

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

Hiroyuki Ogata^{1,*}, Jessica Ray², Kensuke Toyoda^{3,†}, Ruth-Anne Sandaa², Keizo Nagasaki³,
Gunnar Bratbak², Jean-Michel Claverie¹

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, Kanagawa 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 and 1X Coraload 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

Ogata et al., Table S3

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

Ogata et al., Table S3

JCVI_SITE_GS029	Coastal	Galapagos Islands	North James Bay, Santiago 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

Ogata et al., Table S3

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.



Fig. S1



Fig. S2

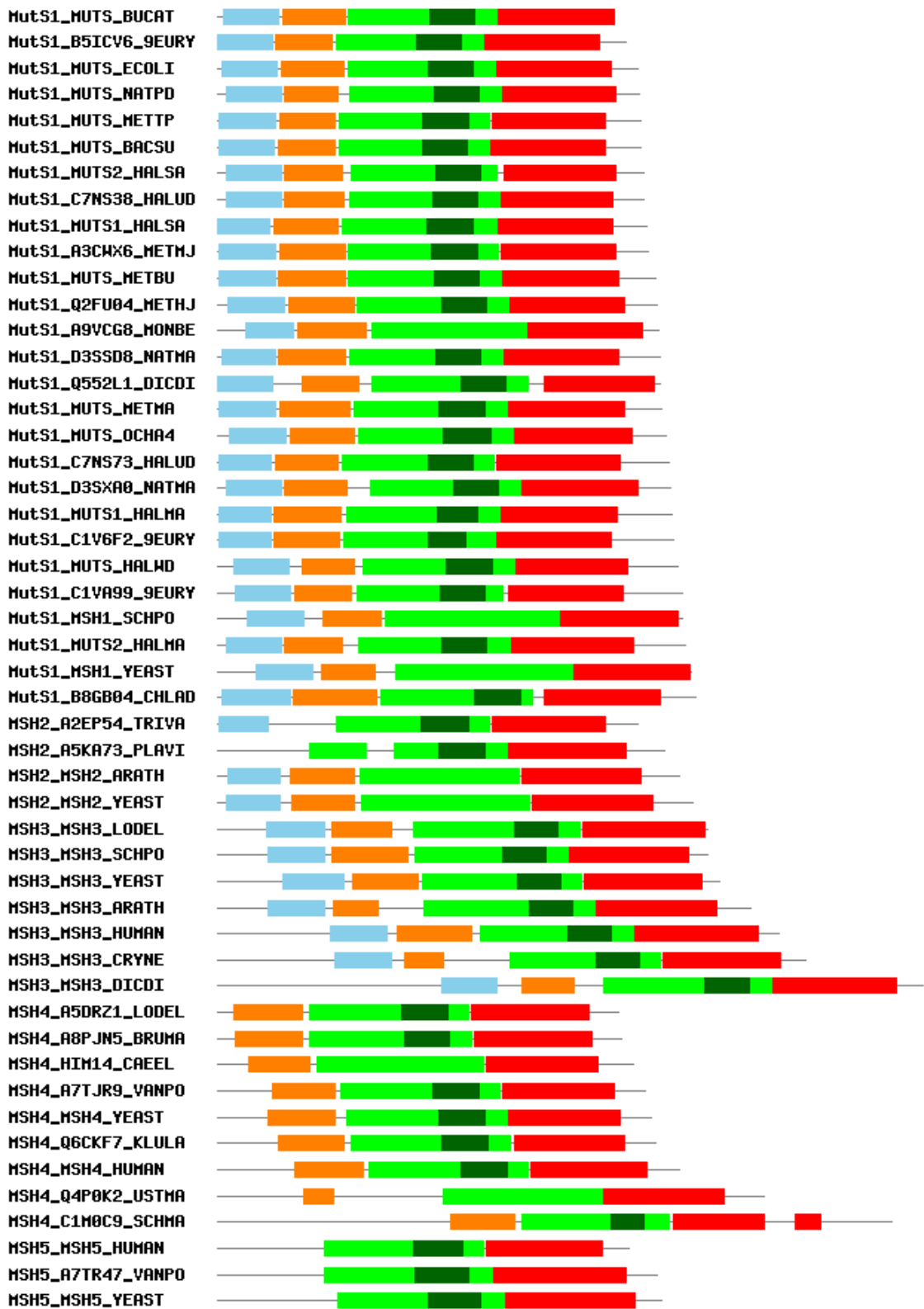


Fig. S3 (1/3)

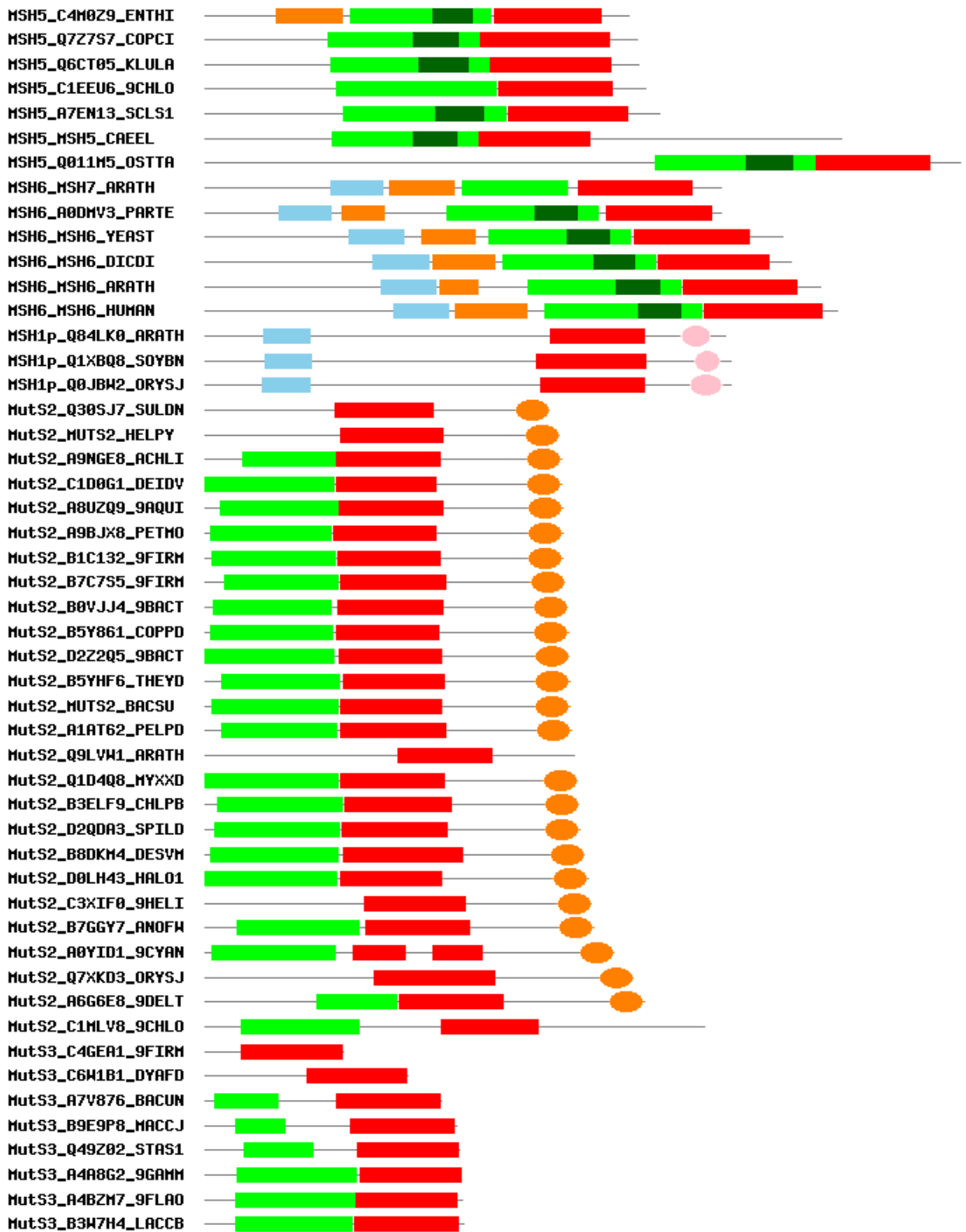


Fig. S3 (2/3)

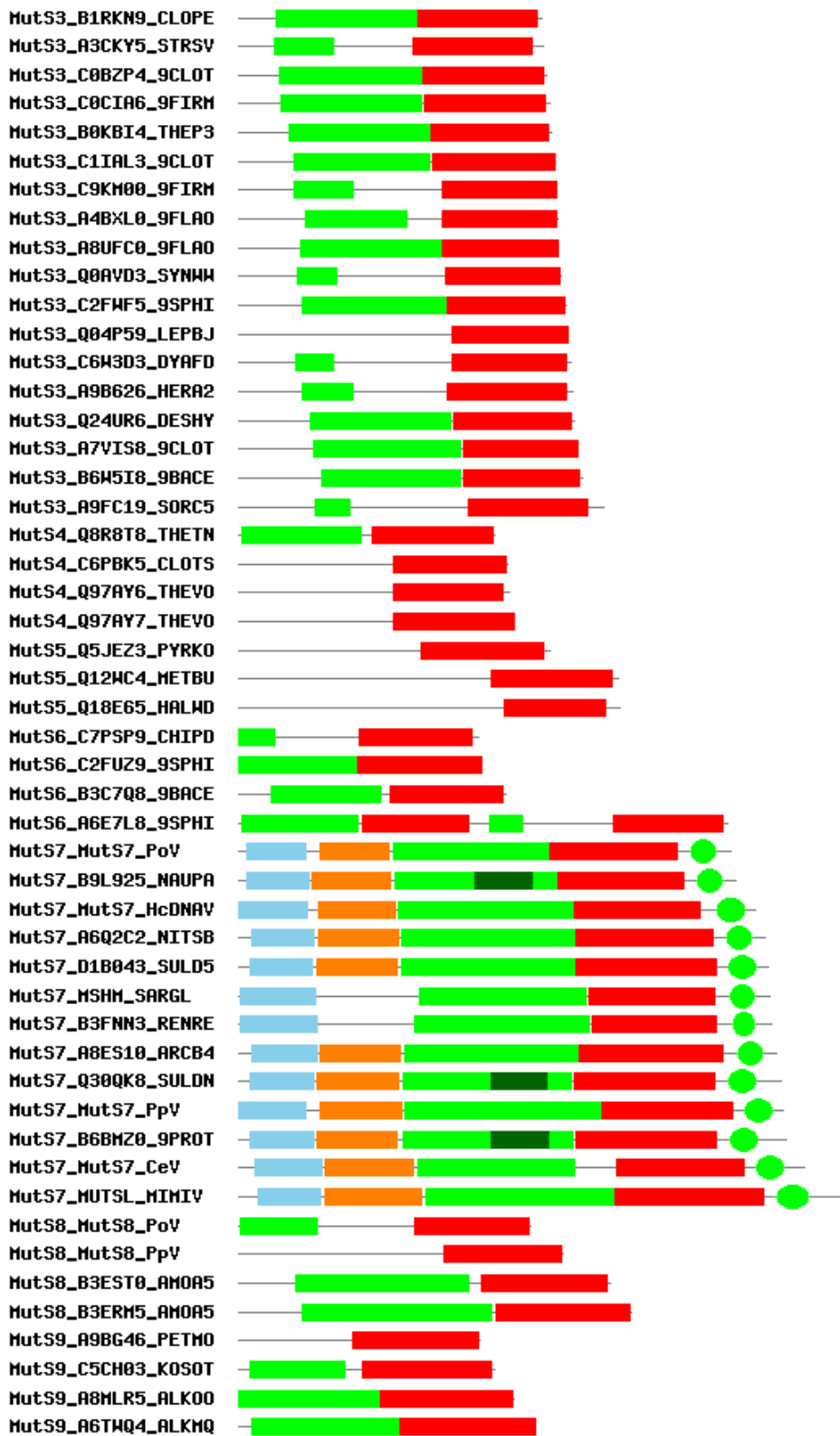


Fig. S3 (3/3)

```

*           :           *           : : : * * : :
MutS7_HcDNAV M-----TMLREYBELNLNLEKKYGKLSVILMEVGSFYEIINGVNTLKTNI-G-NI 47
MutS7_PoV M-----IYDDYVVYTEKKKSVYGDKTIVFIEIGSFPEIYGVNNDNVG-GANM 47
MutS7_PpV M-----SLISRYFDATRHHKRTNGEKTIVLIQVGSFPECYAMVEKDGSGSGSSI 49
MutS7_CeV MLQEMGNNGVRVHNLDFK-INLKRVCICXLMITVKEIYLDLTKKKEVGEKTLVLMQVGSFPEAYGLLDKDDIYGSDI 79
MutS7_APM MVSHLIFTDF--YKKFKRMKYSDSEDNKKETIYTYFTEKNAKIYGSKLVFQIGKFFVEAYCTRRK-GYV---NL 74
MutS7_MSHM_SARGL MN-----QITQYFNWMEENYSNYGLSVIQLIIGKFFVELWHEPDT-SSK-Q-QA 47
MutS7_B3FNN3_RENRE MS-----QITQYFNWMEENYSNYGLSVIQLIIGKFFVELWHEPDA-PCI-Q-QS 47
MutS7_A6Q2C2_NITSB MFDK-----EYLSNLLNDKKLLTQIYFDLQKYPEKLYGDNTVVLMEVGSFPEVVELNTPQEKI-G-KA 62
MutS7_D1B043_SULD5 MLEN-----IAKLLNSKELLTEIYFDLQKYPEKLYGKNTIVFMVEIGSFPEVVEVNNETHQI-G-KA 60
MutS7_B9L925_NAUPA MDI-----NKLAS-KKLLTEIYFDLQKYPVDSIYP-NAVVLMEVCTPFEVVEAG---GV-G-KA 52
MutS7_A8ES10_ARCB4 MDRSAKIV--R-EEVAELLENRGELLTIYFKLQKLPFKKYGGSNALVLMIECTPFEVVEVNNDEEKI-G-KA 67
MutS7_Q30QK8_SULDN MFSS-----DIDSILNNKDKLLTQIYFDLQKYPFKKYGKDTVVFIEICTPFEVVEVNNDEBQV-G-KA 61
MutS7_B6BMZ0_9PROT MLSS-----DVSSILNNKDKLLTQIYFDLQKYPFDKYGGSNTVVFVEICTPFEVVEVNNDEBQV-G-KA 61
MutS1_MUTS_ECOLI MSAIE-----NFDA-HTPMMQYLRLLK---AQHPETLLFVRMGDFYELFYDD---A 44
1.....10.....20.....30.....40.....50.....60.....70.....80

```



Fig. S4

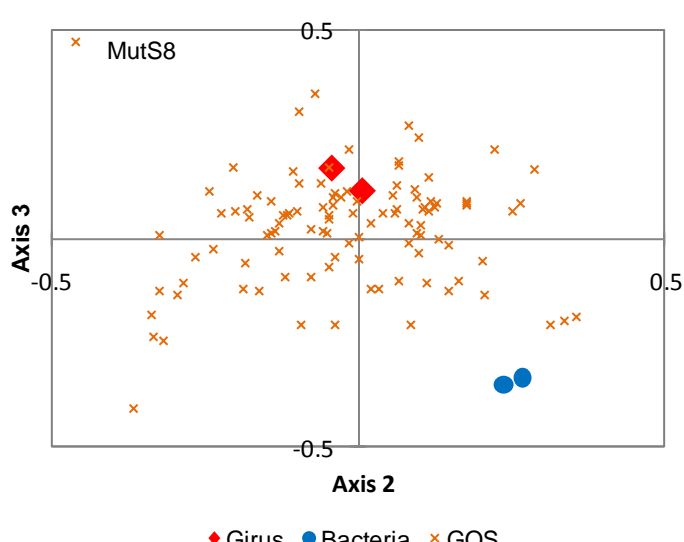
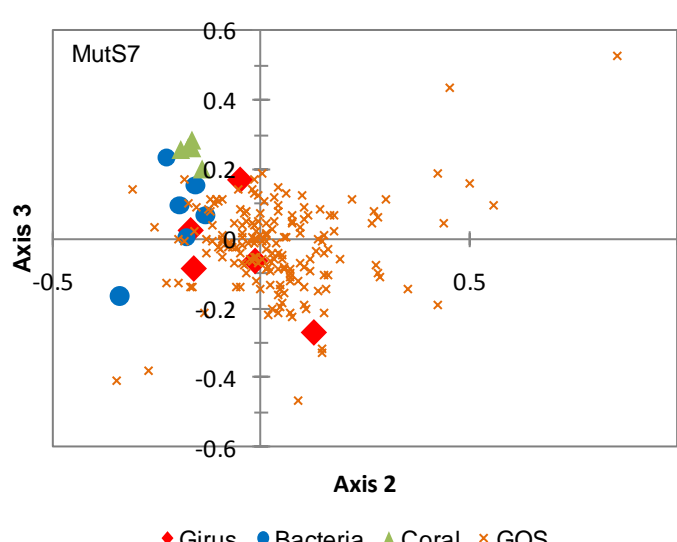
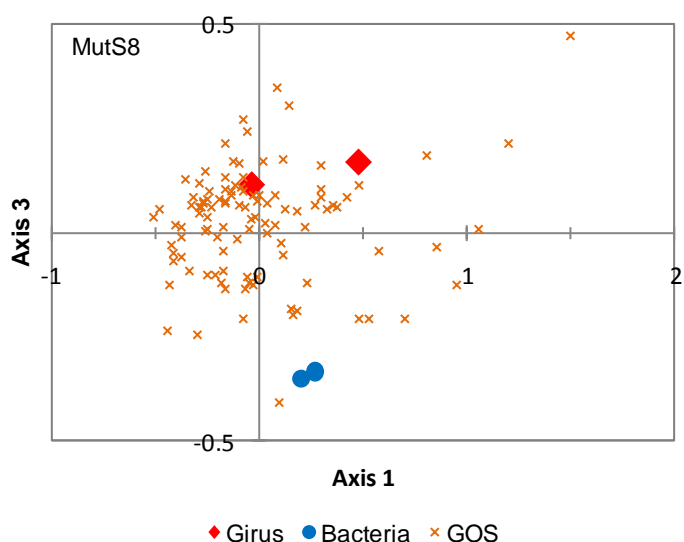
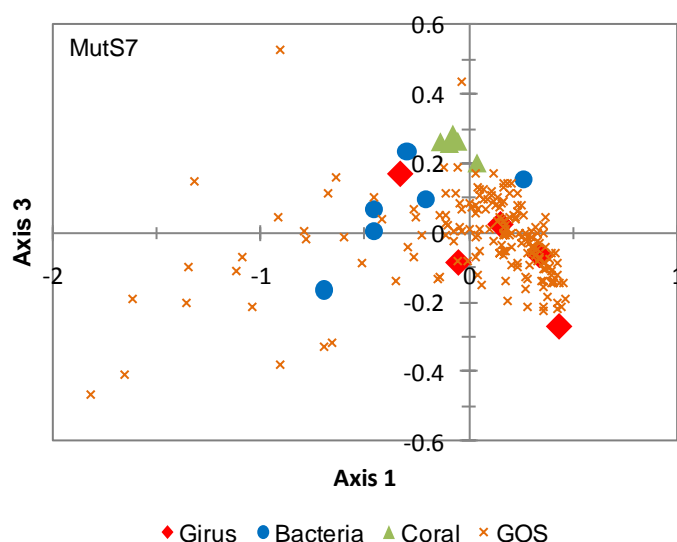
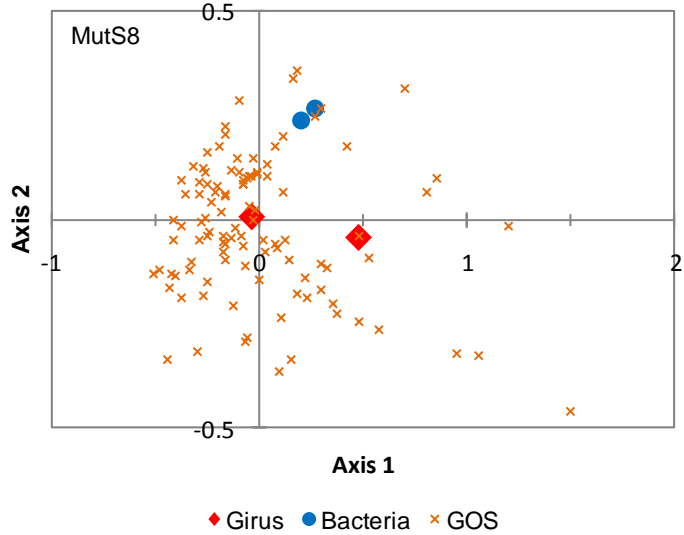
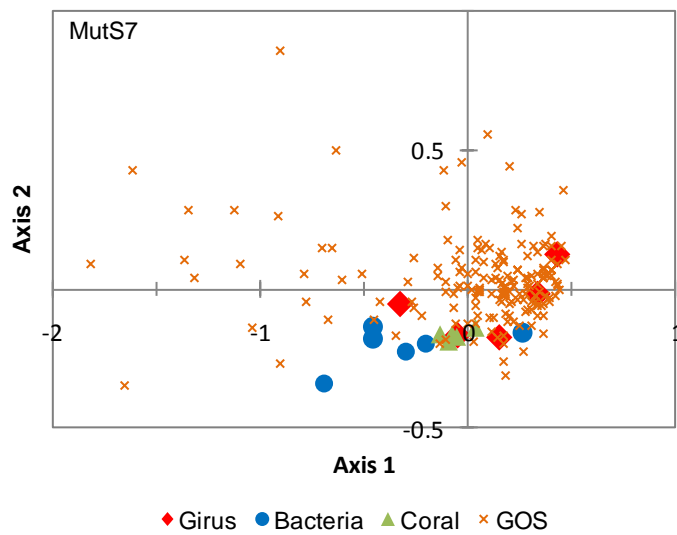


Fig. S5