain the MutSac domain near the C-terminus and some of them also have the MutSd domain (Figure 1).

Members of the fourth bacterial *MutS* subfamily, *MutS4*, are also encoding MutSac domain-containing proteins (Figure 1). *MutS4* gene were only detected in five distantly related bacterial species and four of them contain two copies, *MutS4A* and *MutS4B* (Table 2, Supplementary Table S1). Like the *MutS3* genes, the two *MutS4* genes could also be generated by duplication in the ancestral bacteria and lost in most bacterial species (Figure 2). Alternatively, HGT of *MutS4A* and *MutS4B* between bacterial species is also possible. In these bacterial genomes, the two *MutS4* genes are adjacent and the stop codon of *MutS4A* overlaps with the initiation codon of *MutS4B*. The conserved gene organization suggests that the two *MutS4* genes could be produced by tandem duplication in one lineage, followed by HGT to other distantly related bacteria.

Although the biological functions of *MutS3* and *MutS4* genes have not been studied, the presence of the MutSac domain in these proteins suggests that they might be involved in DNA metabolism in these species. However, their absence from most bacteria suggests that since they are not essential, they gradually became diversified or lost during evolution. Furthermore, presence of two duplicate *MutS3* or *MutS4* genes might have accelerated their diversification in some species, as supported by the relatively long branches associated with the duplicates. This could partially explain why they are not as conserved as the *MutS1* genes.

### Presence and origins of *MutS* homologs in archaeal genomes

The *MutS1* orthologs were only detected from nine archaeal species, which all belong to the Phylum Euryarchaeota (Supplementary Table S1). These nine species could be classified into two groups: halophiles and methanogens. Notably, two similar *MutS1* genes were present in the halophiles. The archaeal MutS1 proteins shared identical domain structure with the bacterial MutS1 proteins, suggesting that they are closely related to their bacterial counterparts. Surprisingly, according to the phylogenetic trees based on the MutSac domain (Figure 2), the archaeal *MutS1* genes were nested within bacterial *MutS1* genes and formed two separated groups instead of a monophyletic group. This topology was further confirmed by another phylogenetic analysis using all four domains shared by all prokaryotic MutS1 proteins (Figure 3A). One of archaeal *MutS1* group, including all methanogen *MutS1* genes, are closely related to *MutS1* from Firmicutes, a group of Gram-positive bacteria, with strong bootstrap supports. The halophile *MutS1A* and *MutS1B* genes were likely produced by gene duplication before divergence of these halophiles, formed the other archaeal group (Figure 3A).

Considering the fact that the archaeal *MutS1* genes are only present in nine species, and the phylogenetic topology is incongruent with the universal tree of life based on the small subunit rRNA (38), it is likely that HGT has occurred with *MutS1* genes. The special affinity of *MutS1* genes between the methanogen archaea and Firmicutes strongly supports a possible HGT from Firmicutes to the methanogens. However, it is unclear about the bacterial donor of the two halophile *MutS1* genes due to insufficient phylogenetic evidence. The uncertain origin of the halophile *MutS1* genes could be because of the sequence divergence following the duplication in ancestral halophiles and the consequent reduction of sequence similarity to their bacterial donor genes. The halophile *MutS1* genes were most similar to Firmicutes *MutS1* among bacterial species in BLASTP searches, suggesting a possible Firmicutes origin of halophile *MutS1* genes.

We did not detect any gene that is significantly similar to the bacterial *MutS2* and *MutS3* genes from archaeal genomes. However, two copies of *MutS4*-like genes were found in each of the two closely related thermophilic archaeal species *Thermoplasma volcanium* and *Ferroplasma acidarmanus*. The two archaeal *MutS4*-like genes were grouped into *MutS4A* and *MutS4B* subgroup, respectively (Figure 2). Therefore, each of the *MutS4A* and *MutS4B* groups contains both bacterial and archaeal members, suggesting a possible duplication event prior to the divergence of bacteria and archaea. Alternatively, the archaeal species might have acquired the *MutS4* genes from bacteria through HGT or vice versa, because archaeal *MutS4A* and *MutS4B* are also neighboring genes similar to their bacterial counterparts. The highly similar gene organization between bacterial and archaeal species strongly suggests HGT between them. Because the four *MutS4*-containing bacterial species are distributed in distantly related taxonomic groups, the *MutS4* genes should exist in bacteria prior to their divergence, if we do not consider HGT in this case. In contrast, the two *MutS4*-containing archaeal species are taxonomically closely related, suggesting a more recent origin of *MutS4*. Therefore, HGT of *MutS4* from bacteria to archaea is more favored under the parsimonious assumption. However, since the distance between the two archaeal species are similar to or larger than that of bacteria, the opposite scenario is also possible if these genes have evolved with similar rates.

In addition to the *MutS1* and *MutS4* genes, a novel type of *MutS* genes were identified in 14 archaeal species and designated here as the *MutS5* subfamily. The deduced archaeal MutS5 proteins share significant similarity with other MutS-like proteins in the MutSac domain (Figure 1). The *Pyrococcus furiosus MutS5* gene was previously regarded as a *MutS2*-like gene (19). *PfMutS5* encodes a protein possessing thermostable ATPase and nonspecific DNA-binding activities, but no detectable mismatch-specific DNA-binding activity, suggesting that the *MutS5* genes might be involved in other DNA metabolic activities in archaea. The *MutS5* genes formed a separate clade in the tree shown in Figure 2, suggesting that *MutS5* genes have diverged from other *MutS*-like genes during early cellular evolution.

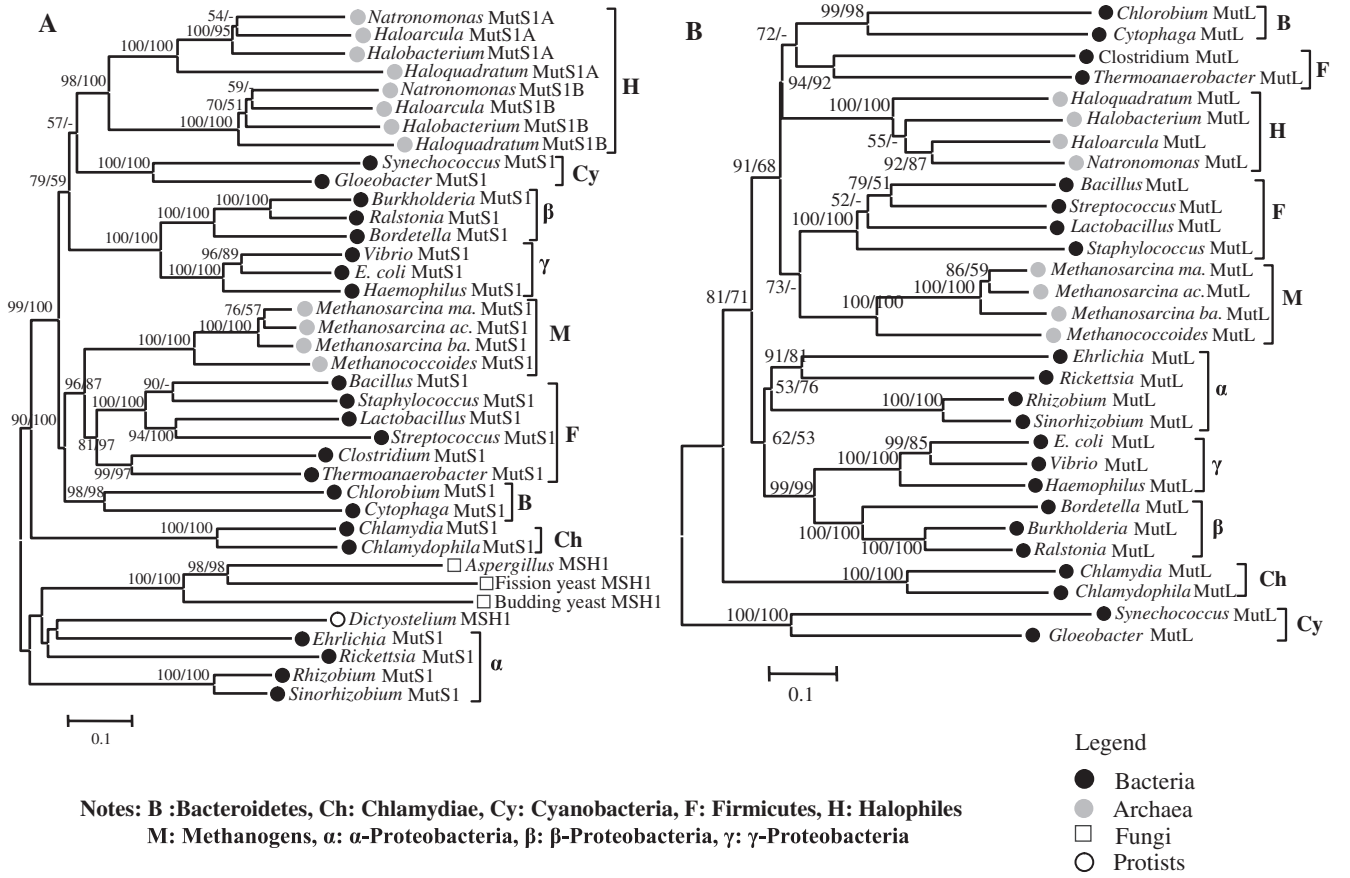


**Figure 3.** Phylogenetic trees of prokaryotic-type of *MutS1* and *MutL* genes. (**A**) A phylogenetic tree of prokaryotic *MutS1* genes and fungal/protist *MSH1* genes. Phylogenetic trees were constructed using all four MutS domain sequences. (**B**) A phylogenetic tree of bacterial and archaeal MutL genes. The tree was reconstructed based on the full length of the MutL protein sequences. The methods used in tree reconstruction and percent bootstrap values are given as in Figure 2.

### Origin and evolution of eukaryotic *MutS1*-like genes

In eukaryotes, only *MutS1* and *MutS2* orthologs were detected (Figure 2 and Supplementary Table S2). The eukaryotic *MutS1*-like genes include MMR genes *MSH1*, *MSH2*, *MSH3*, *MSH6*, as well as nonMMR genes *MSH4* and *MSH5*. *MSH4* and *MSH5* are the most divergent eukaryotic *MutS1* genes in terms of both functions and sequences and thus were classified as *MutS2* genes in a previous study (7). Our phylogenetic analysis indicated that all *MSH* genes, including *MSH4* and *MSH5*, were grouped within the *MutS1* subfamily by both NJ and ML methods with strong bootstrap supports (98 and 86%, respectively). The *MSH4* and *MSH5* genes were located at the most basal position in the *MutS1* subfamily (Figure 2), probably because of long-branch attraction. Without a third group, the *MSH4* and *MSH5* could be mistakenly considered to be closely related to the *MutS2* genes, and this problem was resolved in this study when the other three *MutS* subfamilies were included.

Because the phylogenetic tree shown in Figure 2 was reconstructed based on the ~200 amino acid MutSac domain that is the only domain shared by all MutS proteins, the sequence information could be insufficient to resolve interior branches of the *MutS1* subfamily. To gain further understanding about the origin and evolution of eukaryotic *MutS1*-like genes, additional phylogenetic analysis of the *MutS1* subfamily was performed. To maximize available information, a new phylogenetic tree was generated based on MutSd and MutSac domains that were shared by all MutS1 proteins (Figures 1 and 4A). As shown in Figure 4A, eukaryotic *MSH* genes formed six major paralogous groups (*MSH1–MSH6*). *MSH2–MSH6* genes were found in all major eukaryotic lineages, including animals, plants, fungi and protists. *MSH1* was previously found only in fungal species, but we also detected an *MSH1* ortholog in the slime mold *Dictyostelium discoideum*. Therefore, all six *MSH* genes are present in multiple eukaryotic lineages, suggesting that they were generated by duplication before the divergence of major eukaryotic lineages and the *MSH1* genes were likely lost in animals and plants (Figure 4A).

In addition to the six major *MSH* genes, a number of other *MutS1*-like genes have been identified in some specific lineages, such as the plant *MSH1* and *MSH7* genes and the coral *mtMSH1* gene (39–41). The *MSH7* genes were only detected in plants and are highly similar to the *MSH6* genes. As shown in Figure 4A, the *MSH7* genes were likely resulted from duplication of the plant *MSH6* gene, consistent with a previous preliminary

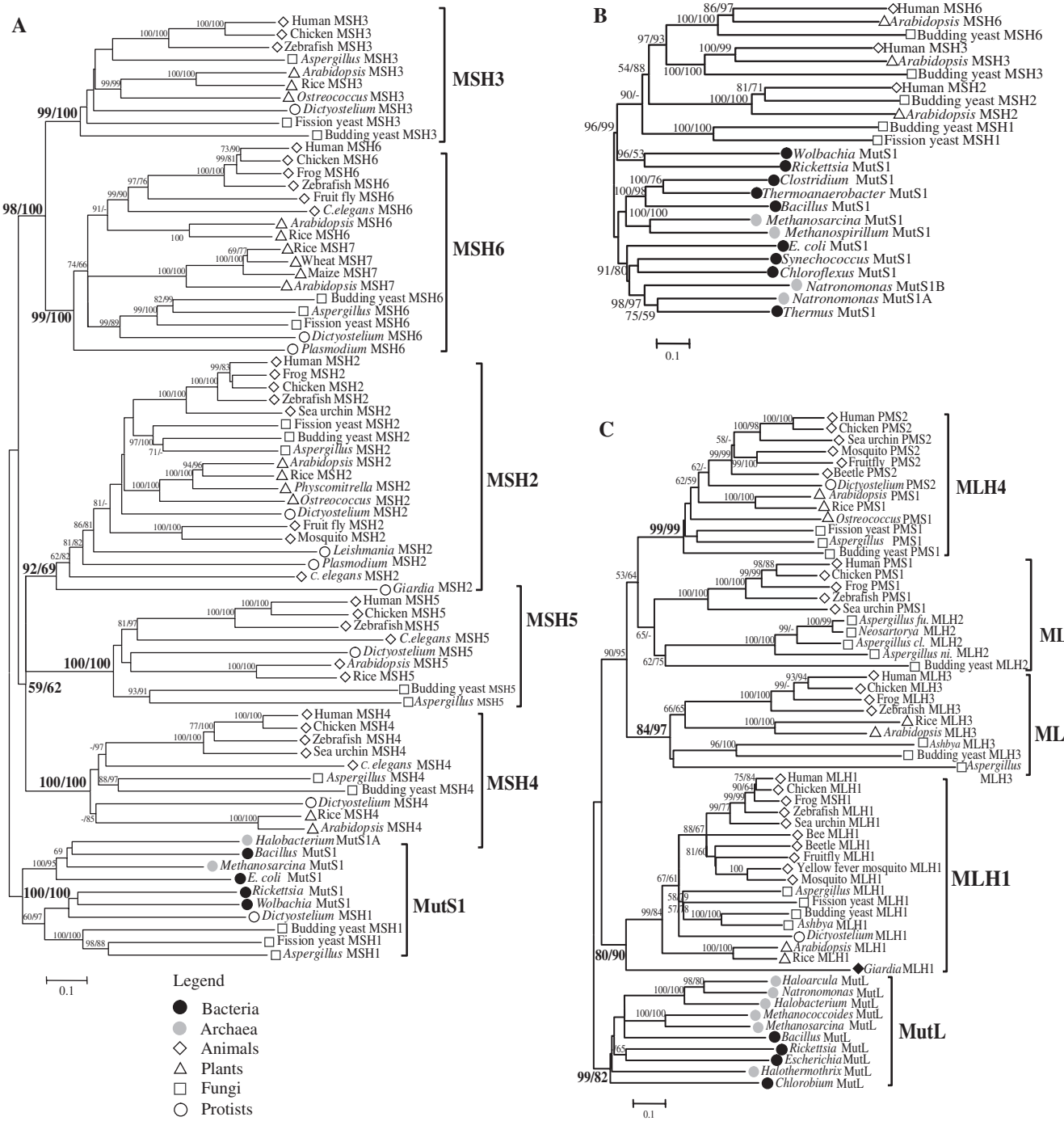


**Figure 4.** Phylogenetic trees of *MutS1* (*MSH*) and *MutL* (*MLH*) genes in eukaryotes. (**A**) A phylogenetic tree of the *MutS1* subfamily in eukaryotes. The evolutionary history was inferred using the MutSd and MutSac domain regions. (**B**) A phylogenetic tree indicating the likely origin of eukaryotic *MSH* genes, constructed using all four MutS domains. The *MSH4* and *MSH5* genes were excluded in this tree due to the lack of MutS-I and/or MutS-II domains. (**C**) A phylogenetic tree of the eukaryotic *MutL* (*MLH*) genes. The methods used in tree reconstruction and percent bootstrap values are given as in Figure 2.

phylogenetic analysis (40). Due to considerably diverged sequences of plant *MSH1* (distinct from the fungal *MSH1* genes) and coral *mtMSH1* genes, their origins and evolutionary relationships with other *MutS1* genes were not elucidated in this study (results not shown).

Notably, the fungal/protist *MSH1* genes are closely related to the prokaryotic *MutS1* genes (Figure 4A), suggesting that *MSH1* genes were likely the most primitive eukaryotic *MutS1* members. Specifically, the *MSH1* genes grouped with the α-proteobacterial *MutS1* in the phylogenetic tree containing prokaryotic *MutS1* and *MSH1* genes using all four conserved MutS1 domains (Figure 3A). The unusual affinity between eukaryotic and α-proteobacterial *MutS1* homologs strongly suggests

the occurrence of HGT between them. It is widely accepted that the eukaryotic mitochondria originated from an α-proteobacterium-like endosymbiont (42). To test whether the other eukaryotic *MSH* genes *(MSH2–MSH7)* originated from *MSH1* or another bacterial lineage, a separate phylogenetic analysis was conducted using all four MutS domains with representative prokaryotic *MutS1* genes and eukaryotic *MSH* genes (Figure 4B). Relatively strong support was obtained for the hypothesis that the other eukaryotic *MSH* genes were derived from *MSH1*, not one of the other bacterial lineages. Thus, it is reasonable to postulate that ancestral eukaryotes acquired *MutS1*-type genes from the α-proteobacterium-like precursor of mitochondria. This scenario is further supported by the findings that the fungal *MSH1* is involved in repairing mitochondrial DNA mismatches (13). The ancestral *MutS1(MSH1)* gene, similar to many other organelle genes (43), was translocated from the mitochondrial genome to the nuclear chromosome during early eukaryotic evolution. Multiple gene duplication events on the ancestral eukaryotic *MutS1* (*MSH1*) occurred, probably after its integration into nuclear genome, and produced at least six additional *MSH* genes.

Although the origin of eukaryotes is still controversial, it has been shown that archaea are more closely related to eukaryotes with regard to genes involved in DNA replication and repair, transcription and translation, all of which are called informational genes (44). Genome-wide comparisons between yeast, bacteria and archaea also suggested that informational genes of eukaryotes were derived almost exclusively from archaea (45). For example, among DNA repair genes, the eukaryotic recombinational repair gene *RAD51* is more closely related to the archaeal *RADA* than to the bacterial *recA* (46). Remarkably, as another major group of DNA repair genes, the eukaryotic *MutS1* (*MSH*) genes apparently originated from bacteria, instead of archaea. Therefore, our study provides a prominent example for an alternative origin for eukaryotic informational genes.

### Functional diversification of *MSH* genes and implication on eukaryotic evolution

As described above, the eukaryotic *MHS* genes experienced multiple gene duplication events (Figure 4A). Duplicated gene copies provide extra genetic materials for functional specialization and innovation (47–49). In *E. coli*, the MutS1 proteins form asymmetric homodimers for DNA repair, suggesting that the two subunits play non-identical roles in MMR (35,36). In eukaryotes, different types of DNA mismatches are repaired by two different heterodimers, MSH2/MSH3 and MSH2/MSH6, instead of asymmetric homodimers (3). Therefore, the expansion of the *MutS1* subfamily in eukaryotes probably allowed functional specialization of the duplicated *MSH* genes in two ways. First, the asymmetric MutS1 homodimer was replaced by MSH heterodimers, allowing additional freedom to specialize in each subunit of the heterodimers. Second, the MutS1 homodimer was replaced by two different heterodimers, making it possible for each heterodimer to evolve functionally to repair specific type of DNA errors. The duplication and subsequent functional divergence might have enhanced the efficiency of MMR in eukaryotes.

Furthermore, *MSH4* and *MSH5* are not required for MMR, but are indispensable for stabilizing heteroduplex formation between nonidentical homologous sequences during meiotic recombination (50,51). The emergence of *MSH4* and *MSH5* might have facilitated the evolution of meiosis by allowing the interaction of homologous, yet mismatched, sequences. Therefore, the specialization and innovation of *MSH* gene functions could have contributed to evolution of MMR and meiosis, which are critical for the evolutionary success of eukaryotes.

Interestingly, the evolution of eukaryotic *MSH* genes is similar to that of the recombinational repair gene *RAD51* in several regards (46). First, both gene families have experienced multiple gene duplication in the ancestral eukaryote. Second, each duplicate gene has been maintained as single copy over vast evolutionary distances after the divergence of major lineages of eukaryotes, suggesting a very strong selection for a single copy. Third, meiosis-specific genes were generated in both the gene families. These similarities suggest that a certain class of eukaryotic multiple-gene families, which are important for DNA metabolism, might have evolved through similar mechanism(s).

### Origin and evolution of plant *MutS2* homologs

The eukaryotic *MutS2* genes are only found in chloroplast-containing species, such as plants and green algae (Supplementary Table S2). We detected two copies of *MutS2*-like genes in each of the nuclear genomes of the flowering plants *Arabidopsis thaliana* and rice (*Oryza sativa japonica*), and three copies of *MutS2*-like genes from the genome of the moss *Physcomitrella patens*. Phylogenetic trees of *MutS2* subfamily show that all eukaryotic *MutS2* genes, except for the moss *MutS2C* gene, formed a well-supported clade and were most closely related to the cyanobacterial *MutS2* (Figure 5). It is well

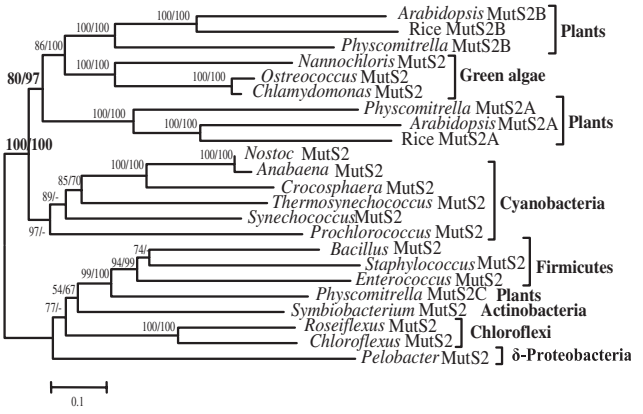


**Figure 5.** A phylogenetic tree of *MutS2* subfamily from representative species showing horizontal gene transfer between bacteria and plants. Phylogenetic trees were constructed using full-length MutS2 protein sequences. The methods used in tree reconstruction and percent bootstrap values are given as in Figure 2.

accepted that the plant chloroplasts were derived from an ancestral endosymbiont related to cyanobacteria (42). Therefore, the eukaryotic *MutS2* gene was apparently transferred from the ancestral chloroplast genome to the nuclear genome and it also explains the presence of eukaryotic *MutS2*-like genes only in chloroplast-containing species. Furthermore, the *MutS2* gene was duplicated before the divergence of land plants and green algae, producing two similar paralogs, *MutS2A* and *MutS2B* (Figure 5).

In addition to the HGT from cyanobacteria to plants, a separate HGT event might have occurred from Firmicutes to *Physcomitrella*, resulting in the presence of *MutS2C* in its genome (Figure 5). An intron is present in the *MutS2C* genomic sequence, so it is not likely that *MutS2C* is a microbial contaminant. Although HGT between Firmicutes to moss is unusual, it is not unique to the *MutS2C* gene, because a similar HGT event has been reported for the *MIP* gene (52). It is not clear how HGT occurred from Firmicutes to the moss and what biological roles *MutS2C* plays, but it is worthwhile to postulate that HGT could occur more frequently and play more important roles than previously recognized during the evolution of multicellular organisms.

### A model for the early evolution of the *MutS* gene family

The evolutionary history of each *MutS* subfamily was elucidated after detailed analyses of their distributions and phylogenies. One of the key points is that HGT events apparently have frequently occurred in several *MutS1* subfamilies from bacteria to eukaryotes and/or archaea. These included separate events for *MutS1* from bacteria to eukaryotes and archaea, and *MutS2* from bacteria to plant and algae via the endosymbiosis of the chloroplast. Furthermore, the *MutS3* and *MutS4* genes might also be transferred between different lineages of bacteria or between bacteria and archaea. MMR genes (*MutS1* and *MutL*) have been suggested to be frequently transferred between different strains of *E. coli* (53,54). Such frequent HGT between closely related bacteria could also serve as an alternative to gene duplication in the production of new copies in a gene family. Our study further indicates that HGT of *MutS* family genes also have frequently occurred between distantly related species through various pathways. The *MutS1* gene was involved in preventing homologous recombination between divergent sequences (1). As a consequence, we could expect elevated recombination rate in those lost of *MutS1* prokaryotic species, consistent with the observation that many *E. coli* populations experience frequent losses and reacquisitions of MMR genes (53). We could speculate that certain archaea lineages reacquired *MutS1* from bacteria after the loss of *MutS1* in ancestral archaea. Similarly, there could also be HGT of *MutS1* between distantly related bacteria groups. Nevertheless, the possible HGT of *MutS1* between different bacteria lineages would have no impact on our results and discussion about the origins of eukaryotic and archaeal *MutS1* genes.

If we assume that the *MutS4* genes were transferred from bacteria to archaea and then exclude genes produced by HGT from the phylogenetic tree shown in Figure 2, it can be significantly simplified as a tree including four bacterium-specific groups (*MutS1-4*) and one archaea-specific group (*MutS5*). Because each of *MutS1-4* subfamilies is present in divergent bacterial groups, it is reasonable to postulate that they were produced by several gene duplication events before the divergence of bacteria. However, it is difficult to determine the evolutionary relationships among the five subfamilies without knowing the true root of the phylogenetic tree. Theoretically, the root could be designated at any point between two major sister clades. Among these possibilities, the most parsimonious scenario is to root the tree between *MutS5* and the joint clade of the other four subfamilies. According to this hypothesis, the ancestral *MutS* gene was present in the common ancestor of bacteria and archaea. Since the split of bacteria and archaea, the *MutS* gene evolved differently in the two groups (Figure 6). The ancestral bacteria *MutS* were duplicated and produced four subfamilies (*MutS1-4*). In contrast, the ancestral archaeal *MutS* gene was maintained as single copy (*MutS5*). *MutS5* was lost in most archaeal species possibly due to appearance of new mismatch repair gene (55) or acquisition of MMR from bacteria (this study). As a result of the fading of the archaeal *MutS5* gene, the eukaryotes, which are believed to share a last common ancestor with archaea, did not inherit the *MutS1* gene from an archaea-like ancestor, but from their bacterial endosymbionts.

In addition, other scenarios are also theoretically possible. One of the hypotheses is to locate the root between *MutS1* and the other four subfamilies. In this case, if the gene duplication generating the *MutS1* and the ancestor of the other four subfamilies occurred before the divergence of archaea and bacteria, multiple gene loss events of *MutS1-3* should have occurred in archaea. If the gene duplication occurred after the divergence of archaea

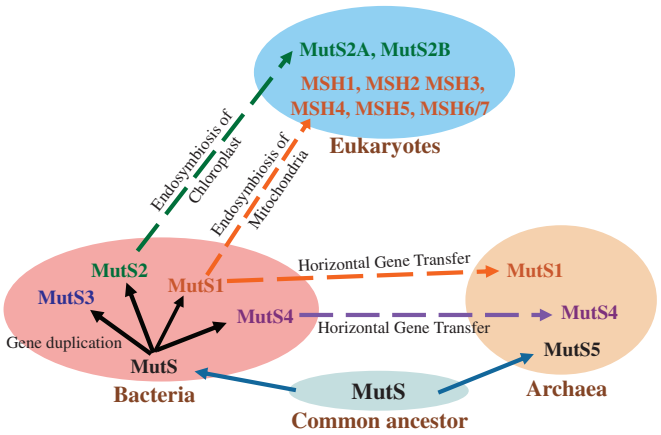


**Figure 6.** A model of the evolutionary history of the *MutS* gene family. Multiple gene duplication events have occurred on the ancestral *MutS* gene before the divergence of bacteria, producing four lineages, *MutS1*, *MutS2*, *MutS3* and *MutS4*. The ancestral *MutS* gene evolved to *MutS5* gene in archaea. Eukaryotic *MutS1* homologs were acquired from ancient α-proteobacteria through endosymbiosis of mitochondria. The archaeal *MutS1* genes were likely originated from Firmicutes by HGT. The plant and green algae *MutS2* genes were obtained from ancient cyanobacteria by endosymbiosis of chloroplasts. Some archaeal species acquired the *MutS4* gene from bacteria by HGT.

and bacteria, it means that the *MutS* genes were present only in ancestral bacteria and all archaeal *MutS* genes, including *MutS5*, were acquired by HGT from bacteria. However, the *MutS5* genes do not share significant affinity to any of the bacterial *MutS* genes; therefore, it is unreasonable to propose that *MutS5* originated from a specific subfamily of bacterial *MutS* genes by HGT. Although other rooting possibilities cannot be ruled out, the first scenario is most favored according to the current data. Furthermore, the position of the root does not affect our major conclusions that the *MutS* family has five subfamilies and multiple HGT have occurred from bacteria to eukaryotes and archaea.

### Evolution of the *MutL* gene family

Our searches of *MutL* homologs uncovered a most intriguing result that they are only present strictly in *MutS1*-containing species and vice versa (Table 2 and Supplementary Tables S1 and S2). All MutL proteins share two highly conserved domains, the HATPase and DNA mismatch repair domains (Supplementary Figure S1). Therefore, the phylogeny of the *MutL* gene family was reconstructed based on these two domains (Figure 4C). As shown in Figure 4C, the eukaryotic *MutL* homologs formed four well-supported clades, and archaeal and bacterial *MutL* genes formed the fifth clade. Three of the eukaryotic subgroups contain sequences from fungi, plants and animals, indicating that the four eukaryotic *MutL* homologs were generated by gene duplication prior to the divergence of major eukaryotic lineages.

Our phylogenetic analysis showed that plant and fungal *PMS1* genes are grouped with animal *PMS2*, indicating that plant and fungal *PMS1* genes are the orthologs of the animal *PMS2*, rather than the animal *PMS1* genes. This is consistent with previous functional studies on these genes (21,56). In addition, another clade contains the fungal *MLH2* and animal *PMS1* genes. To avoid confusion between gene names and orthologous relationships, we designated the group with the fungal *MLH2* and animal *PMS1* genes as the *MLH2* group, and the group with plant and fungal *PMS1* and animal *PMS2* genes as the *MLH4* group (Figure 4C). The *MLH2* genes can only be identified in vertebrate animals and some fungal species, indicating that their orthologs were lost in many eukaryotic organisms.

Like the *MSH* genes, the available functional data of *MLH* genes support the idea that the duplicated *MLH* genes have experienced functional specialization and innovation. For example, MLH1 and MLH4 proteins form heterodimers (MutL$\alpha$) and function in MMR during the mitotic cycle, analogous to the prokaryotic MutL homodimers (57,58). Furthermore, the MutL$\alpha$ are also required for MMR of the heteroduplex formed by the meiotic recombination (59,60), indicating that the MutL$\alpha$ have acquired new meiotic roles during evolution. In addition, MLH3 is important for meiotic recombination, particularly the formation of double Holliday junction (61). In summary, the generation and functional diversification of multiple eukaryotic *MLH* genes might also have contributed significantly to the evolution of eukaryotes and meiosis, similar to their functional partner *MutS1* genes.

Similar to the archaeal *MutS1* genes, archaeal *MutL* genes did not form a monophyletic clade (Figure 3B). The *MutL* genes from archaeal methonagens were grouped with Firmicutes with strong bootstrap supports, indicating that methanogen *MutL* genes might also have originated from Firmicutes by HGT. In addition, halophile *MutL* genes form the other archaeal *MutL* clade and were nested in the bacterial *MutL* groups. Therefore, it is likely that halophile *MutL* were acquired from bacteria through HGT, although we were unable to infer their bacterial donor based on current data. Therefore, all *MutL* genes present in archaea were likely transferred from bacteria, suggesting that the *MutL* genes were not present or were lost in ancestral archaea. As a consequence, the eukaryotic *MLH* genes were apparently originated from bacteria, but not archaea. It is reasonable to postulate that the first eukaryotic *MutL* homolog was transferred from $\alpha$-proteobacterium-like endosymbionts along with the *MutS1* gene, although mitochondria-targeted *MutL* homologs were not detected in eukaryotes. One explanation is that the mitochondria-targeted *MutL* homologs have been lost in eukaryotes, consistent with the observation that the mitochondria-targeted *MSH1* genes have been lost in most eukaryotes.

### Co-transfer of *MutS1* and *MutL* genes from bacteria to archaea

To further elucidate the origin of archaeal *MutS1* and *MutL* genes, their genomic locations were compared between bacteria and archaea. The positions of *MutS1* and *MutL* genes were obtained for each species from NCBI genomic database. In most groups of bacteria, the two genes are distantly located, separating by at least 10 000 nucleotides (Figure 7). In contrast, *MutS1* and *MutL* are neighboring genes in most Firmicutes. Coincidentally, the physical proximity of *MutS1* and *MutL* genes was also found in the methanogens of archaea (Figure 7). Our phylogenetic studies showed that *MutS1* and *MutL* genes in methanogens were likely acquired from Firmicutes by HGT. The conservation of the unusual neighboring *MutS1* and *MutL* in both Firmicutes and archaea not only supports the HGT from Firmicutes to archaea, but also suggests that ancient methanogens acquired both *MutS1* and *MutL* genes via a single HGT event. In halophile archara, one of *MutS1* genes, *MutS1B*, is also closely linked to *MutL*. Considering that the best hits of halophilic *MutS1* and *MutL* are from Firmicutes in bacteria, and halophile *MutS1B* and *MutL* are also neighboring genes, we postulate that halophile *MutS1* and *MutL* were also simultaneously transferred from Firmicutes, although this hypothesis lacks support from phylogenetic analysis (Figure 3).

Our analysis indicated that *MutS1* and *MutL* are absent in most archaea species. The mutation rates of genomes could significantly increase without an efficient repair mechanism, especially for those archaea that live in harsh environments. However, the mutation rates are not

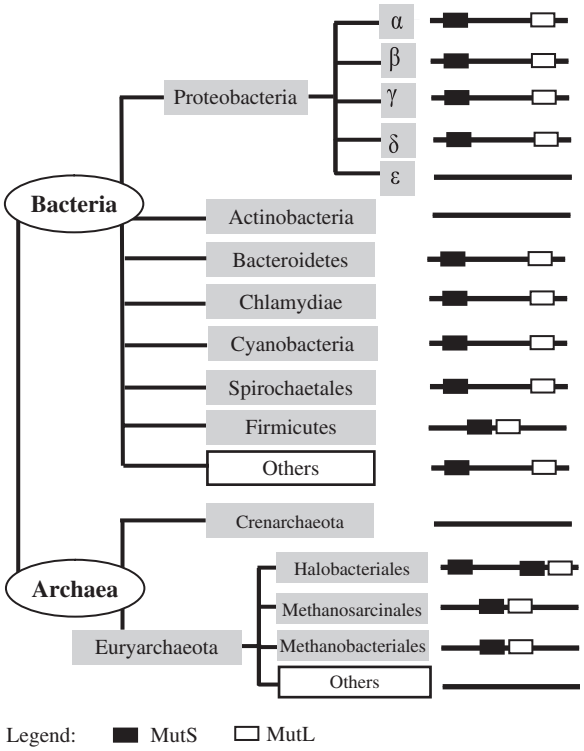


**Figure 7.** The co-occurrence and co-absence of the *MutS1* and *MutL* genes and their genomic locations in bacteria and archaea. Both *MutS* and *MutL* genes are present in most bacteria and a number of archaeal species. The *MutS* and *MutL* genes are closely linked only in Firmicutes of bacteria and archaeal species, indicating that archaeal *MutS* and *MutL* were probably acquired simultaneously from Firmicutes bacteria.

enhanced in some archaeal species without the MutS1/MutL-dependent MMR pathway (62). Therefore, alternative repair pathways should exist in these archaeal species to correct replication errors. Previous genomic context analyses have shown that there is a putative DNA repair system specific for thermophilic archaea and bacteria (55). This putative repair mechanism is not fully analogous to the MMR system, so the understanding of the full extent of DNA mismatch repair pathways awaits future investigations.

### Molecular co-evolution of the MMR duo

We have observed co-occurrence patterns of *MutS1* and *MutL* homologs in cellular organisms, suggesting that a loss of one gene could subsequently lead to the loss of the other gene in a genome. Study on the origins of archaeal *MutS1* and *MutL* genes suggests co-acquisition of these two genes from bacteria. Phylogenetic analysis further indicates that these two gene families share very similar evolution profiles. For example, both gene families have experienced multiple gene duplication events during early evolution of eukaryotes. As a result of gene duplication, different mismatch repair heterodimers and meiosis-specific proteins appeared in both the families. These observations strongly suggest that, as physically interacting duo in MMR, a heritable change in one gene could become selective force for a complementary change in the other one. Therefore, the *MutS1* and *MutL* gene families have evolved in a correlated fashion. Co-evolution at molecular level has been commonly observed between host and parasites, and between ligands and receptors (63–67). However, it has not been reported on DNA metabolic genes to our knowledge. Our study provides a prominent example, suggesting that co-evolution might also play important roles in the gene network of DNA metabolism.

## CONCLUSIONS

This study provides an overall picture of the evolutionary history of *MutS* and *MutL* gene families that play crucial roles in maintaining genome stability. We identified three new subfamilies in the *MutS* family, and showed that the *MutS* gene family has experienced many gene duplication, loss and HGT events during early evolution. Our data suggest that the archaeal *MutS1* and *MutL* genes were originated from bacteria by HGT. The eukaryotic *MutS* and *MutL* homologs were also originated from bacteria, indicating that bacteria could be an important source for eukaryotic informational genes. The results provide direct evidence that genomes are highly dynamic, in part, because they can acquire genes from even very distant organisms. We also showed that the *MutS1* and *MutL* genes display a pattern of strict co-presence and co-absence, indicating that they have evolved in a correlated way. Our results about the origins of *MutL* and *MutS1* homologs of eukaryotes and archaea further support the co-evolution between the *MutL* and *MutS1* genes during a long-term evolutionary history. In summary, our phylogenetic results have established an evolutionary foundation for future studies on the contributions of *MutS* and *MutL* genes to genome stability and the functions of the newly recognized *MutS* subfamilies.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Masafumi Nozawa, Sabyasachi Das, Dimitra Chalkia, Edward Holmes, Hongzhi Kong and Alexandra Surcel for critical reading and valuable comments on the manuscript. Funding for this work was provided by a NIH grant (GM63871) and a grant from the Tobacco Settlement Funds to H.M. and an NIH grant (GM020293) to M.N., and by funds from the Biology Department and the Huck Institutes of the Life Sciences, the Pennsylvania State University. Funding to pay the Open Access publication charges for this article was provided by the Biology Department, the Pennsylvania State University.

*Conflict of interest statement*. None declared.

## REFERENCES


1. Modrich,P. and Lahue,R. (1996) Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu. Rev. Biochem.*, **65**, 101–133.
2. Umar,A. and Kunkel,T.A. (1996) DNA-replication fidelity, mismatch repair and genome instability in cancer cells. *Eur. J. Biochem.*, **238**, 297–307.
3. Iyer,R.R., Pluciennik,A., Burdett,V. and Modrich,P.L. (2006) DNA mismatch repair: functions and mechanisms. *Chem. Rev.*, **106**, 302–323.
4. Reenan,R.A. and Kolodner,R.D. (1992) Characterization of insertion mutations in the *Saccharomyces cerevisiae MSH1* and *MSH2* genes: evidence for separate mitochondrial and nuclear functions. *Genetics*, **132**, 975–985.
5. Harfe,B.D. and Jinks-Robertson,S. (2000) DNA mismatch repair and genetic instability. *Annu. Rev. Genet.*, **34**, 359–399.
6. Surtees,J.A., Argueso,J.L. and Alani,E. (2004) Mismatch repair proteins: key regulators of genetic recombination. *Cytogenet. Genome Res.*, **107**, 146–159.
7. Eisen,J.A. (1998) A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res.*, **26**, 4291–4300.
8. Culligan,K.M., Meyer-Gauen,G., Lyons-Weiler,J. and Hays,J.B. (2000) Evolutionary origin, diversification and specialization of eukaryotic MutS homolog mismatch repair proteins. *Nucleic Acids Res.*, **28**, 463–471.
9. Bjorkholm,B., Sjolund,M., Falk,P.G., Berg,O.G., Engstrand,L. and Andersson,D.I. (2001) Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA*, **98**, 14607–14612.
10. Fukui,K., Masui,R. and Kuramitsu,S. (2004) *Thermus thermophilus* MutS2, a MutS paralogue, possesses an endonuclease activity promoted by MutL. *J. Biochem.*, **135**, 375–384.
11. Rossolillo,P. and Albertini,A.M. (2001) Functional analysis of the *Bacillus subtilis yshD* gene, a *mutS* paralogue. *Mol. Gen. Genet.*, **264**, 809–818.
12. Kang,J., Huang,S. and Blaser,M.J. (2005) Structural and functional divergence of MutS2 from bacterial MutS1 and eukaryotic MSH4-MSH5 homologs. *J. Bacteriol.*, **187**, 3528–3537.
13. Chi,N.W. and Kolodner,R.D. (1994) Purification and characterization of MSH1, a yeast mitochondrial protein that binds to DNA mismatches. *J. Biol. Chem.*, **269**, 29984–29992.
14. Marsischky,G.T., Filosi,N., Kane,M.F. and Kolodner,R. (1996) Redundancy of *Saccharomyces cerevisiae MSH3* and *MSH6* in *MSH2*-dependent mismatch repair. *Genes Dev.*, **10**, 407–420.
15. Sugawara,N., Paques,F., Colaiacovo,M. and Haber,J.E. (1997) Role of *Saccharomyces cerevisiae* Msh2 and Msh3 repair proteins in double-strand break-induced recombination. *Proc. Natl Acad. Sci. USA*, **94**, 9214–9219.
16. Zalevsky,J., MacQueen,A.J., Duffy,J.B., Kemphues,K.J. and Villeneuve,A.M. (1999) Crossing over during *Caenorhabditis elegans* meiosis requires a conserved MutS-based pathway that is partially dispensable in budding yeast. *Genetics*, **153**, 1271–1283.
17. Wu,S.Y., Culligan,K., Lamers,M. and Hays,J. (2003) Dissimilar mispair-recognition spectra of Arabidopsis DNA-mismatch-repair proteins MSH2*MSH6 (MutSalpha) and MSH2*MSH7 (MutSgamma). *Nucleic Acids Res.*, **31**, 6027–6034.
18. Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R. *et al.* (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J. Bacteriol.*, **179**, 7135–7155.
19. Vijayvargia,R. and Biswas,I. (2002) MutS2 family protein from *Pyrococcus furiosus*. *Curr. Microbiol.*, **44**, 224–228.
20. Kolodner,R. (1996) Biochemistry and genetics of eukaryotic mismatch repair. *Genes Dev.*, **10**, 1433–1442.
21. Flores-Rozas,H. and Kolodner,R.D. (1998) The *Saccharomyces cerevisiae MLH3* gene functions in MSH3-dependent suppression of frameshift mutations. *Proc. Natl Acad. Sci. USA*, **95**, 12404–12409.
22. Alou,A.H., Jean,M., Domingue,O. and Belzile,F.J. (2004) Structure and expression of *AtPMS1*, the Arabidopsis ortholog of the yeast DNA repair gene *PMS1*. *Plant Sci.*, **167**, 447–456.
23. Cummings,L., Riley,L., Black,L., Souvorov,A., Resenchuk,S., Dondoshansky,I. and Tatusova,T. (2002) Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. *FEMS Microbiol. Lett.*, **216**, 133–138.
24. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–251.
25. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
26. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
27. Tamura,K., Dudley,J., Nei,M. and Kumar,S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol. Biol. Evol.*, **24**, 1596–1599.
28. Nicholas,K.B., Nicholas,H.B. Jr and Deerfield,D.W. II (1997) GeneDoc: analysis and visualization of genetic variation. *EMBNET News*, **4**, 1–4.
29. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
30. Russo,C.A., Takezaki,N. and Nei,M. (1996) Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.*, **13**, 525–536.
31. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
32. Schmidt,H.A., Strimmer,K., Vingron,M. and von Haeseler,A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.
33. New,L., Liu,K. and Crouse,G.F. (1993) The yeast gene MSH3 defines a new class of eukaryotic MutS homologues. *Mol. Gen. Genet.*, **239**, 97–108.
34. Tachiki,H., Kato,R., Masui,R., Hasegawa,K., Itakura,H., Fukuyama,K. and Kuramitsu,S. (1998) Domain organization and functional analysis of *Thermus thermophilus* MutS protein. *Nucleic Acids Res.*, **26**, 4153–4159.
35. Lamers,M.H., Perrakis,A., Enzlin,J.H., Winterwerp,H.H., de Wind,N. and Sixma,T.K. (2000) The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch. *Nature*, **407**, 711–717.
36. Obmolova,G., Ban,C., Hsieh,P. and Yang,W. (2000) Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature*, **407**, 703–710.
37. Moreira,D. and Philippe,H. (1999) Smr: a bacterial and eukaryotic homologue of the C-terminal region of the MutS2 family. *Trends Biochem. Sci.*, **24**, 298–300.
38. Woese,C.R., Kandler,O. and Wheelis,M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA*, **87**, 4576–4579.
39. Pont-Kingdon,G.A., Okada,N.A., Macfarlane,J.L., Beagley,C.T., Wolstenholme,D.R., Cavalier-Smith,T. and Clark-Walker,G.D. (1995) A coral mitochondrial *mutS* gene. *Nature*, **375**, 109–111.
40. Culligan,K.M. and Hays,J.B. (2000) Arabidopsis MutS homologs-AtMSH2, AtMSH3, AtMSH6, and a novel AtMSH7-form three distinct protein heterodimers with different specificities for mismatched DNA. *Plant cell*, **12**, 991–1002.
41. Abdelnoor,R.V., Yule,R., Elo,A., Christensen,A.C., Meyer-Gauen,G. and Mackenzie,S.A. (2003) Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. *Proc. Natl Acad. Sci. USA*, **100**, 5968–5973.
42. Andersson,S.G., Zomorodipour,A., Andersson,J.O., Sicheritz-Ponten,T., Alsmark,U.C., Podowski,R.M., Naslund,A.K., Eriksson,A.S., Winkler,H.H. *et al.* (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 133–140.
43. Timmis,J.N., Ayliffe,M.A., Huang,C.Y. and Martin,W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev.*, **5**, 123–135.
44. Brown,J.R. and Doolittle,W.F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.*, **61**, 456–502.

45. Rivera,M.C., Jain,R., Moore,J.E. and Lake,J.A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl Acad. Sci. USA*, **95**, 6239–6244.
46. Lin,Z., Kong,H., Nei,M. and Ma,H. (2006) Origins and evolution of the *recA/RAD51* gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc. Natl Acad. Sci. USA*, **103**, 10328–10333.
47. Bridges,C.B. (1935) Salivary chromosome maps: with a key to the banding of the chromosomes of *Drosophila melanogaster*. *J. Hered.*, **26**, 60–64.
48. Lewis,E.B. (1951) Pseudoallelism and gene evolution. *Cold Spring Harbor Symp. Quant. Biol.*, **16**, 159–174.
49. Ohno,S. (1970) *Evolution by gene duplication* Springer-Verlag, Berlin.
50. Hollingsworth,N.M., Ponte,L. and Halsey,C. (1995) *MSH5*, a novel MutS homolog, facilitates meiotic reciprocal recombination between homologs in *Saccharomyces cerevisiae* but not mismatch repair. *Genes Dev.*, **9**, 1728–1739.
51. Snowden,T., Acharya,S., Butz,C., Berardini,M. and Fishel,R. (2004) hMSH4-hMSH5 recognizes Holliday Junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Mol. Cell*, **15**, 437–451.
52. Gustavsson,S., Lebrun,A.S., Norden,K., Chaumont,F. and Johanson,U. (2005) A novel plant major intrinsic protein in *Physcomitrella patens* most similar to bacterial glycerol channels. *Plant Physiol.*, **139**, 287–295.
53. Denamur,E., Lecointre,G., Darlu,P., Tenaillon,O., Acquaviva,C., Sayada,C., Sunjevaric,I., Rothstein,R., Elion,J. *et al.* (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell*, **103**, 711–721.
54. Brown,E.W., LeClerc,J.E., Li,B., Payne,W.L. and Cebula,T.A. (2001) Phylogenetic evidence for horizontal transfer of mutS alleles among naturally occurring *Escherichia coli* strains. *J. Bacteriol.*, **183**, 1631–1644.
55. Makarova,K.S., Aravind,L., Grishin,N.V., Rogozin,I.B. and Koonin,E.V. (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.*, **30**, 482–496.
56. Lipkin,S.M., Wang,V., Jacoby,R., Banerjee-Basu,S., Baxevanis,A.D., Lynch,H.T., Elliott,R.M. and Collins,F.S. (2000) *MLH3*: a DNA mismatch repair gene associated with mammalian microsatellite instability. *Nat. Genet.*, **24**, 27–35.
57. Kramer,B., Kramer,W., Williamson,M.S. and Fogel,S. (1989) Heteroduplex DNA correction in *Saccharomyces cerevisiae* is mismatch specific and requires functional *PMS* genes. *Mol. Cell. Biol.*, **9**, 4432–4440.
58. Strand,M., Prolla,T.A., Liskay,R.M. and Petes,T.D. (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, **365**, 274–276.
59. Baker,S.M., Plug,A.W., Prolla,T.A., Bronner,C.E., Harris,A.C., Yao,X., Christie,D.M., Monell,C., Arnheim,N. *et al.* (1996) Involvement of mouse Mlh1 in DNA mismatch repair and meiotic crossing over. *Nat. Genet.*, **13**, 336–342.
60. Edelmann,W., Cohen,P.E., Kane,M., Lau,K., Morrow,B., Bennett,S., Umar,A., Kunkel,T., Cattoretti,G. *et al.* (1996) Meiotic pachytene arrest in MLH1-deficient mice. *Cell*, **85**, 1125–1134.
61. Santucci-Darmanin,S., Neyton,S., Lespinasse,F., Saunieres,A., Gaudray,P. and Paquis-Flucklinger,V. (2002) The DNA mismatch-repair MLH3 protein interacts with MSH4 in meiotic cells, supporting a role for this MutL homolog in mammalian meiotic recombination. *Hum. Mol. Genet.*, **11**, 1697–1706.
62. Sniegowski,P. (2001) Evolution: constantly avoiding mutation. *Curr. Biol.*, **11**, R929–R931.
63. Moyle,W.R., Campbell,R.K., Myers,R.V., Bernard,M.P., Han,Y. and Wang,X. (1994) Co-evolution of ligand-receptor pairs. *Nature*, **368**, 251–255.
64. Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, **271**, 511–523.
65. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
66. Goh,C.S., Bogan,A.A., Joachimiak,M., Walther,D. and Cohen,F.E. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.
67. Ramani,A.K. and Marcotte,E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.*, **327**, 273–284.