The ISME Journal (2011) 5, 1143–1151

www.nature.com/ismej

ORIGINAL ARTICLE

Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment

Hiroyuki Ogata[1], Jessica Ray[2], Kensuke Toyoda[3,4], Ruth-Anne Sandaa[2], Keizo Nagasaki[3], Gunnar Bratbak[2] and Jean-Michel Claverie[1]

[1]*Information Génomique et Structurale, CNRS-UPR2589, Institut de Microbiologie de la Méditerranée, Parc Scientifique de Luminy, Aix-Marseille Université, Marseille Cedex 9, France;* [2]*Department of Biology, University of Bergen, Bergen, Norway and* [3]*Harmful Algal Bloom Division, National Research Institute of Inland Sea, Fisheries Research Agency, Hiroshima, Japan*

MutS proteins are ubiquitous in cellular organisms and have important roles in DNA mismatch repair or recombination. In the virus world, the amoeba-infecting Mimivirus, as well as the recently sequenced *Cafeteria roenbergensis* virus are known to encode a MutS related to the homologs found in octocorals and ε-*proteobacteria*. To explore the presence of MutS proteins in other viral genomes, we performed a genomic survey of four giant viruses ('giruses') (*Pyramimonas orientalis* virus (PoV), *Phaeocystis pouchetii* virus (PpV), *Chrysochromulina ericina* virus (CeV) and *Heterocapsa circularisquama* DNA virus (HcDNAV)) that infect unicellular marine algae. Our analysis revealed the presence of a close homolog of Mimivirus MutS in all the analyzed giruses. These viral homologs possess a specific domain structure, including a *C*-terminal HNH-endonuclease domain, defining the new MutS7 subfamily. We confirmed the presence of conserved mismatch recognition residues in all members of the MutS7 subfamily, suggesting their role in DNA mismatch repair rather than DNA recombination. PoV and PpV were found to contain an additional type of MutS, which we propose to call MutS8. The MutS8 proteins in PoV and PpV were found to be closely related to homologs from '*Candidatus* Amoebophilus asiaticus', an obligate intracellular amoeba-symbiont belonging to the *Bacteroidetes*. Furthermore, our analysis revealed that MutS7 and MutS8 are abundant in marine microbial metagenomes and that a vast majority of these environmental sequences are likely of girus origin. Giruses thus seem to represent a major source of the underexplored diversity of the MutS family in the microbial world.

The ISME Journal (2011) **5,** 1143–1151; doi:10.1038/ismej.2010.210; published online 20 January 2011
Subject Category: evolutionary genetics
Keywords: mimivirus; girus; virus; DNA repair; MutS

Introduction

Large DNA viruses carry genes for their own DNA repair apparatus to enhance the accuracy of genome replication (Furuta *et al.*, 1997; Srinivasan and Tripathy, 2005; Redrejo-Rodriguez *et al.*, 2009; Bogani *et al.*, 2010). The amoeba-infecting Mimivirus (*Acanthamoeba polyphaga* mimivirus, APMV) with the largest genome (1.2 Mb) of all known viruses encodes eight putative genes for DNA repair enzymes capable of correcting mismatches or errors induced by oxidation, UV irradiation and alkylating agents (Raoult *et al.*, 2004). Most of these genes were never found in a viral genome until their discovery in Mimivirus. One of these corresponds to a MutS homolog (open reading frame (ORF) L359) predicted to function in DNA mismatch repair (MMR) or recombination. MMR recognizes and corrects base–base mismatches and small insertion or deletion loops introduced during replication, leading to 50- to 1000-folds enhancement of replication fidelity in cellular organisms (Schofield and Hsieh, 2003; Iyer *et al.*, 2006). The best-studied MMR system is the *Escherichia coli* MutS–MutL–MutH pathway. In the first step in this pathway, the MutS homodimer binds the site of a mismatch (or a loop) in double-strand DNA. The MutS protein recruits the 'linker protein' MutL and together activates the endonuclease MutH, which nicks specifically the newly synthesized DNA strand to initiate DNA excision and resynthesis pathway. Homologs of *E. coli* MutS have been found in many species of bacteria, archaea and eukaryotes, and together classified in

Correspondence: H Ogata, Information Génomique et Structurale, CNRS-UPR2589, Institut de Microbiologie de la Méditerranée, Parc Scientifique de Luminy, Aix-Marseille Université, 163 Avenue de Luminy, Case 934, Marseille Cedex 9 13288, France.
E-mail: Hiroyuki.Ogata@igs.cnrs-mrs.fr
[4]Current address: Department of Botany, Keio University, Hiyoshi, Kohoku-ku, Yokohama, Kangawa 223-8521, Japan.
Received 28 September 2010; revised 9 December 2010; accepted 12 December 2010; published online 20 January 2011

the MutS family (Eisen, 1998; Lin *et al.*, 2007). In viruses, MutS homologs have only been found in Mimivirus, the closely related Mamavirus (La Scola *et al.*, 2008; Yutin *et al.*, 2009), and more recently in the giant marine virus, *Cafeteria roenbergensis* virus (CroV) with a 730-kb genome (Fischer *et al.*, 2010).

The phyletic distribution of the close homologs of Mimivirus MutS is notable. The Mimivirus MutS homolog is most closely related to the homologs found in the mitochondrial genomes of a group of animals (that is, octocorals) and several genomes of the ε-*Proteobacteria* such as *Sulfurimonas*, *Nitratiruptor* and *Arcobacter* (Claverie *et al.*, 2006, 2009). Octocorals (phylum *Cnidaria*, class *Anthozoa*, subclass *Octocorallia*) include diverse species of corals (for example, soft corals, sea fans, sea pens), representing important members of marine communities from shallow tropical coral reefs to the deep sea (McFadden *et al.*, 2006). A *mutS* homolog has been found encoded in the mitochondria of all octocorals, including the three major orders *Alcyoncea*, *Helioporacea* and *Pennatulacea*, but not in the mitochondrial genomes of any other eukaryotes, including those of the sister subclass *Hexacorallia* (for example, stony corals, sea anemones) (Pont-Kingdon *et al.*, 1995; Brugler and France, 2008). The ε-*proteobacteria Sulfurimonas* and *Nitratiruptor* are sulfur-oxidizing chemoautotrophs and often found in deep-sea hydrothermal vent or coastal sediments (Nakagawa *et al.*, 2007; Sievert *et al.*, 2008). *Arcobacter* includes species of water-borne pathogens and taxonomically close to *Campylobacter jejuni* and *Helicobacter pylori* (Miller *et al.*, 2007). The common origin of these MutS homologs is further suggested by their atypical domain organization. Distinct from all other MutS family proteins, the MutS homologs in Mimivirus, octocorals and the ε-*proteobacteria* are fused with a *C*-terminal HNH nicking endonuclease domain (Malik and Henikoff, 2000; Claverie *et al.*, 2009). The domain fusion was predicted to make these enzymes a 'self-contained' single polypeptide having both mismatch recognition (MutS) and nicking (MutH) functions (Malik and Henikoff, 2000). The distribution of these unique MutS homologs is thus limited to a few totally unrelated lineages (that is, Mimivirus, a single subclass of animals, and the ε-*Proteobacteria*) and suggests the occurrence of gene transfer between their ancestors (Claverie *et al.*, 2009). We introduce the MutS7 subfamily to denote this specific group of MutS proteins.

DNA viruses with genomes greater than 300 kb up to 1.2 Mb are being discovered with increasing frequency from diverse ecosystems, with many of them now being subject to genome sequencing analysis (La Scola *et al.*, 2010; Van Etten *et al.*, 2010). The double-stranded DNA (dsDNA) genomes of these giant viruses (often called 'giruses' (Claverie *et al.*, 2006; Claverie and Ogata, 2009)) show a high coding potential with more than several hundred of genes densely packed in their genomes. To investigate the presence of Mimivirus-like *mutS* gene in other giruses, we have undertaken a genomic sequencing survey of four giruses previously isolated from marine environments. The four giruses investigated are *Pyramimonas orientalis* virus (PoV-01B, 560-kb genome), *Phaeocystis pouchetii* virus (PpV-01B, 485-kb genome), *Chrysochromulina ericina* virus (CeV-01B, 510-kb genome) and *Heterocapsa circularisquama* virus (HcDNAV, 356-kb genome) (Jacobsen *et al.*, 1996; Sandaa *et al.*, 2001; Tarutani *et al.*, 2001). The hosts of these viruses are phylogenetically distant and ecologically distinct unicellular marine algae. *C. ericina* and *P. pouchetii* are haptophytes classified in different orders, that is, *Prymnesiales* and *Phaeocystales*, respectively. *C. ericina* has a worldwide distribution; it occurs most commonly in low numbers but has been observed to form blooms together with other *Chrysochromulina* species (Simonsen and Moestrup, 1997). *P. pouchetii* may both be a free-swimming flagellated cell and a non-flagellated cell embedded in gelatinous colonies that form dense blooms in polar and sub-polar regions. *P. orientalis* is a non-blooming prasinophyte belonging to the green algae (Chlorophyta). *H. circularisquama* is a small thecate dinoflagellate which frequently forms large-scale red tides in Japan causing mass mortality of shellfish (Tarutani *et al.*, 2001). These four giruses are all lytic viruses belonging to the nucleo-cytoplasmic large DNA virus (NCLDV) superfamily (Yutin *et al.*, 2009). Phylogenetic analysis of DNA polymerase and major capsid protein sequences has revealed that PoV, PpV and CeV form a monophyletic clade that clusters together with Mimivirus (Larsen *et al.*, 2008; Monier *et al.*, 2008). In contrast, the DNA polymerase sequence of HcDNAV has been found to be closely related to that of African swine fever virus (Ogata *et al.*, 2009), suggesting that HcDNAV is phylogenetically distant from the other viruses included in this study.

With the accumulation of genome sequences and the following phylogenetic studies during the last decade, a significant advance has been made regarding the classification of diverse MutS proteins (Eisen, 1998; Lin *et al.*, 2007). In this study, we use the following naming of the MutS subfamilies, which is adapted from the recent study by Lin *et al.* (Lin *et al.*, 2007). In total, 12 subfamilies were previously described to compose the MutS family: 'MutS1/MSH1' including *E. coli* MutS and the mitochondria-targeted fungal MutS homolog 1 (MSH1); 'MutS2', known to inhibit recombination in *H. pylori* (Pinto *et al.*, 2005) and to possess a *C*-terminal endonuclease domain called the small MutS-related (Smr) domain (Moreira and Philippe, 1999; Fukui *et al.*, 2008); 'MSH2', 'MSH3', 'MSH4', 'MSH5' and 'MSH6/7', found in most eukaryotes (with the exception of MSH7 being a plant-specific paralogous group of MSH6 (Wu *et al.*, 2003)); another plant-specific MSH1 (called 'plt-MSH1' hereafter) with the

GIY-YIG endonuclease domain at their *C*-terminus (Abdelnoor *et al.*, 2006); 'MutS3', 'MutS4' and 'MutS5', recently described but functionally uncharacterized prokaryotic homologs (Lin *et al.*, 2007), and the above mentioned 'MutS7' subfamily represented by the Mimivirus MutS homolog.

In this report, we analyze the MutS sequences newly identified in four giruses, and assess the abundance of their homologs in environmental sequence databases.

Materials and methods

Girus MutS sequences

As part of an ongoing genome-sequencing project, we obtained assembled contigs for three previously isolated dsDNA viruses, *Pyramimonas orientalis* virus (PoV-01B, 141 contigs), *Phaeocystis pouchetii* virus (PpV-01B, 287 contigs) and *Chrysochromulina ericina* virus (234 contigs, CeV-01B) (Larsen *et al.*, 2008). This sequence information will be published elsewhere. In this study presented here, we scanned these girus contigs for the presence of MutS homologs. Two complete MutS ORFs were readily identified in each of the PoV and PpV contigs. Part of a contig corresponding to the CeV MutS ORF was targeted for PCR amplifications using overlapping sets of primers and re-sequenced to resolve ambiguities in the contig (see Supplementary Table S1). A fragmented ORF for the HcDNAV MutS was identified in the previously-described low coverage shotgun sequencing data (Ogata *et al.*, 2009). We obtained a complete ORF for the HcDNAV MutS after several trials of TAIL-PCR and sequencing. These sequences were submitted to public DNA databases (DDBJ: AB587728; EMBL: FR691705-FR691709). The MutS sequence from CroV (crov486, YP_003970119), that only became recently available, was partially included in our analysis.

Bioinformatics analysis

Reference MutS sequences except the girus sequences determined in this study were retrieved from the UniProt protein sequence database (as of April 27, 2010) (UniProtConsortium, 2010). The selection of sequences was performed to maximize the coverage of diverse MutS subfamilies, referring to previous publications (Eisen, 1998; Lin *et al.*, 2007), through iterative process involving clustering by BLASTCLUST (Altschul *et al.*, 1997), inspection of sequence alignment and phylogenetic reconstruction. We used T-Coffee version 8.06 (Notredame *et al.*, 2000) for multiple sequence alignment. We used ClustalX (Larkin *et al.*, 2007) for the visualization of alignments. All gap-containing sites were removed from the alignments for the following phylogenetic analyses. Maximum-likelihood phylogenetic analyses were performed using PhyML version 3.0 (Guindon and Gascuel, 2003) using LG substitution matrix (Le and Gascuel, 2008) and a gamma low (four rate categories). We used ProtTest version 2.2 (Abascal *et al.*, 2005) to determine the best substitution model (that is, LG) for our phylogenetic reconstruction based on the MutS domain V sequences. Phylogenetic trees were drawn using MEGA version 4 (Kumar *et al.*, 2008). For the delineation of the sequence domains, we used HMMER/HMMSEARCH version 2.3.2 (Eddy, 1996) and PSI-BLAST (Altschul *et al.*, 1997). The assignment of environmental sequences on the MutS7 and MutS8 subtrees was performed using a maximum-likelihood method implemented in the 'phylogenetic placement' software developed by Matsen *et al.* (pplacer version 1.0; http://matsen.fhcrc.org/pplacer/). The results were visualized using Archaeopteryx version 0.957 (http://www.phylosoft.org/archaeopteryx/) (Han and Zmasek, 2009). Correspondence analysis of codon usages was performed using CodonW version 1.3 (http://codonw.sourceforge.net/).

Results

Two types of MutS homologs in giruses

We identified six ORFs similar to known MutS family proteins in the analyzed viral genomic sequences. These ORFs were classified into two groups according to their length and sequence similarity. The first group of ORFs was relatively long and was found in all the analyzed giruses (PoV, 910 amino-acid residues (aa); PpV, 1004 aa; CeV, 1043 aa; HcDNAV, 953 aa). When searched against the NCBI non-redundant protein sequence database using BLAST (Altschul *et al.*, 1997), these girus MutS homologs showed the most significant sequence similarities to the MutS7 homologs in Mimivirus (amino-acid sequence identity 31–38%; alignment coverage 34–99%; E-value $= 10^{-63} \sim 10^{-120}$), ε-*proteobacteria* (29–37%; 95–99%; $10^{-100} \sim 10^{-153}$) and octocorals (26–28%; 96–99%; $10^{-67} \sim 10^{-84}$). Like previously reported MutS7 homologs, these four predicted proteins were found to possess a *C*-terminal HNH endonuclease domain (Supplementary Figure S1). The second group of shorter ORFs similar to MutS proteins was found in PoV (539 aa) and PpV (600 aa). These PoV and PpV MutS homologs showed the most significant sequence similarity in '*Candidatus* Amoebophilus asiaticus' (Aasi_0916; amino-acid sequence identity 38%; alignment coverage 39%; E-value $= 2 \times 10^{-26}$) and *Clostridium perfringens* (YP_694765.1; 34%; 32%; 2×10^{-17}), respectively.

Girus MutS homologs correspond to two distinct subfamilies

To classify the newly identified girus MutS homologs, we compiled a reference sequence set containing 150 MutS homologs, representing diverse MutS subfamilies, and performed phylogenetic analyses.

Our analyses revealed 15 distinct clades, 12 of which corresponded to the previously described MutS subfamilies (Figure 1a and Supplementary Figure S2). The newly identified four girus MutS homologs of the first group (that is, those with longer amino-acid sequences) were found within the MutS7 group (Figure 1b). The other MutS homologs of the second group (with shorter amino-acid sequences) were grouped in none of the previously documented subfamilies but with two paralogous sequences from 'Amoebophilus asiaticus' (Figure 1c). This bacterium is an obligate intracellular amoeba symbiont belonging to the *Bacteroidetes*. We use MutS8 to denote this new group of MutS homologs. In addition, we identified two previously undescribed subfamilies found only in bacteria. These subfamilies are referred to as MutS6 and MutS9.

Next, we determined the sequence domain architecture of MutS subfamilies with the use of position-specific scoring matrices corresponding to eight domains known to be present in MutS homologs (Figure 2 and Supplementary Figure S3). Sequence length and domain architecture were found to be comparable within individual MutS subfamilies but could differ greatly across subfamilies. Outside of these identified domains, no residual similarity was found between different subfamilies (BLAST E-value $<10^{-5}$), corroborating the classification of MutS proteins based on our phylogenetic analysis.

MutS7 was found to contain at least five known domains including the *N*-terminal MutS domain I. The domain I of bacterial MutS1 is known to directly interact with and recognize mismatched bases. The mismatch recognition by the domain I involves a phenylalanine residue (Phe 36 in *E. coli*) and a glutamic acid residue (Glu 38 in *E. coli*) in a conserved motif 'FXE' within this domain (Natrajan *et al.*, 2003). The MutS domain I is also present in the eukaryotic MSH1, MSH2, MSH3, MSH6 and plt-MSH1 subfamilies. They also exhibit conserved residues at the same location, albeit with different patterns from 'FXE' for MSH2 and MSH3 (Culligan *et al.*, 2000). Remarkably, all the members of MutS7 sequences were found to show the conserved 'FXE' motif (that is, 'FYE' for Mimivirus, HcDNAV and octocorals; 'FHE' for CroV; 'FFE' for PoV, PpV, CeV and ε-*proteobacteria*) (Supplementary Figure S4). This suggests that MutS7 may be involved in MMR rather than DNA recombination. We noted that the Mimivirus *mutS* gene showed the same intermediate expression pattern as other genes involved in

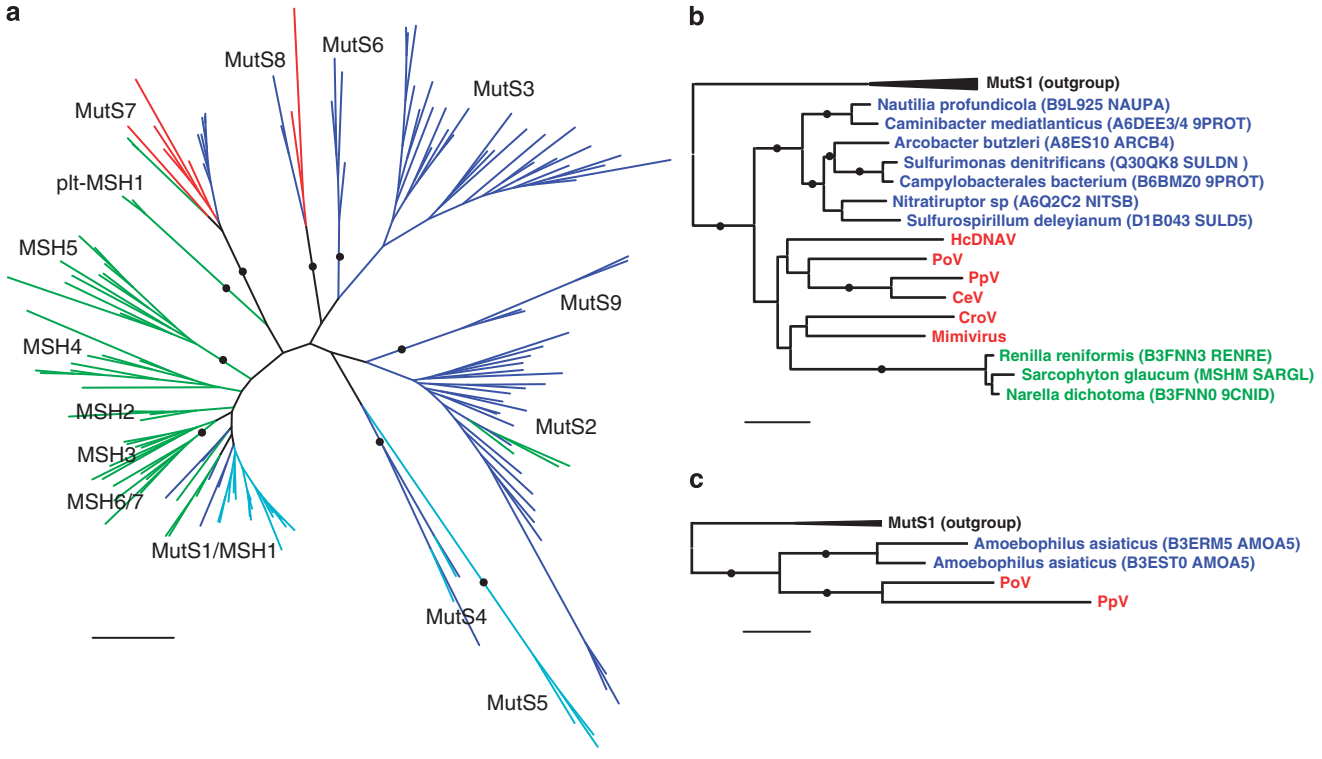


Figure 1 Maximum-likelihood phylogenetic tree of MutS family proteins. (**a**) Phylogenetic tree covering diverse MutS subfamilies including the newly identified MutS6, MutS8 and MutS9. The tree is based on the alignment of the MutS domain V sequences. (**b**) Phylogenetic tree of MutS7 homologs based on the conserved sequences between MutS1 and MutS7. (**c**) Phylogenetic tree of MutS8 homologs based on the conserved sequences between MutS1 and MutS8. The trees in the panel **b** and **c** are rooted with MutS1 sequences as the outgroup. Statistically supported branches are indicated by black dots if bootstrap values are >75%. Color code for branches and sequence names are as follows: Bacteria (blue), Archaea (light blue), Eukaryotes (green), Giruses (Red). Scale bars correspond to 0.5 substitutions per site. The recently described CroV MutS was only included in **b**.

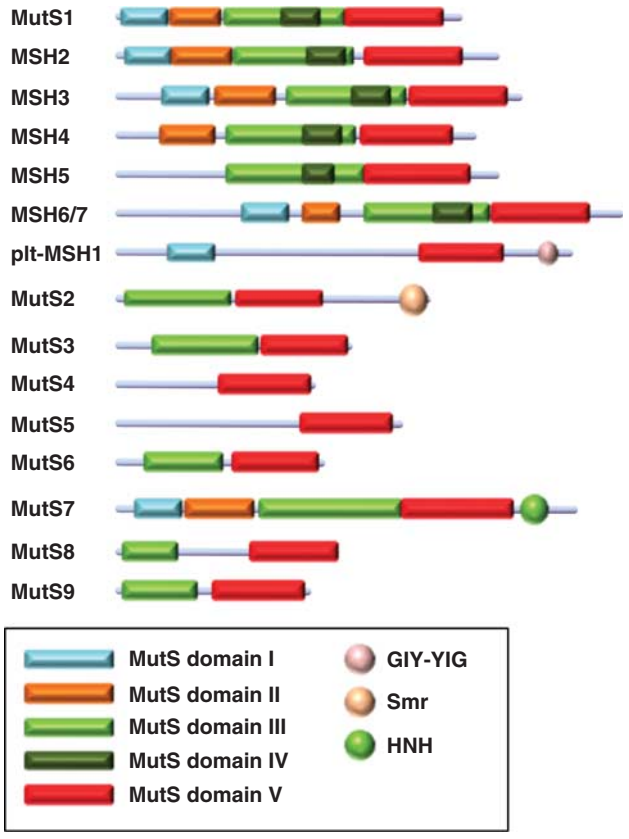


Figure 2 Domain architecture of MutS family proteins. The drawing represents the typical sequence domain organizations of MutS subfamilies (approximately scaled). A larger set of sequences is depicted in Supplementary Figure S3. Position-specific scoring matrices used for the delineation of sequence domains are as follows: MutS domain I (pfam01624), II (pfam05188), III (pfam05192), IV (pfam05190), V (pfam00488), GIY-YIG endonuclease (pfam01541), Smr (pfam01713) and HNH-endonuclease (pfam01844).

DNA replication (with the highest level of expression between 3 and 5 h after infection) (Legendre *et al*., 2010). The newly identified MutS8, MutS6 and MutS9 lacked the MutS domain I but they possess the domain III and V. A similar domain configuration can be seen in the members of the previously described MutS3 subfamily of unknown function.

MutS7 and MutS8 are abundant in marine metagenomic sequence data sets

We next used the 150 reference MutS sequences to assess the abundance of the MutS subfamilies in a standard protein sequence database (that is, UniProt), as well as in an environmental sequence collection (that is, NCBI/Env_Nr) using BLAST. We first collected MutS homologs from UniProt with the use of a position-specific scoring matrices corresponding to the MutS domain V sequences extracted from the reference sequence set. This resulted in a set of 4028 MutS homologs including the six MutS

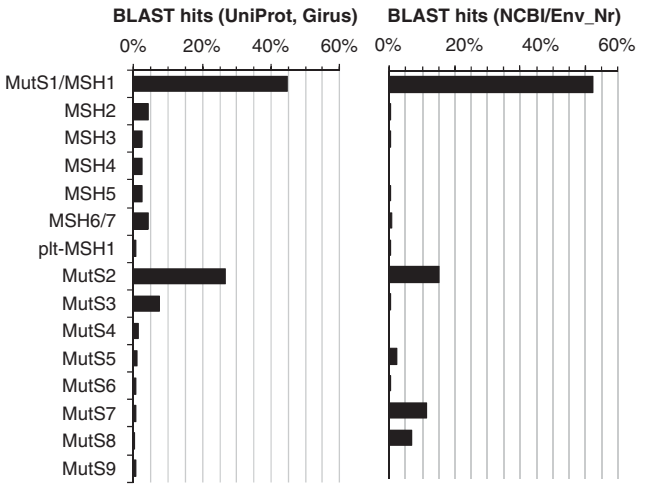


Figure 3 Representation of the different MutS subfamilies in the curated UniProt database (left panel) versus the environmental sequence data set, NCBI/Env_Nr (right panel).

homologs from PoV, PpV, CeV and HcDNAV. These 4028 sequences were searched against the 150 reference sequences with BLASTP (E-value $<10^{-5}$), and best hits were used for subfamily assignment. The relative abundance of the predicted subfamilies is shown in Figure 3 and Supplementary Table S2. Being consistent with their ubiquitous presence in prokaryotes, the most abundant subfamily was the MutS1/MSH1 subfamily (45%), which was followed by MutS2 representing 27% of MutS homologs in UniProt. Each of the remaining 13 subfamilies accounted for less than 5% of the total MutS subfamily assignments. The two subfamilies, MutS7 and MutS8, containing viral homologs were ranked at twelfth (0.7%) and fifteenth (0.1%), respectively. This analysis also confirmed the presence of MutS7 exclusively in giruses, the ε-*Proteobacteria* and octocoral mitochondria. The MutS8 subfamily was found to contain only PpV, PoV and 'Amoebophilus asiaticus' sequences. MutS6 was found exclusively in the *Bacteroidetes* (*Bacteroides*, *Chitinophaga*, *Dyadobacter*, *Pedobacter*, *Sphingobacterium*). MutS9 was found in the *Bacteroidetes*, *Firmicutes* (*Clostridia*), *Fusobacteria*, *Thermotogae* and '*Candidatus* Cloacamonas (candidate division WWE1)'. Eukaryotic MutS sequences were found in nine subfamilies (that is, MutS1/MSH1, MSH2, MSH3, MSH4, MSH5, MSH6/7, plt-MSH1, MutS2, MutS7). Bacterial sequences were present in eight subfamilies (that is, MutS1/MSH1, MutS2, MutS3, MutS4, MutS6, MutS7, MutS8, MutS9). Archaeal MutS sequences were found in three subfamilies (that is, MutS1/MSH1, MutS4, MutS5).

As the current database is highly biased towards model organisms that have been cultured and targeted for genomic analysis, we applied the same procedure to an environmental protein sequence data set (NCBI/Env_Nr) to reduce such a bias.

The position-specific scoring matrices corresponding to the MutS domain V identified 1568 MutS homologs in NCBI/Env_Nr. The subfamily assignments of these environmental sequences are shown in Figure 3 and Supplementary Table S2. Again MutS1/MSH1 (62%) and MutS2 (15%) subfamilies were the most highly represented groups. However, the MutS7 and MutS8 subfamilies, which include giral MutS homologs, were now ranked at third (176 environmental protein sequences; 11%) and fourth (106 environmental protein sequences; 7%), respectively. Each of the remaining 11 subfamilies accounted for less than 2% of the total assignments. The environmental protein sequences classified in MutS7 or MutS8 were all from a marine microbial metagenomic study, the global ocean sampling expedition (GOS) (Rusch *et al.*, 2007). The GOS reads associated with these protein sequences (441 reads for MutS7; 262 reads for MutS8) were found to originate in different geographical sampling sites (38 sites for MutS7; 35 sites for MutS8; Supplementary Table S3). Therefore, the MutS7 and MutS8 subfamily members are relatively abundant in marine microbial communities, and presently underrepresented in the curated sequence database (that is, UniProt).

Environmental MutS7 and MutS8 are likely of 'girus-origin'

An inspection of the BLAST results of the MutS7-like or MutS8-like environmental sequences immediately suggests that most of them are likely of girus origin. Of the 176 environmental MutS7 homologs, 152 (86%) sequences showed their BLAST best hit to girus MutS7 sequences (79 sequences to CeV; 48 to PoV; 18 to PpV; 5 to HcDNAV; 2 to Mimivirus). The remaining 24 sequences showed best hit to MutS7 sequences from ε-*proteobacteria*. There was no environmental sequence having a best hit to the octocoral MutS7 group. Of the 106 environmental MutS8 homologs, 95 (89%) sequences showed their best hit to girus MutS8 (69 sequences to PpV; 26 to PoV). The remaining 11 sequences best matched to 'Amoebophilus asiaticus'. To verify the evolutionary relatedness between the environmental sequences and girus MutS homologs, we used a maximum-likelihood method implemented in the 'phylogenetic placement' software developed by Matsen *et al.* (pplacer; http://matsen.fhcrc.org/pplacer/). Again, a majority (88% for MutS7 and 96% for MutS8) of the environmental sequences were positioned on the branches leading to giruses in the reference MutS7 and MutS8 phylogenetic trees (Figure 4). Finally, we compared the nucleotide compositions of these MutS homologs. Most of the *mutS7* and *mutS8* genes were found to be A+T-rich (girus-MutS7: A+T=64–82%; ε-*proteobacteria*-MutS7: 58–73%; octocoral-MutS7: 74–78%; girus-MutS8: 64–74%; Amoebophilus-MutS8: 64–66%). The environmental sequences assigned to these subfamilies were also found to be A+T-rich in average: 69% for MutS7 and 71% for MutS8. Despite this similarity in nucleotide composition, however, a correspondence analysis of the codon usages revealed that a large proportion of environmental sequences showed codon usages close to those of girus sequences for both MutS7 and MutS8 (Supplementary Figure S5). Overall, these results suggest that most of the MutS7 and MutS8 homologs in the GOS metagenomic data set probably belong to marine giruses.

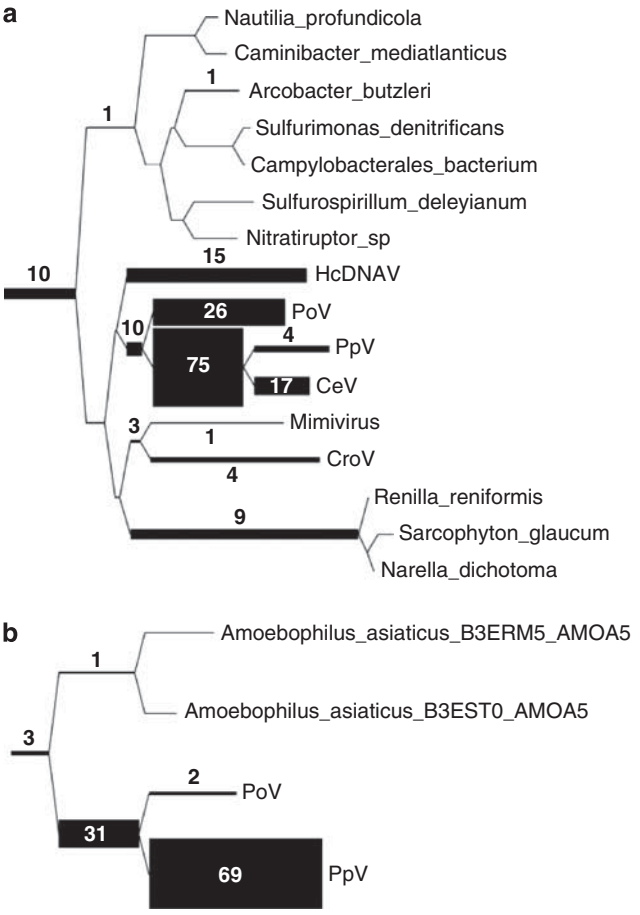


Figure 4 Taxonomic placement of the environmental MutS7 (**a**) and MutS8 (**b**) homologs. The number of environmental sequences mapped on each branch is indicated. The width of the branch is proportional to the number of mapped sequences.

Discussion

The recent accumulation of genomic and metagenomic sequence data revolutionized our understanding of the diversity and evolution of genes in microorganisms. With over 1000 sequenced genomes from cells and over 1500 genomes from DNA viruses, the available sequence data now cover a wide spectrum of species, which have already helped advancing our understanding of the functions

and evolution of protein families such as the MutS family (Eisen, 1998; Lin *et al.*, 2007). However, given the huge diversity of girus genomes (Ogata and Claverie, 2007), they seem to be still underrepresented in this sequencing effort (Claverie *et al.*, 2006; Claverie and Abergel, 2010); out of the 1500 available viral genomes, only a handful of genomes exceed 350 kb (for example, Mimivirus (1.2 Mb), CroV (730 kb), *Emiliania huxleyi* virus (407 kb), *Paramecium bursaria* Chlorella virus NY2A (369 kb), Marseillevirus (368 kb), Canarypox virus (360 kb)). In this study, we analyzed four distantly related marine giruses representing a relatively large class of giruses with estimated genome size from 356 kb up to 560 kb and identified new MutS homologs in all of the four giruses.

We showed that these girus-encoded MutS proteins fell into two subfamilies: MutS7 and MutS8. The recently reported MutS sequence from the largest marine girus, CroV, was classified in the MutS7 subfamily (Figure 1b) and was found to share the typical domain organization of this subfamily. Most unexpectedly, close homologs of the girus-encoded MutS7 and MutS8 were found to be highly abundant in marine metagenomic sequence data sets. Giruses thus seem to represent one of the major sources of the diversity of MutS family proteins. Our phylogenetic reconstruction strongly suggests the occurrence of horizontal gene transfers between giruses and cellular organisms for both the MutS7 and MutS8 subfamilies. The abundance of 'girus-like' MutS7 in the marine environment favors the previously proposed scenario that an ancestor of marine giruses had a central role in transferring MutS7 to the octocoral mitochondrial genome. Consistently, the branch to the octocoral MutS7 sequences was placed within the girus clade in the MutS7 phylogenetic tree (Figure 1b). The self-contained nature of the *mutS7* gene (with both recognition and cutting functions) might have facilitated such a gene transfer between distantly related organisms. Similar gene transfer from a virus to the ancestor of mitochondria has been proposed for the mitochondrial RNA/DNA polymerases and DNA primase; in this case, the source is likely to be a cryptic prophage (related to T3/T7) and the mitochondrial enzymes are encoded in the nuclear genome (Filee and Forterre, 2005). The possible gene transfer for MutS8 (found in PoV, PpV and the obligate intracellular amoeba-symbiont 'Amoebophilus asiaticus' isolated from lake sediment) reinforces the previously proposed idea that amoebae (or other phagocytic protists) function as 'genetic melting pots' to enhance the evolution of intracellular bacteria and viruses infecting these eukaryotes by providing ample opportunities for gene exchanges (Ogata *et al.*, 2006). Given the apparent specificity of virally encoded MutS for viruses with the largest genomes, these MutS sequences will be useful to probe metagenomic sequences for the presence of unknown giruses.

DNA viruses show a tremendous variation in genome size from a few kilobases for the oncogenic polyomaviruses, to more than a megabase for the giant Mimivirus (Monier *et al.*, 2007). Drake's rule states that the mutation rate per genome per strand copying is roughly constant across DNA-based microorganisms including bacteria, unicellular eukaryotes and DNA viruses (Drake, 1991; Sanjuan *et al.*, 2010). Mutation rate per nucleotide per replication is thus negatively correlated with the genome size. In fact, the loss of DNA repair functions is a common trend in bacterial with reduced genomes, which exhibit higher mutation rate than other bacteria with larger genomes (Moran and Wernegreen, 2000; Moran *et al.*, 2009). Experimental data for mutation rate is currently unavailable for giruses. However, given the large amount of coding DNA that they need to protect from mutations, giruses may be under a specific selective pressure for efficient DNA repair systems (such as MutS7), which may be less crucial for smaller viruses. The identification of MutS homologs in all of the four giruses tested in this study, as well as a wealth of other DNA repair genes in Mimivirus and CroV are consistent with this view.

Although the organisms with MutS7 or MutS8 had many opportunities to exchange these genes (Claverie *et al.*, 2009), the reason for the sporadic and limited phyletic distribution of these MutS subfamilies still remains unclear. One might presume that the functions of these MutS proteins are somehow associated with A+T-rich genomes. However, the presence of ε-*proteobacteria* with A+T-rich genomes (such as *H. pylori*, A+T=62%) lacking these MutS subfamily members contradicts this hypothesis. *E. coli* MutH distinguishes the nascent DNA strand from the template DNA strand through the hemi-methylation of bases. It would be interesting to examine the presence of hemi-methylated bases in girus genomes and octocoral mitochondrial genomes. We have started to clone and purify the Mimivirus MutS7 for functional characterization.

The unique presence of a MutS homolog in Mimivirus was already noticed during the initial genome annotation (Raoult *et al.*, 2004). We then recognized the surprising relationship between the Mimivirus MutS and its homologs uniquely found in the mitochondria of all octocorals (Claverie *et al.*, 2009), all belonging to the newly defined MutS7 subfamily. The finding of these MutS homologs in CroV, CeV, PoV, PpV and HcDNAV definitely confirms their association with large DNA viruses in marine environments. These findings strongly suggest that the presence of MutS in Mimivirus is not merely an example of an eccentric lateral gene transfer, but probably requires a more subtle explanation. We believe that much deeper experimental investigation of these girus MutS homologs would help provide a holistic view on the evolution of gene families in the light of evolutionary interactions between the viral and cellular gene pools.

Acknowledgements

We thank Dr Stéphane Audic for his technical assistance in an early stage of this work. The IGS laboratory is supported, in part, by CNRS and the French National Research Agency (Grant # ANR-09-PCS-GENM-218, ANR-08-BDVA-003). The FRA laboratory is partially supported by Grants-in-Aid for Scientific Research (A) (No. 20247002) from the Ministry of Education, Science and Culture of Japan. The University of Bergen received financial support from the Norwegian Research Council for research programmes 'Viral lysis and programmed cell death in marine phytoplankton' (VIPMAP, 186142/V40) and 'Diversity and dynamics of marine Haptophytes' (HAPTODIV, 190307/S40).'

References

Abascal F, Zardoya R, Posada D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104–2105.

Abdelnoor RV, Christensen AC, Mohammed S, Munoz-Castillo B, Moriyama H, Mackenzie SA. (2006). Mitochondrial genome dynamics in plants and animals: convergent gene fusions of a MutS homologue. *J Mol Evol* **63**: 165–173.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Bogani F, Corredeira I, Fernandez V, Sattler U, Rutvisuttinunt W, Defais M *et al.* (2010). Association between the Herpes Simplex Virus-1 DNA Polymerase and Uracil DNA Glycosylase. *J Biol Chem* **285**: 27664–27672.

Brugler MR, France SC. (2008). The mitochondrial genome of a deep-sea bamboo coral (Cnidaria, Anthozoa, Octocorallia, Isididae): genome structure and putative origins of replication are not conserved among octocorals. *J Mol Evol* **67**: 125–136.

Claverie JM, Abergel C. (2010). Mimivirus: the emerging paradox of quasi-autonomous viruses. *Trends Genet* **26**: 431–437.

Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J *et al.* (2009). Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. *J Invertebr Pathol* **101**: 172–180.

Claverie JM, Ogata H. (2009). Ten good reasons not to exclude giruses from the evolutionary picture. *Nat Rev Microbiol* **7**: 615; author reply 615.

Claverie JM, Ogata H, Audic S, Abergel C, Suhre K, Fournier PE. (2006). Mimivirus and the emerging concept of 'giant' virus. *Virus Res* **117**: 133–144.

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

Drake JW. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* **88**: 7160–7164.

Eddy SR. (1996). Hidden Markov models. *Curr Opin Struct Biol* **6**: 361–365.

Eisen JA. (1998). A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res* **26**: 4291–4300.

Filee J, Forterre P. (2005). Viral proteins functioning in organelles: a cryptic origin? *Trends Microbiol* **13**: 510–513.

Fischer MG, Allen MJ, Wilson WH, Suttle CA. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA* **107**: 19508–19513.

Fukui K, Nakagawa N, Kitamura Y, Nishida Y, Masui R, Kuramitsu S. (2008). Crystal structure of MutS2 endonuclease domain and the mechanism of homologous recombination suppression. *J Biol Chem* **283**: 33417–33427.

Furuta M, Schrader JO, Schrader HS, Kokjohn TA, Nyaga S, McCullough AK *et al.* (1997). Chlorella virus PBCV-1 encodes a homolog of the bacteriophage T4 UV damage repair gene denV. *Appl Environ Microbiol* **63**: 1551–1556.

Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Han MV, Zmasek CM. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10**: 356.

Iyer RR, Pluciennik A, Burdett V, Modrich PL. (2006). DNA mismatch repair: functions and mechanisms. *Chem Rev* **106**: 302–323.

Jacobsen A, Bratbak G, Heldal M. (1996). Isolation and characterization of a virus infecting Phaeocystis pouchetii (Prymnesiophyceae). *J Phycol* **32**: 923–927.

Kumar S, Nei M, Dudley J, Tamura K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* **9**: 299–306.

La Scola B, Campocasso A, N'Dong R, Fournous G, Barrassi L, Flaudrops C *et al.* (2010). Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. *Intervirology* **53**: 344–353.

La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G *et al.* (2008). The virophage as a unique parasite of the giant mimivirus. *Nature* **455**: 100–104.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Larsen JB, Larsen A, Bratbak G, Sandaa RA. (2008). Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. *Appl Environ Microbiol* **74**: 3048–3057.

Le SQ, Gascuel O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol* **25**: 1307–1320.

Legendre M, Audic S, Poirot O, Hingamp P, Seltzer V, Byrne D *et al.* (2010). mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res* **20**: 664–674.

Lin Z, Nei M, Ma H. (2007). The origins and early evolution of DNA mismatch repair genes–multiple horizontal gene transfers and co-evolution. *Nucleic Acids Res* **35**: 7591–7603.

Malik HS, Henikoff S. (2000). Dual recognition-incision enzymes might be involved in mismatch repair and meiosis. *Trends Biochem Sci* **25**: 414–418.

McFadden CS, France SC, Sanchez JA, Alderslade P. (2006). A molecular phylogenetic analysis of the Octocorallia (Cnidaria: Anthozoa) based on mitochondrial protein-coding sequences. *Mol Phylogenet Evol* **41**: 513–527.

Miller WG, Parker CT, Rubenfield M, Mendz GL, Wosten MM, Ussery DW *et al.* (2007). The complete genome sequence and analysis of the epsilonproteobacterium Arcobacter butzleri. *PLoS ONE* **2**: e1358.

Monier A, Claverie JM, Ogata H. (2007). Horizontal gene transfer and nucleotide compositional anomaly in large DNA viruses. *BMC Genomics* **8**: 456.

Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, Ogata H. (2008). Marine mimivirus relatives are probably large algal viruses. *Virol J* **5**: 12.

Moran NA, McLaughlin HJ, Sorek R. (2009). The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* **323**: 379–382.

Moran NA, Wernegreen JJ. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol Evol* **15**: 321–326.

Moreira D, Philippe H. (1999). Smr: a bacterial and eukaryotic homologue of the C-terminal region of the MutS2 family. *Trends Biochem Sci* **24**: 298–300.

Nakagawa S, Takaki Y, Shimamura S, Reysenbach AL, Takai K, Horikoshi K. (2007). Deep-sea vent epsilonproteobacterial genomes provide insights into emergence of pathogens. *Proc Natl Acad Sci USA* **104**: 12146–12150.

Natrajan G, Lamers MH, Enzlin JH, Winterwerp HH, Perrakis A, Sixma TK. (2003). Structures of Escherichia coli DNA mismatch repair enzyme MutS in complex with different mismatches: a common recognition mode for diverse substrates. *Nucleic Acids Res* **31**: 4814–4821.

Notredame C, Higgins DG, Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205–217.

Ogata H, Claverie JM. (2007). Unique genes in giant viruses: regular substitution pattern and anomalously short size. *Genome Res* **17**: 1353–1361.

Ogata H, La Scola B, Audic S, Renesto P, Blanc G, Robert C *et al.* (2006). Genome sequence of Rickettsia bellii illuminates the role of amoebae in gene exchanges between intracellular pathogens. *PLoS Genet* **2**: e76.

Ogata H, Toyoda K, Tomaru Y, Nakayama N, Shirai Y, Claverie JM *et al.* (2009). Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. *Virol J* **6**: 178.

Pinto AV, Mathieu A, Marsin S, Veaute X, Ielpi L, Labigne A *et al.* (2005). Suppression of homologous and homeologous recombination by the bacterial MutS2 protein. *Mol Cell* **17**: 113–120.

Pont-Kingdon GA, Okada NA, Macfarlane JL, Beagley CT, Wolstenholme DR, Cavalier-Smith T *et al.* (1995). A coral mitochondrial mutS gene. *Nature* **375**: 109–111.

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H *et al.* (2004). The 1.2-megabase genome sequence of Mimivirus. *Science* **306**: 1344–1350.

Redrejo-Rodriguez M, Ishchenko AA, Saparbaev MK, Salas ML, Salas J. (2009). African swine fever virus AP endonuclease is a redox-sensitive enzyme that repairs alkylating and oxidative damage to DNA. *Virology* **390**: 102–109.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Sandaa RA, Heldal M, Castberg T, Thyrhaug R, Bratbak G. (2001). Isolation and characterization of two viruses with large genome size infecting Chrysochromulina ericina (Prymnesiophyceae) and Pyramimonas orientalis (Prasinophyceae). *Virology* **290**: 272–280.

Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. (2010). Viral mutation rates. *J Virol* **84**: 9733–9748.

Schofield MJ, Hsieh P. (2003). DNA mismatch repair: molecular mechanisms and biological function. *Annu Rev Microbiol* **57**: 579–608.

Sievert SM, Scott KM, Klotz MG, Chain PS, Hauser LJ, Hemp J *et al.* (2008). Genome of the epsilonproteobacterial chemolithoautotroph Sulfurimonas denitrificans. *Appl Environ Microbiol* **74**: 1145–1156.

Simonsen S, Moestrup O. (1997). Toxicity tests in eight species of Chrysochromulina (Haptophyta). *Can J Bot* **75**: 129–136.

Srinivasan V, Tripathy DN. (2005). The DNA repair enzyme, CPD-photolyase restores the infectivity of UV-damaged fowlpox virus isolated from infected scabs of chickens. *Vet Microbiol* **108**: 215–223.

Tarutani K, Nagasaki K, Itakura S, Yamaguchi M. (2001). Isolation of a virus infecting the novel shellfish-killing dinoflagellate Heterocapsa circularisquama. *Aquat Microb Ecol* **23**: 103–111.

UniProtConsortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**: D142–D148.

Van Etten JL, Lane LC, Dunigan DD. (2010). DNA Viruses: The Really Big Ones (Giruses). *Annu Rev Microbiol* **64**: 83–99.

Wu SY, Culligan K, Lamers M, 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.

Yutin N, Wolf YI, Raoult D, Koonin EV. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* **6**: 223.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment

Hiroyuki Ogata[1,*], Jessica Ray[2], Kensuke Toyoda[3,†], Ruth-Anne Sandaa[2], Keizo Nagasaki[3], Gunnar Bratbak[2], Jean-Michel Claverie[1]

1: Information Génomique et Structurale, CNRS-UPR2589, Institut de Microbiologie de la Méditerranée, Parc Scientifique de Luminy, Aix-Marseille Université, 163 Avenue de Luminy, Case 934, 13288 Marseille Cedex 9, France

2: Department of Biology, University of Bergen, PO Box 7800, N-5020 Bergen, Norway

3: Harmful Algal Bloom Division, National Research Institute of Inland Sea, Fisheries Research Agency, 2-17-5 Maruishi, Hatsukaichi, Hiroshima 739-0452, Japan

*: Correspondence: Hiroyuki Ogata (E-mail: Hiroyuki.Ogata@igs.cnrs-mrs.fr)

†: K. Toyoda's present address is "Department of Botany, Keio University, Hiyoshi, Kohoku-ku, Yokohama, Kangawa 223-8521, Japan" (E-mail: toyoda@diatom.jp)

Supplementary materials

Table S1. Primers used to resolve sequence ambiguities in the CeV MutS7 region. Five sets of overlapping DNA primers were designed to resolve sequence ambiguities within a CeV contig containing a MutS7 homolog. Amplification reactions (50µl) contained 1X Hot Star Taq Plus Master Mix (Qiagen, Hilden, Germany), 0.4 µM each of forward and reverse primers, 0.2 mg ml-1 BSA (Promega, Madison, Wisconsin), 3 mM MgCl2 and1X Coralload buffer (Qiagen). CeV lysates were diluted 10-fold with sterile MQ water and subjected to 2 x 2min incubation at 99°C with an intervening 2min incubation on ice. Template for PCR amplification reactions consisted of 2µl of the freeze-thaw diluted CeV lysate. The thermal cycling program consisted of an initial 5min denaturation at 95°C, 30 cycles of 95°C for 30sec, 55°C for 30sec and 72°C for 90sec, and a final 10min elongation at 72°C. When necessary, PCR products were stored overnight at 4°C until analysis by agarose gel electrophoresis to confirm amplification specificity. Correct molecular weight bands were excised from agarose gels and purified using the illustra GFX PCR purification kit (GE Healthcare, Amersham, UK) according to the manufacturer's protocol. 30-40ng of purified PCR product was used as template for forward and reverse sequencing reactions using the BigDye Terminator v3.1 cycle sequencing chemistry (Applied Biosystems, Carlsbad, California) and 3.2pmol primer. Sequencing was performed at the DNA Sequencing Facility at the University of Bergen, Norway.

Primer name	**5'-3' sequence**
cev_002_181-200	TGGCCTGGGCCACCAAATCC
cev_002_1010-989	AAACAATGGGTGCGTGTCC
cev_002_697-716	TGGGGTAATCCAAATCCTGCCA
cev_002_1807-1784	ACTCACGTTTACCCATAGGTGTT
cev_002_1779-1806	TGTTTAACACCTATGGGTAAACGTGAG
cev_002_2844-2821	ACCATTAATTCATAACCATTAGTCACC
cev_002_2388-2413	ATAAGAGCGTGTCTTAGTCCCGT
cev_002_3405-3379	TTCCAGGAACTGAAAGTGATTCGGCA
cev_002_3179-3204	GCAACATTACAAGCAACAAAACGACG
cev_002_3782-3756	CCATACATACTTTCTCCAGCTCCTGGT

Table S2. MutS homologs in UniProt, girus genomes and NCBI/Env_Nr.

MutS subfamily	**UniProt and girus MutS homologs**					**NCBI EnvNr**
	Total BLAST hits	Bacteria	Archaea	Eukaryota	Virus	
MutS1/MSH1	1803	1680	38	85		976
MSH2	176			176		7
MSH3	102			102		4
MSH4	93			93		
MSH5	103			103		2
MSH6/7	174			174		11
plt-MSH1	29			29		4
MutS2	1070	1037		33		237
MutS3	304	304				6
MutS4	51	47	4			
MutS5	42		42			34
MutS6	24	24				5
MutS7	28	8		15	5*	176
MutS8	4	2			2	106
MutS9	25	25				
Total	4028	3127	84	810	7	1568

* A recently reported viral MutS7 (CroV MutS) is not included in this table.

Table S3. GOS reads related to MutS7/MutS8

ID	Habitat Type	Geographic Location	Sample Location	Sample Depth (m)	Water Depth (m)	T (oC)	S (ppt)	Size Fraction (?m)	Chl a Sample Month (Annual±SE) mg m-3	Reads	MutS7 Reads	MutS8 Reads
JCVI_SITE_GS000_S13	Open Ocean	Sargasso Sea	Sargasso Sea, Station 13	5	>4200	20	36.6	0.1-0.8	0.17 (0.09±0.02)	644551	16	8
JCVI_SITE_GS000_S13	Open Ocean	Sargasso Sea	Sargasso Sea, Station 13	5	>4200	20	36.6	0.22-0.8	0.17 (0.09±0.02)	317180	16	8
JCVI_SITE_GS000_S11	Open Ocean	Sargasso Sea	Sargasso Sea, Station 11	5	>4200	20.5	36.7	0.1-0.8	0.17 (0.09±0.02)	644551	9	0
JCVI_SITE_GS000_S11	Open Ocean	Sargasso Sea	Sargasso Sea, Station 11	5	>4200	20.5	36.7	0.22-0.8	0.17 (0.09±0.02)	317180	9	0
JCVI_SITE_GS000_S03	Open Ocean	Sargasso Sea	Sargasso Sea, Station 3	5	>4200	19.8	36.7	0.22-0.8	0.17 (0.09±0.02)	368835	4	4
JCVI_SITE_GS000_S13	Open Ocean	Sargasso Sea	Sargasso Sea, Station 13	5	>4200	20	36.6	0.22-0.8	0.17 (0.09±0.02)	332240	16	8
JCVI_SITE_GS001	Open Ocean	Sargasso Sea	Sargasso Sea, Hydrostation S	5	>4200	22.9	36.7	3.0-20	0.10 (0.10±0.01)	142352	8	4
JCVI_SITE_GS001	Open Ocean	Sargasso Sea	Sargasso Sea, Hydrostation S	5	>4200	22.9	36.7	0.8-3.0	0.10 (0.10±0.01)	90905	8	4
JCVI_SITE_GS001	Open Ocean	Sargasso Sea	Sargasso Sea, Hydrostation S	5	>4200	22.9	36.7	0.1-0.8	0.10 (0.10±0.01)	92351	8	4
JCVI_SITE_GS002	Coastal	North American East Coast	Gulf of Maine	1	106	18.2	29.2	0.1-0.8	1.4 (1.12±0.19)	121590	48	13
JCVI_SITE_GS003	Coastal	North American East Coast	Browns Bank, Gulf of Maine	1	119	11.7	29.9	0.1-0.8	1.4 (1.12±0.19)	61605	27	55
JCVI_SITE_GS004	Coastal	North American East Coast	Outside Halifax, Nova Scotia	2	142	17.3	28.3	0.1-0.8	0.4 (0.78±0.17)	52959	4	0
JCVI_SITE_GS005	Embayment	North American East Coast	Bedford Basin, Nova Scotia	1	64	15	30.2	0.1-0.8	6 (6.76±0.98)	61131	5	3
JCVI_SITE_GS006	Estuary	North American East Coast	Bay of Fundy, Nova Scotia	1	11	11.2		0.1-0.8	2.8 (1.87±0.18)	59679	15	0
JCVI_SITE_GS007	Coastal	North American East Coast	Northern Gulf of Maine	1	139	17.9	31.7	0.1-0.8	1.4 (1.12±0.19)	50980	7	4
JCVI_SITE_GS008	Coastal	North American East Coast	Newport Harbor, RI	1	12	9.4	26.5	0.1-0.8	2.2 (1.59±0.17)	129655	6	0

JCVI_SITE_GS009	Coastal	North American East Coast	Block Island, NY	1	32	11	31	0.1-0.8	4.0 (2.72±0.24)	79303	10	6
JCVI_SITE_GS010	Coastal	North American East Coast	Cape May, NJ	1	10	12	31	0.1-0.8	2.0 (2.75±0.33)	78304	4	7
JCVI_SITE_GS011	Estuary	North American East Coast	Delaware Bay, NJ	1	8	11		0.1-0.8	4.8 (9.23±1.02)	124435	22	10
JCVI_SITE_GS012	Estuary	North American East Coast	Chesapeake Bay, MD	13.2	25	1	3.5	0.1-0.8	21.0 (15.0±1.01)	126162	15	16
JCVI_SITE_GS013	Coastal	North American East Coast	Off Nags Head, NC	2.1	20	9.3		0.1-0.8	3.0 (2.24±0.25)	138033	10	6
JCVI_SITE_GS014	Coastal	North American East Coast	South of Charleston, SC	1	31	18.6		0.1-0.8	1.70 (1.92±0.25)	128885	2	4
JCVI_SITE_GS015	Coastal	Caribbean Sea	Off Key West, FL	1.7	47	25	36	0.1-0.8	0.2 (0.27±0.09)	127362	15	4
JCVI_SITE_GS016	Coastal Sea	Caribbean Sea	Gulf of Mexico	2	3333	26.4	35.8	0.1-0.8	0.16 (0.11±0.01)	127122	12	8
JCVI_SITE_GS017	Open Ocean	Caribbean Sea	Yucatan Channel	2	4513	27	35.8	0.1-0.8	0.13 (0.09±0.01)	257581	32	23
JCVI_SITE_GS018	Open Ocean	Caribbean Sea	Rosario Bank	1.7	4470	27.4	35.4	0.1-0.8	0.14 (0.09±0.01)	142743	6	2
JCVI_SITE_GS019	Coastal	Caribbean Sea	Northeast of Colón	1.7	3336	27.7	35.4	0.1-0.8	0.23 (0.15±0.02)	135325	13	2
JCVI_SITE_GS020	Fresh Water	Panama Canal	Lake Gatun	2	4.2	28.6	0.1	0.1-0.8		296355	21	11
JCVI_SITE_GS021	Coastal	Eastern Tropical Pacific	Gulf of Panama	1.6	76	27.6	30.7	0.1-0.8	0.50 (0.73±0.22)	131798	3	4
JCVI_SITE_GS022	Open Ocean	Eastern Tropical Pacific	250 miles from Panama City	2	2431	29.3	32.3	0.1-0.8	0.33 (0.28±0.02)	121662	6	4
JCVI_SITE_GS023	Open Ocean	Eastern Tropical Pacific	30 miles from Cocos Island	2	1139	28.7	32.6	0.1-0.8	0.07 (0.19±0.02)	133051	4	8
JCVI_SITE_GS025	Fringing Reef	Eastern Tropical Pacific	Dirty Rock, Cocos Island	1.1	30	28.3	31.4	0.8-3.0	0.11 (0.19±0.01)	120671	0	4
JCVI_SITE_GS026	Open Ocean	Galapagos Islands	134 miles NE of Galapagos	2	2386	27.8	32.6	0.1-0.8	0.22 (0.28±0.02)	102708	0	2
JCVI_SITE_GS027	Coastal	Galapagos Islands	Devil's Crown, Floreana Island	2.2	2.3	25.5	34.9	0.1-0.8	0.40 (0.38±0.03)	222080	16	2
JCVI_SITE_GS028	Coastal	Galapagos Islands	Coastal Floreana	2	156			0.1-0.8	0.35 (0.35±0.02)	189052	17	11

JCVI_SITE_GS029	Coastal	Galapagos Islands	North James Bay, Santigo Island	2.1	12	26.2	34.5	0.1-0.8	0.40 (0.39±0.03)	131529	4	4
JCVI_SITE_GS030	Warm Seep	Galapagos Islands	Warm seep, Roca Redonda	19	19	26.9		0.1-0.8		359152	2	2
JCVI_SITE_GS031	Coastal upwelling	Galapagos Islands	Upwelling, Fernandina Island	12	19.6	18.6		0.1-0.8	0.35 (0.39±0.03)	436401	25	12
JCVI_SITE_GS032	Mangrove	Galapagos Islands	Mangrove on Isabella Island	0.1	1.6	25.4		0.1-0.8		148018	19	2
JCVI_SITE_GS033	Hypersaline	Galapagos Islands	Punta Cormorant, Hypersaline Lagoon, Floreana Island	0.2	0.3	37.6	63.4	0.1-0.8		692255	8	3
JCVI_SITE_GS034	Coastal	Galapagos Islands	North Seamore Island	2.1	35	27.5		0.1-0.8	0.36 (0.35±0.02)	134347	12	4
JCVI_SITE_GS035	Coastal	Galapagos Islands	Wolf Island	1.7	71	21.8	34.5	0.1-0.8	0.28 (0.31±0.02)	140814	2	2
JCVI_SITE_GS036	Coastal	Galapagos Islands	Cabo Marshall, Isabella Island	2.1	67	25.8	34.6	0.1-0.8	0.65 (0.45±0.05)	77538	4	0
JCVI_SITE_GS037	Open Ocean	Eastern Tropical Pacific	Equatorial Pacific TAO Buoy	1.8	3334	28		0.1-0.8	0.21 (0.24±0.02)	65670	0	6
JCVI_SITE_GS038	Open Ocean	Tropical South Pacific	Tropical South Pacific	1.8	>4000	28.4		0.1-0.8		741	0	0
JCVI_SITE_GS039	Open Ocean	Tropical South Pacific	Tropical South Pacific	2	>4000	28.6		0.1-0.8		759	0	0
JCVI_SITE_GS040	Open Ocean	Tropical South Pacific	Tropical South Pacific	2.2	>4000	27.8		0.1-0.8		736	0	0
JCVI_SITE_GS041	Open Ocean	Tropical South Pacific	Tropical South Pacific	2	>4000	28	35	0.1-0.8		678	0	0
JCVI_SITE_GS042	Open Ocean	Tropical South Pacific	Tropical South Pacific	1.7	>4000	27.6		0.1-0.8		699	0	0
JCVI_SITE_GS043	Open Ocean	Tropical South Pacific	Tropical South Pacific	1.9	>4000	27.6	35.9	0.1-0.8		711	0	0
JCVI_SITE_GS044	Open Ocean	Tropical South Pacific	600 miles from F. Polynesia	2	>4000	27.6		0.1-0.8		678	0	0
JCVI_SITE_GS045	Open Ocean	Tropical South Pacific	400 miles from F. Polynesia	1.7	>4000	28.3	37	0.1-0.8		730	0	0
JCVI_SITE_GS046	Open Ocean	Tropical South Pacific	300 miles from F. Polynesia	1.9	>4000	28.7	35.3	0.1-0.8		626	0	0

JCVI_SITE_GS047	Open Ocean	Tropical South Pacific	201 miles from F. Polynesia	30	2400	28.6	37.3	0.1-0.8		66023	4	2
JCVI_SITE_GS048	Coral Reef	Polynesia Archipelagos	Moorea, Cooks Bay	1.4	34	28.9	35.1	0.1-0.8		744	0	0
JCVI_SITE_GS049	Coastal	Polynesia Archipelagos	Moorea, Outside Cooks Bay	1.4	900	28.8	32.6	0.1-0.8		735	0	0
JCVI_SITE_GS050	Coral Atoll	Polynesia Archipelagos	Tikehau Lagoon	1.2	24	27.8		0.1-0.8		715	0	0
JCVI_SITE_GS051	Coral Reef Atoll	Polynesia Archipelagos	Rangirora Atoll	1	10	27.3	34.2	0.1-0.8		128982	4	0

Legend for supplementary figures

Figure S1. Multiple sequence alignment of HNH endonuclease domains of the MutS7 subfamily proteins. The positions of four conserved residues around the endonuclease active site are marked by red triangles.

Figure S2 Maximum likelihood phylogenetic tree of MutS family proteins. The tree is based on the conserved MutS domain V sequences. The tree is mid-point rooted. Bootstrap values < 50% are not shown. Taxon names are composed of a MutS family name, a sequence identifier, a domain classification (B for Bacteria, A for Archaea, E for Eukaryote, V for Viruses), followed by the species name. Color code for branches are as follows: Bacteria (blue), Archaea (light blue), Eukaryotes (green), Giruses (Red). MutS subfamilies introduced in this study (MutS6, MutS7, MutS8, MutS9) are indicated in red.

Figure S3 Domain architecture of MutS family proteins. Sequence domains were identified using NCBI/Cdd profiles and PSI-BLAST (E-value<0.01). This diagram is drawn to scale. For MutS domains (I, II, III, IV, V), PSI-BLAST was used with four iterations. Identified domains were represented as follows: MutS domain I (pfam01624), light blue rectangle; MutS domain II (pfam05188), orange rectangle; MutS domain III (pfam05192), light green rectangle; MutS domain IV (pfam05190), dark green rectangle; MutS domain V (pfam00488), red rectangle; Smr domain (pfam01713), orange oval; GIY-YIG domain (pfam01541), pink oval; HNH domain (pfam01844), green oval. MSH1p corresponds to plant specific MSH1 (plt-MSH1).

Figure S4. Multiple sequence alignment of the N-terminal part of the domain I sequences from MutS7 and *E. coli* MutS1. The conserved "F(X)E" residues are highlighted by a red rectangle.

Figure S5. Correspondence analysis of codon usages of MutS7 and MutS8 homologs. The number of GOS environmental sequences is 176 for MutS7 and 106 for MutS8.

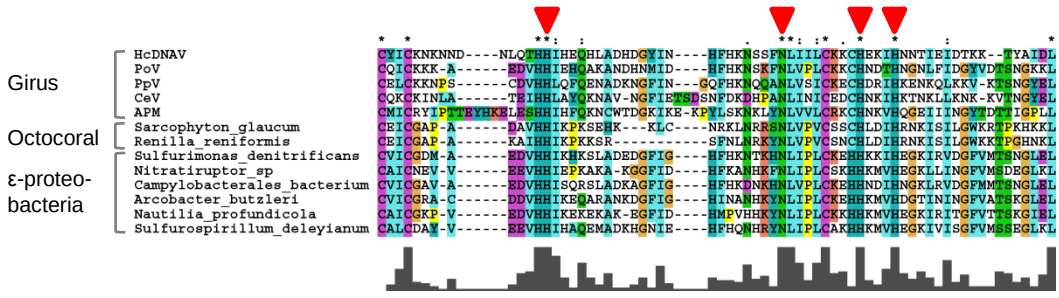

Girus
Octocoral
ε-proteo-
bacteria
HcDNAV
PoV
PpV
CeV
APM
Sarcophyton_glaucum
Renilla_reniformis
Sulfurimonas_denitrificans
Nitratiruptor_sp
Campylobacterales_bacterium
Arcobacter_butzleri
Nautilia_profundicola
Sulfurospirillum_deleyianum

Fig. S1


MutS1/MSH1 (Bacteria/Archaea/Eukaryotes)
MSH6/MSH7 (Eukaryotes)
MSH3 (Eukaryotes)
MSH2 (Eukaryotes)
MSH4 (Eukaryotes)
MSH5 (Eukaryotes)
plt-MSH1 (Plant)
MutS7 (Girus/Epsilonproteobacteria/Octocorals)
MutS8 (Girus/Amoebophilus)
MutS6 (Bacteria)
MutS3A
MutS3B
MutS3 (Bacteria)
MutS9 (Bacteria)
MutS2 (Bacteria/Eukaryotes)
MutS5 (Archaea)
MutS4B
MutS4A
MutS4 (Bacteria/Archaea)
0.5

Fig. S2

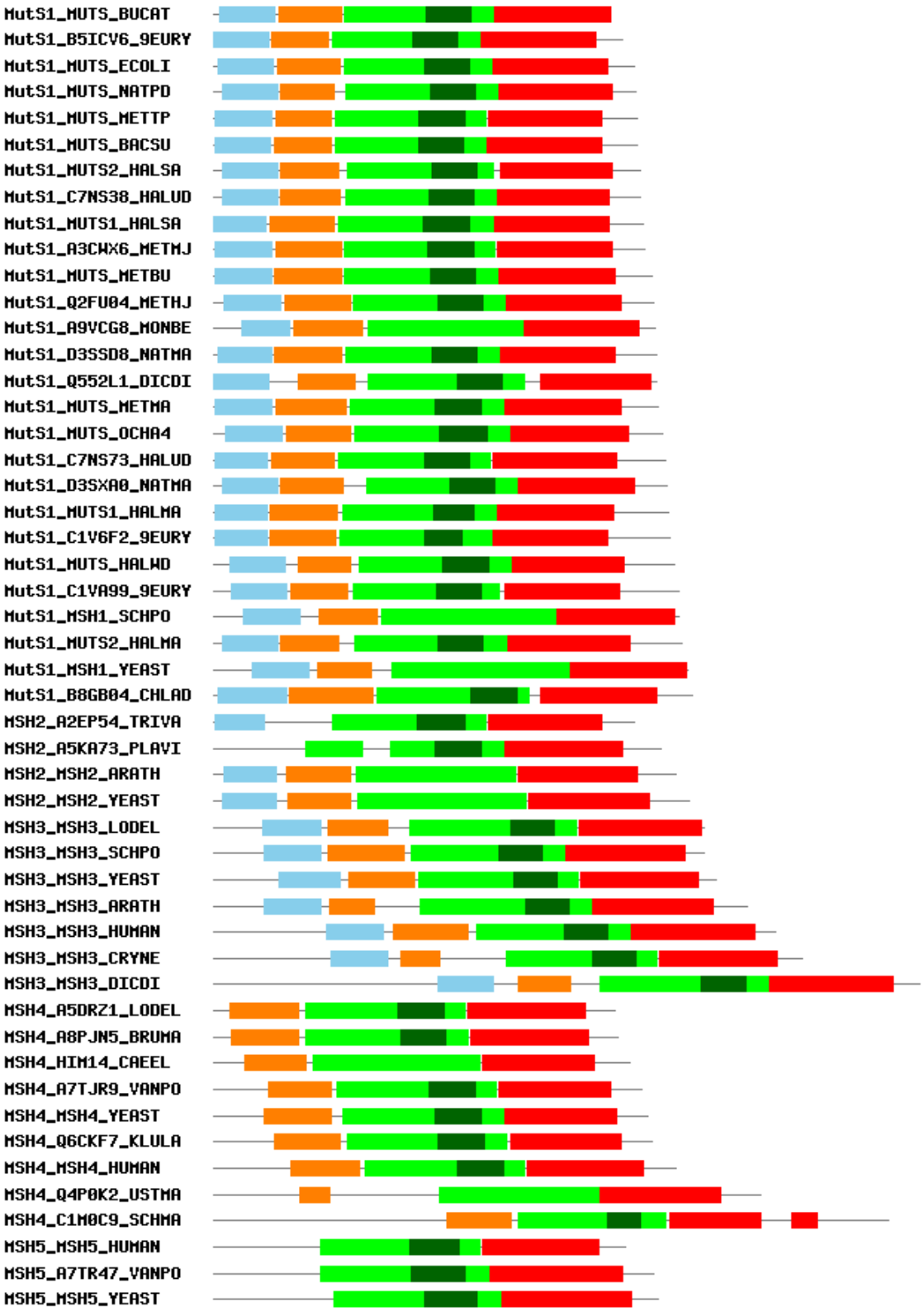

MutS1_MUTS_BUCAT
MutS1_B5ICV6_9EURY
MutS1_MUTS_ECOLI
MutS1_MUTS_NATPD
MutS1_MUTS_METTP
MutS1_MUTS_BACSU
MutS1_MUTS2_HALSA
MutS1_C7NS38_HALUD
MutS1_MUTS1_HALSA
MutS1_A3CWX6_METMJ
MutS1_MUTS_METBU
MutS1_Q2FU04_METHJ
MutS1_A9VCG8_MONBE
MutS1_D3SSD8_NATMA
MutS1_Q552L1_DICDI
MutS1_MUTS_METMA
MutS1_MUTS_OCHA4
MutS1_C7NS73_HALUD
MutS1_D3SXA0_NATMA
MutS1_MUTS1_HALMA
MutS1_C1V6F2_9EURY
MutS1_MUTS_HALWD
MutS1_C1VA99_9EURY
MutS1_MSH1_SCHPO
MutS1_MUTS2_HALMA
MutS1_MSH1_YEAST
MutS1_B8GB04_CHLAD
MSH2_A2EP54_TRIVA
MSH2_A5KA73_PLAVI
MSH2_MSH2_ARATH
MSH2_MSH2_YEAST
MSH3_MSH3_LODEL
MSH3_MSH3_SCHPO
MSH3_MSH3_YEAST
MSH3_MSH3_ARATH
MSH3_MSH3_HUMAN
MSH3_MSH3_CRYNE
MSH3_MSH3_DICDI
MSH4_A5DRZ1_LODEL
MSH4_A8PJN5_BRUMA
MSH4_HIM14_CAEEL
MSH4_A7TJR9_VANPO
MSH4_MSH4_YEAST
MSH4_Q6CKF7_KLULA
MSH4_MSH4_HUMAN
MSH4_Q4P0K2_USTMA
MSH4_C1M0C9_SCHMA
MSH5_MSH5_HUMAN
MSH5_A7TR47_VANPO
MSH5_MSH5_YEAST

Fig. S3 (1/3)


MSH5_C4M0Z9_ENTHI
MSH5_Q7Z7S7_COPCI
MSH5_Q6CT05_KLULA
MSH5_C1EEU6_9CHLO
MSH5_A7EN13_SCLS1
MSH5_MSH5_CAEEL
MSH5_Q011M5_OSTTA
MSH6_MSH7_ARATH
MSH6_A0DMV3_PARTE
MSH6_MSH6_YEAST
MSH6_MSH6_DICDI
MSH6_MSH6_ARATH
MSH6_MSH6_HUMAN
MSH1p_Q84LK0_ARATH
MSH1p_Q1XBQ8_SOYBN
MSH1p_Q0JBW2_ORYSJ
MutS2_Q30SJ7_SULDN
MutS2_MUTS2_HELPY
MutS2_A9NGE8_ACHLI
MutS2_C1D0G1_DEIDV
MutS2_A8UZQ9_9AQUI
MutS2_A9BJX8_PETMO
MutS2_B1C132_9FIRM
MutS2_B7C7S5_9FIRM
MutS2_B0VJJ4_9BACT
MutS2_B5Y861_COPPD
MutS2_D2Z2Q5_9BACT
MutS2_B5YHF6_THEYD
MutS2_MUTS2_BACSU
MutS2_A1AT62_PELPD
MutS2_Q9LVW1_ARATH
MutS2_Q1D4Q8_MYXXD
MutS2_B3ELF9_CHLPB
MutS2_D2QDA3_SPILD
MutS2_B8DKM4_DESVM
MutS2_D0LH43_HALO1
MutS2_C3XIF0_9HELI
MutS2_B7GGY7_ANOFW
MutS2_A0YID1_9CYAN
MutS2_Q7XKD3_ORYSJ
MutS2_A6G6E8_9DELT
MutS2_C1MLV8_9CHLO
MutS3_C4GEA1_9FIRM
MutS3_C6W1B1_DYAFD
MutS3_A7V876_BACUN
MutS3_B9E9P8_MACCJ
MutS3_Q49Z02_STAS1
MutS3_A4A8G2_9GAMM
MutS3_A4BZM7_9FLAO
MutS3_B3W7H4_LACCB

Fig. S3 (2/3)

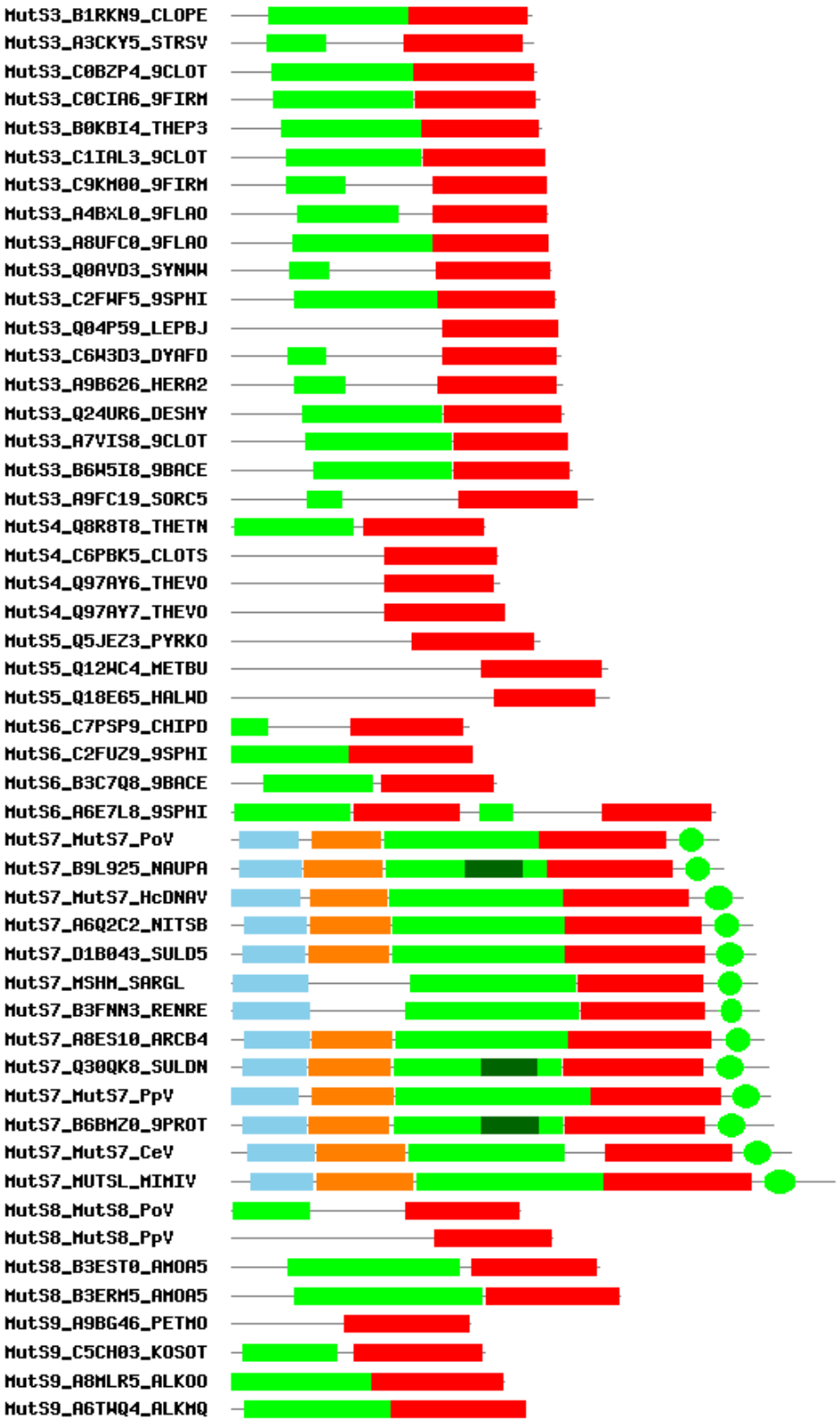

MutS3_B1RKN9_CLOPE
MutS3_A3CKY5_STRSV
MutS3_C0BZP4_9CLOT
MutS3_C0CIA6_9FIRM
MutS3_B0KBI4_THEP3
MutS3_C1IAL3_9CLOT
MutS3_C9KM00_9FIRM
MutS3_A4BXL0_9FLAO
MutS3_A8UFC0_9FLAO
MutS3_Q0AVD3_SYNWW
MutS3_C2FWF5_9SPHI
MutS3_Q04P59_LEPBJ
MutS3_C6W3D3_DYAFD
MutS3_A9B626_HERA2
MutS3_Q24UR6_DESHY
MutS3_A7VIS8_9CLOT
MutS3_B6W5I8_9BACE
MutS3_A9FC19_SORC5
MutS4_Q8R8T8_THETN
MutS4_C6PBK5_CLOTS
MutS4_Q97AY6_THEVO
MutS4_Q97AY7_THEVO
MutS5_Q5JEZ3_PYRKO
MutS5_Q12WC4_METBU
MutS5_Q18E65_HALWD
MutS6_C7PSP9_CHIPD
MutS6_C2FUZ9_9SPHI
MutS6_B3C7Q8_9BACE
MutS6_A6E7L8_9SPHI
MutS7_MutS7_PoV
MutS7_B9L925_NAUPA
MutS7_MutS7_HcDNAV
MutS7_A6Q2C2_NITSB
MutS7_D1B043_SULD5
MutS7_MSHM_SARGL
MutS7_B3FNN3_RENRE
MutS7_A8ES10_ARCB4
MutS7_Q30QK8_SULDN
MutS7_MutS7_PpV
MutS7_B6BMZ0_9PROT
MutS7_MutS7_CeV
MutS7_MUTSL_MIMIV
MutS8_MutS8_PoV
MutS8_MutS8_PpV
MutS8_B3EST0_AMOA5
MutS8_B3ERM5_AMOA5
MutS9_A9BG46_PETMO
MutS9_C5CH03_KOSOT
MutS9_A8MLR5_ALKOO
MutS9_A6TWQ4_ALKMQ

Fig. S3 (3/3)

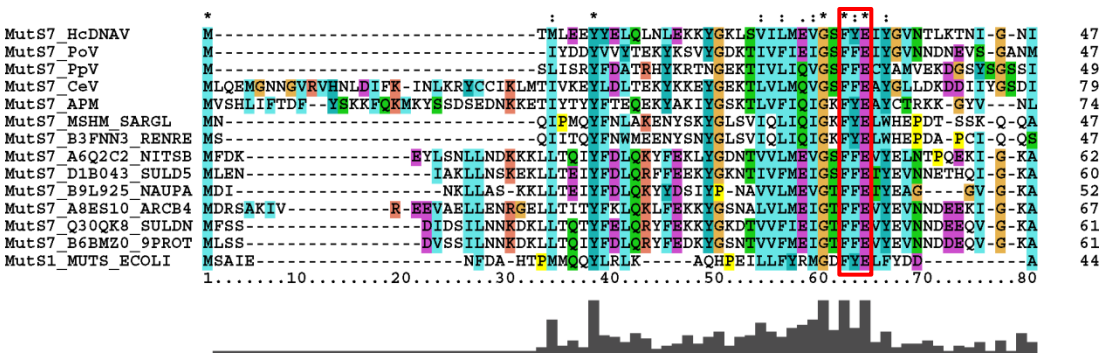
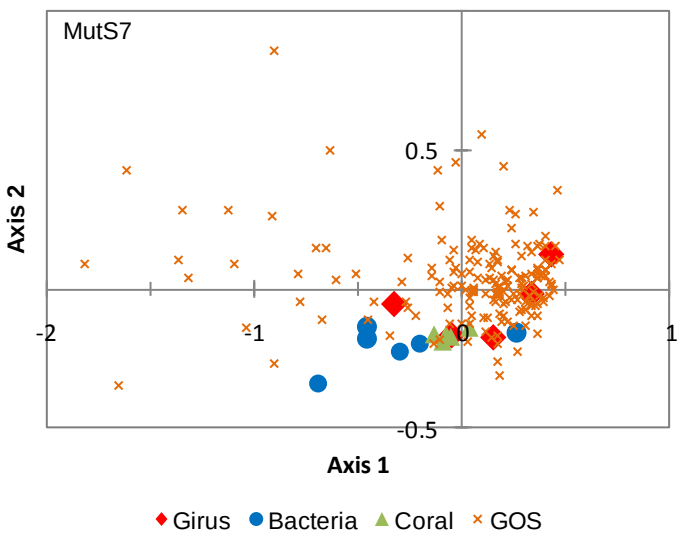
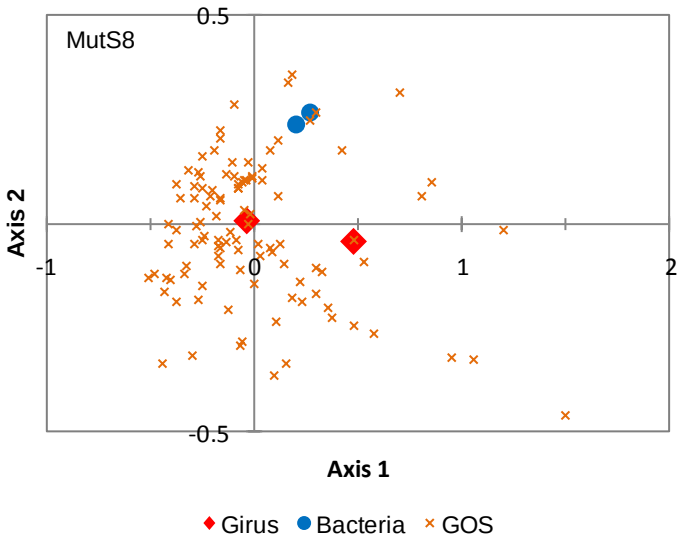
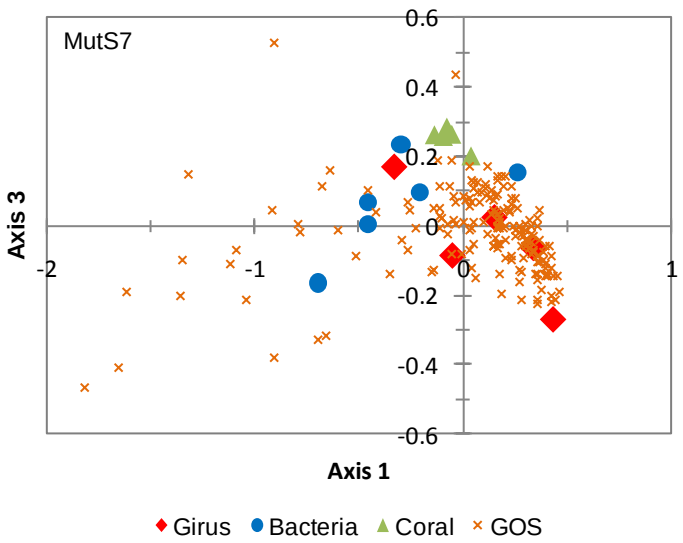
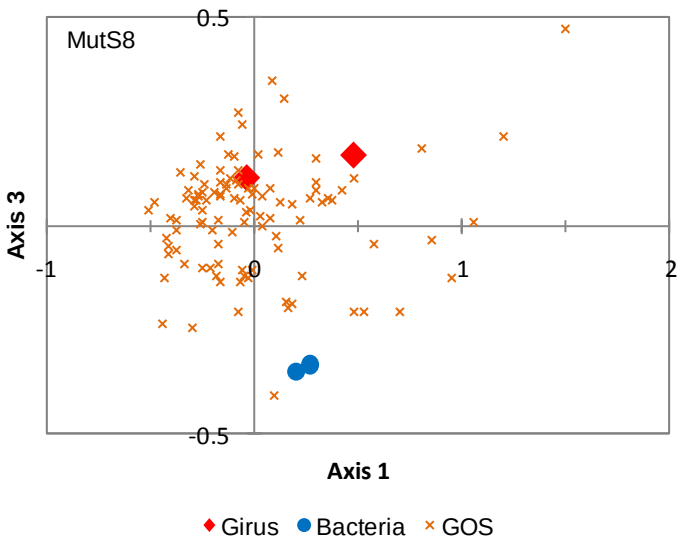

MutS7_HcDNAV
MutS7_PoV
MutS7_PpV
MutS7_CeV
MutS7_APM
MutS7_MSHM_SARGL
MutS7_B3FNN3_RENRE
MutS7_A6Q2C2_NITSB
MutS7_D1B043_SULD5
MutS7_B9L925_NAUPA
MutS7_A8ES10_ARCB4
MutS7_Q30QK8_SULDN
MutS7_B6BMZ0_9PROT
MutS1_MUTS_ECOLI

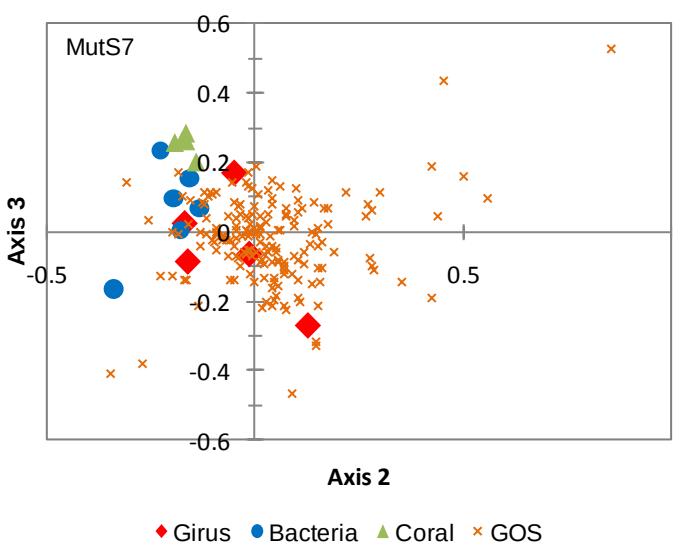
Fig. S4


MutS7
Axis 2
Axis 1
0.5
-0.5
-2
-1
0
1
Girus
Bacteria
Coral
GOS


MutS8
Axis 2
Axis 1
0.5
-0.5
-1
0
1
2
Girus
Bacteria
GOS


MutS7
Axis 3
Axis 1
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-2
-1
0
1
Girus
Bacteria
Coral
GOS


MutS8
Axis 3
Axis 1
0.5
-0.5
-1
0
1
2
Girus
Bacteria
GOS


MutS7
Axis 3
Axis 2
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.5
0.5
Girus
Bacteria
Coral
GOS

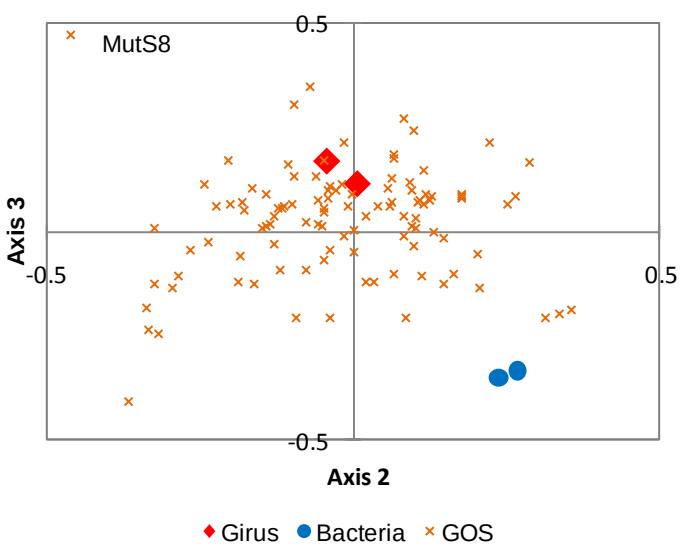

MutS8
Axis 3
Axis 2
0.5
-0.5
-0.5
0.5
Girus
Bacteria
GOS

Fig. S5