**Investigation of the Evolution of MutS Protein Family in Animals**

**EEOB 563 Final Project**

**Mudith Ekanayake**

**Introduction**

Throughout different stages of the cell cycle, many DNA repair pathways act in the cells allowing the cells to repair the DNA damage. Among these pathways, mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER), homologous recombination (HR) and non-homologous end joining (NHEJ) are the major pathways that are active in the cells. More importantly DNA mismatch repair (MMR) is responsible for recognizing and repairing erroneous insertion, deletion and mis-incorporation of bases during DNA replication and recombination. In MMR, there are three major steps including mismatch identification, mismatch excision and DNA re-synthesis. Throughout this whole process several proteins are involved and among those proteins, MutS is incorporated in detecting mismatches in the sequences. These proteins are well conserved in prokaryotes, eukaryotes and even in viruses. There are many homologs of MutS protein including MutS1, MutS2 in bacteria and MSH 1 to 6 in eukaryotes. But In this project, MSH2, MSH3, MSH4, MSH5 and MSH6 were selected for the analysis along with few reference sequences including mitochondrial MutS (mtMutS) from Octocorals (*Dendronephthya gigantea*). There are many theories about the evolution of MSH protein and still there are some confusions which protein evolved first and then how the others got descended from that. Some studies suggest that octocoral mitochondrial MutS does not have a eukaryotic origin and it has evolved due to horizontal gene transfer (HGT) from a large DNA virus into the mitochondrial genome (Bilewitch and Degnan, 2011) while others hypothesize eukaryotic MSH proteins have originated from a HGT event between bacteria and eukaryotes. Since mismatch repair is critical for maintaining genome stability, from this project, evolution of the MutS protein family in animals will be investigated in order to find evidence for the above mentioned theories.

**Methods**

MSH protein Sequences were retrieved from five different groups including Archaea, Bacteria, Metagenomics, Viruses and Eukaryotes excluding animals, plants and fungi. Protein BLAST (blastp) was carried out in NCBI for human MSH proteins from 2 to 6 against above mentioned groups and only the top 10 hits were selected for the analysis. Five separate datasets were generated for five different MSH proteins and to each dataset, a reference dataset was added which included MSH1-6 from *Saccharomyces cerevisiae*, MSH1-6 from *Nematostella*, MSH2-6 from human and mtMutS from *Dendronephthya gigantea*.

All the steps in the phylogenetic analysis were performed five times for five different MSH datasets. MAFFT (Katoh *et al*., 2002) sequence alignment program was used for aligning the protein sequences and the alignments were curated and cleaned using TrimAl (Capella-Gutierrez *et al*., 2009) selecting “automated 1” method. Prior to constructing the phylogeny, model selection step was carried out for all the alignments incorporating SMS: Smart Model Selection tool in PhyML (Lefort *et al*., 2017) using the selection criterion as Akaike Information Criterion (AIC) (Akaike, 1973).

Phylogenetic reconstruction was undertaken under a maximum likelihood framework implemented in RAxML-NG (Kozlov, *et al*., 2019) using the models selected in SMS model selection, and with confidence levels estimated using bootstrap resampling with 1000 replicates. Interactive Tree Of Life (iTOL) v4 tool (Letunic and Bork, 2019) and FigTree v1.4.4 (Rambaut, 2018) was used for the visualization of the tree.

**Results**

After aligning with MAFFT, TrimAl removed poorly aligned regions from multiple sequence alignments and it significantly shortened the alignments. SMS tool selected LG+G+I+F as the best model for all the 5 MSH datasets. Table 1 shows a summary of the model selection step. Complete results are attached in the supplementary materials.

Table 1: Summary of the model selection by SMS smart model selection tool (K: number of model free parameters, Llk: Log-likelihood of the data given the substitution model and the tree, AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MutS Homolog** | **Model** | **Decoration** | **K** | **Llk** | **AIC** | **BIC** |
| MSH2 | LG | +G+I+F | 154 | -104989.82347 | 210287.64694 | 211176.78048 |
| MSH3 | LG | +G+I+F | 154 | -117976.37152 | 236260.74304 | 237165.86220 |
| MSH4 | LG | +G+I+F | 154 | -99178.91815 | 198665.83630 | 199549.16032 |
| MSH5 | LG | +G+I+F | 154 | -93980.49230 | 188268.98460 | 189147.66437 |
| MSH6 | LG | +G+I+F | 154 | -110644.83080 | 221597.66160 | 222490.88671 |

RAxML-NG estimated maximum likelihood values and tree topologies for the best tree found according to the selected models for each and every MutS homolog (20 tree search). The log likelihood values of the best tree and AIC, AICc, BIC scores predicted by RAxML-NG are tabulated in the table 2. The bootstrap analyses with 1000 replicates returned bootstrap support for all the nodes and the values are shown near the nodes of the trees (Figure 1-5). Polar phylogenetic trees with more clear visualizations of bootstrap values are included in supplementary materials (Supplementary figures 1-5).

Table 2: Summary of the RAxML-NG phylogeny reconstruction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MutS Homolog** | **Best Tree logLH** | **AIC score** | **AICc score** | **BIC score** |
| MSH2 | -42591.341315 | 85490.682629 | 85637.574937 | 86133.445689 |
| MSH3 | -59140.561955 | 118589.123909 | 118690.054142 | 119273.275095 |
| MSH4 | -56558.934077 | 113425.868153 | 113522.118153 | 114115.558638 |
| MSH5 | -37632.010775 | 75572.021549 | 75723.577105 | 76211.542384 |
| MSH6 | -45121.922768 | 90551.845537 | 90685.197492 | 91204.848017 |

In the constructed phylogenetic trees, many interesting relationships could be seen. All the viral sequences were clustered in one single clade in each and every tree and interestingly, *Dendronephthya gigantea* sponge mtMutS was also included in all the viral clades (Figure 1-5). The bootstrap support value for the clade was 100 at every occasion. In the figure 1, all the eukaryotic MSH2 proteins were incorporated in a clade with metagenome samples as well as bacteria with 100 percent bootstrap support. More importantly, another bacteria protein was outgrouped with the whole clade. This specific relationship could be seen in the MSH6 phylogenetic tree as well (Figure 5). In addition to that, each MutS homolog reference sequences from *Saccharomyces cerevisiae*, *Nematostella*, and human were clustered together as expected in all the MSH trees giving 100 percent bootstrap support (Figure 1-5).

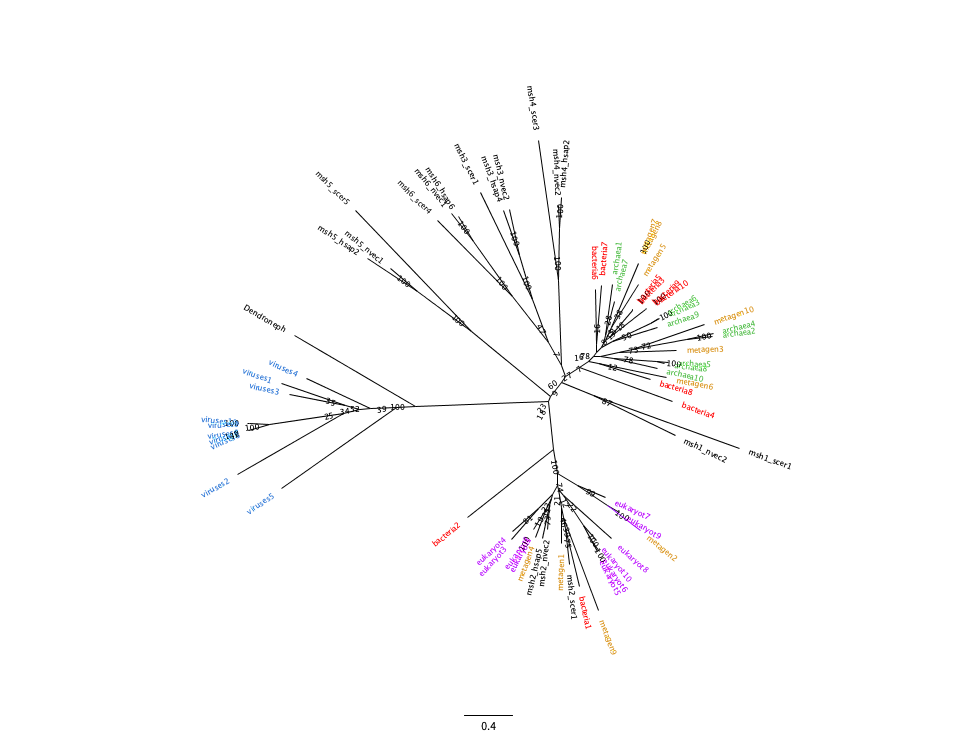


Figure 1: Phylogenetic tree of MSH2 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

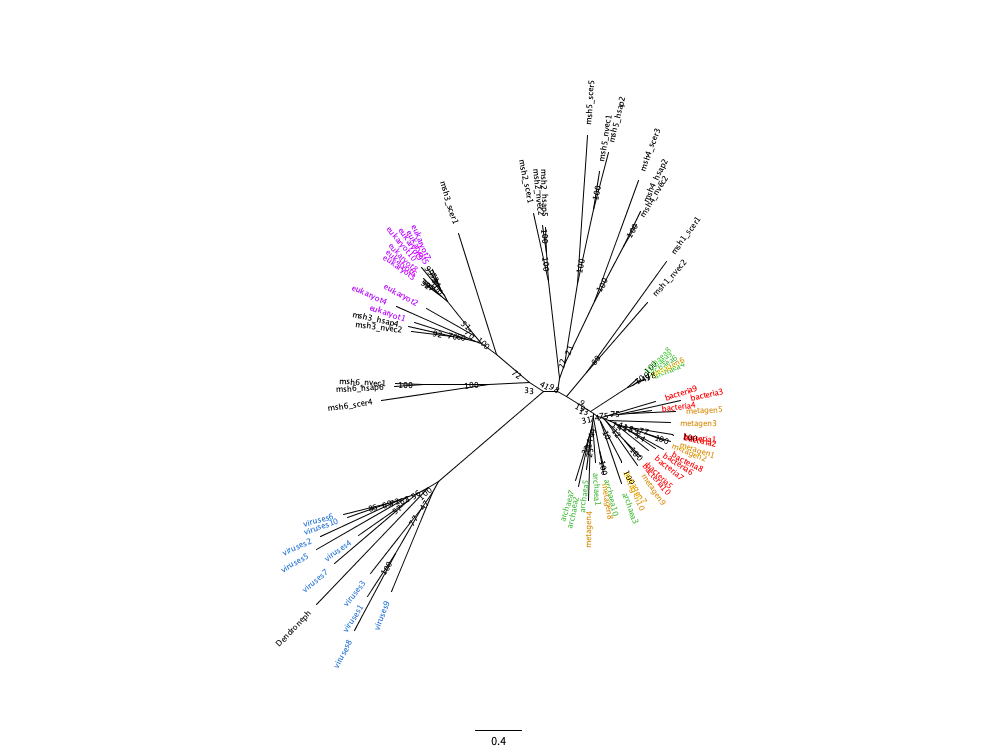


Figure 2: Phylogenetic tree of MSH3 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.



Figure 3: Phylogenetic tree of MSH4 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

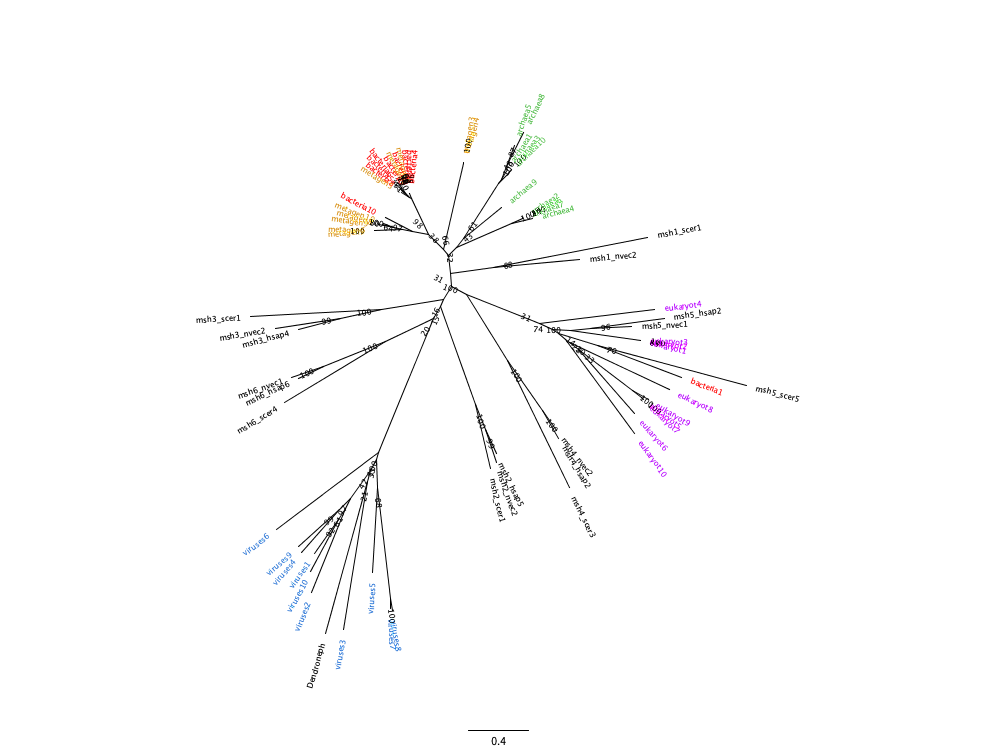


Figure 4: Phylogenetic tree of MSH5 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

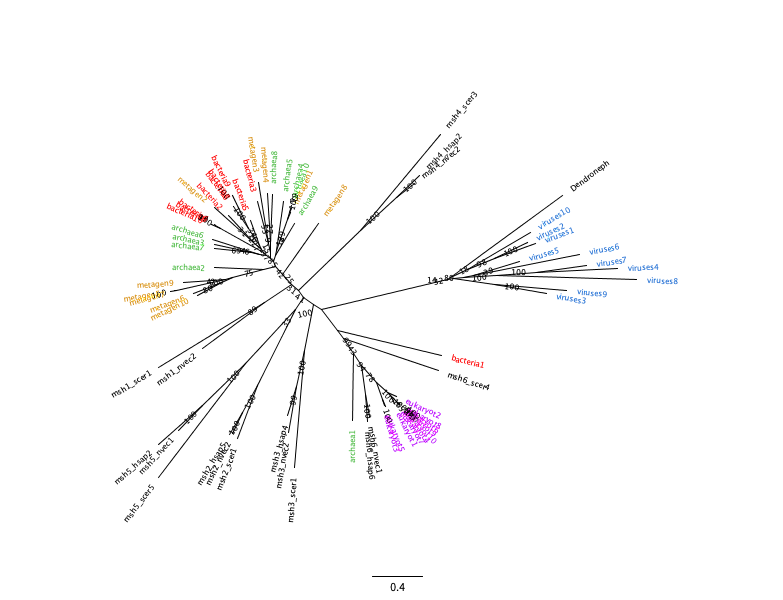


Figure 5: Phylogenetic tree of MSH6 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

**Discussion**

The origin of mtMutS has been unsolved for many years even though many hypotheses were proposed. *Dendronephthya gigantea* mtMutS protein is only found in octocorals and it is not orthologous with any other eukaryotic species. All the constructed phylogenies cluster viral sequences and mtMutS of *Dendronephthya gigantea* together indicating that mtMutS does not belong to the MSH family and also it does not have a eukaryotic origin. Both having this relationship in all the trees and indicating 100 percent bootstrap value provide more support for this phenomena. This observation can be an evidence for the hypothesis that octocoral mtMutS has originated due to a horizontal gene transfer (HGT) event from a large DNA virus.

There are several studies that propose opposing theories about eukaryotic MutS family evolution. While some suggest that MSH homologs form a monophyletic clade in the evolutionary process, many others establish the theory of paraphyletic MSH evolution with some relationship to bacteria (Bilewitch and Degnan, 2011; Muthye and Lavrov, 2020). In this study, phylogenetic tree for MSH2 (Figure 1) clearly shows a clade of eukaryotic sequences including an outgroup of bacteria (bacteria 2) with 100 bootstrap support. It indicates that there are prokaryotic MutS sequences which are more closely related to eukaryotic MSH2 than eukaryotic MSH2 related to other species. The outgroup prokaryote, Bacteria 2 refers to HHG10973.1 MutS family DNA mismatch repair protein [bacterium]. This bacteria sequence is from a hot springs metagenome. The clade includes few metagenome sequences from metagenomes 1, 2, 4 and 9. But these sequences are also from eukaryotes (metagenome1: an aquatic fungi; metagenome2: a marine algae; metagenome4: a unicellular marine eukaryote; metagenome9: a ubiquitous fungi which associate with decaying plant matter). The above stated clade can be an evidence for horizontal gene transfer event in between bacteria and eukaryotic genomes. Figure 6 also illustrates the same relationship, but with small changes. In here, archaea 1 which is a sequence from phyllosphere metagenome is included in eukaryote cluster and the whole cluster is outgrouped with bacteria 1 (MBE7180883.1 hypothetical protein). This protein is from *Terriglobus roseus* plastic metagenome and the bacterial species is commonly found in agricultural soils. In the figure 5 also bacteria 1 includes in the eukaryotic clade (70 percent support for the bacterial sequence and MSH5) providing more evidence for the relationship between eukaryotic MSH and bacteria. In addition to that specific clade, most of the trees cluster MSH sequences with bacterial sequences and it gives the idea that there can be a strong relationship between bacteria and other groups of organisms.

In conclusion, this study provides possible evidence to the hypothesis that the eukaryotic MSH can be evolved due to a horizontal gene transfer event in between bacteria and eukaryotic genome. Furthermore it supports the theory that the origin of octocoral mtMuts has occurred due to another HGT event from a large DNA virus. Even though this analysis gives some evidence about those relationships, further studies are required to confirm these theories.

**References**

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle In: Petrov BN, Csaki F, editors. Second international symposium on information theory. Budapest (Hungary: ): Akademiai Kiado; p. 267–281.

Bilewitch, J. P., Degnan, S. M. (2011). A unique horizontal gene transfer event has provided the octocoral mitochondrial genome with an active mismatch repair gene that has potential for an unusual self-contained function. *BMC Evol Biol* 11, 228. https://doi.org/10.1186/1471-2148-11-228.

Ivica Letunic, Peer Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments, *Nucleic Acids Research*, Volume 47, Issue W1, 02 July 2019, Pages W256–W259, https://doi.org/10.1093/nar/gkz239.

Katoh, et al. “MAFFT: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *OUP Academic*, Oxford University Press, 15 July 2002, academic.oup.com/nar/article/30/14/3059/2904316.

Kozlov, et al. “RAxML-NG: a Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference.” *OUP Academic*, Oxford University Press, 9 May 2019, academic.oup.com/bioinformatics/article/35/21/4453/548.

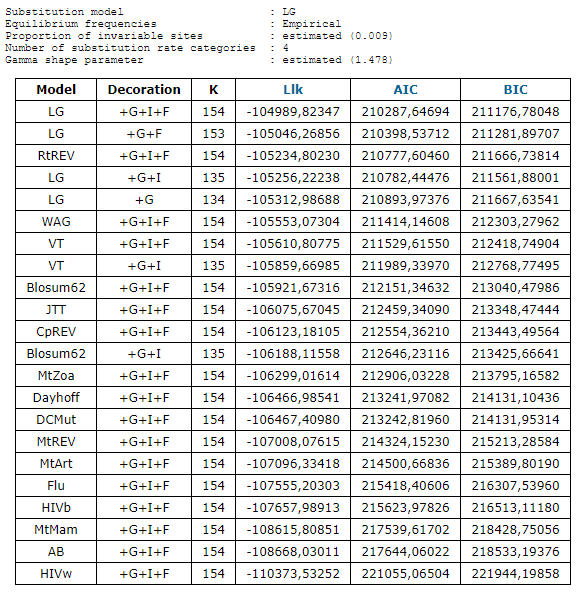
Lefort, V., Longueville, J. E., & Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular biology and evolution, 34*(9), 2422–2424. https://doi.org/10.1093/molbev/msx149.

Salvador Capella-Gutierrez, Jose M. Silla-Martinez, Toni Gabaldon. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25: 1972-1973.

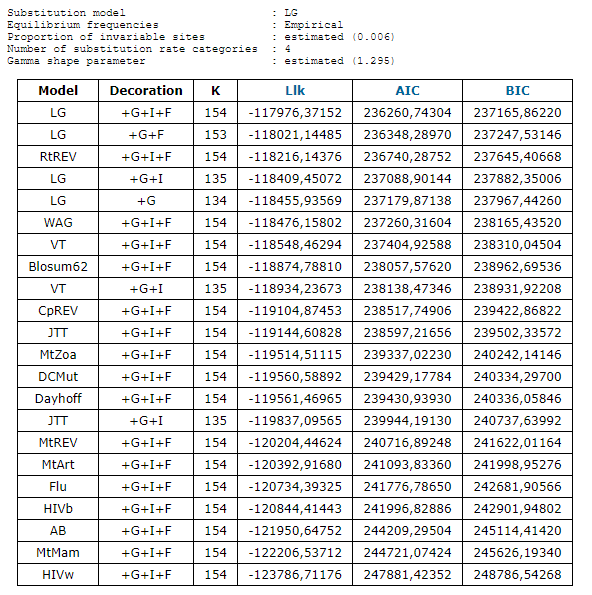
Viraj Muthye, Dennis V. Lavrov. (2020). Dynamic evolution of the MutS family in animals: multiple losses of MSH paralogues and gain of a viral MutS homologue in octocorals. *bioRxiv* 2020.12.22.424024; doi: https://doi.org/10.1101/2020.12.22.424024.

**Supplementary Materials**

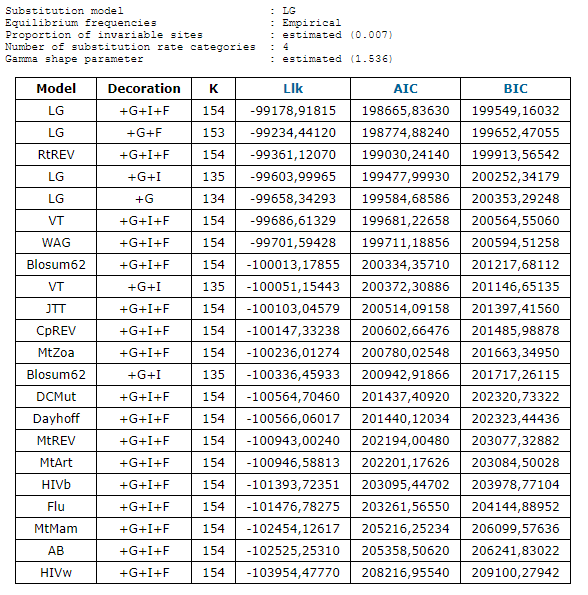
Supplementary Table1: Complete results for MSH2 model selection by SMS smart model selection tool.



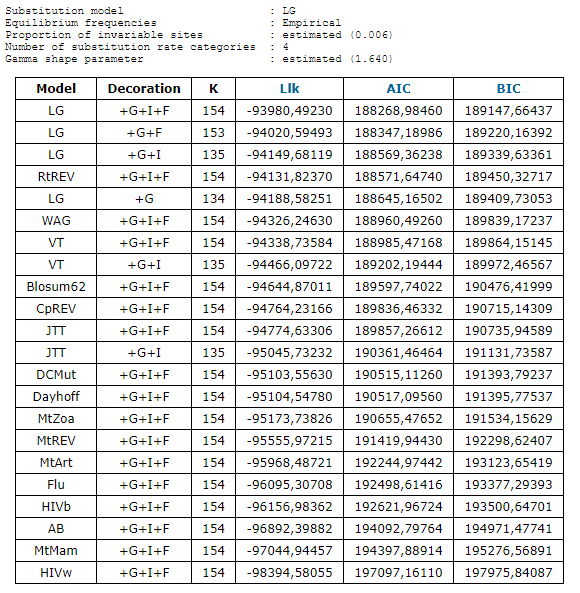
Supplementary Table2: Complete results for MSH3 model selection by SMS smart model selection tool.



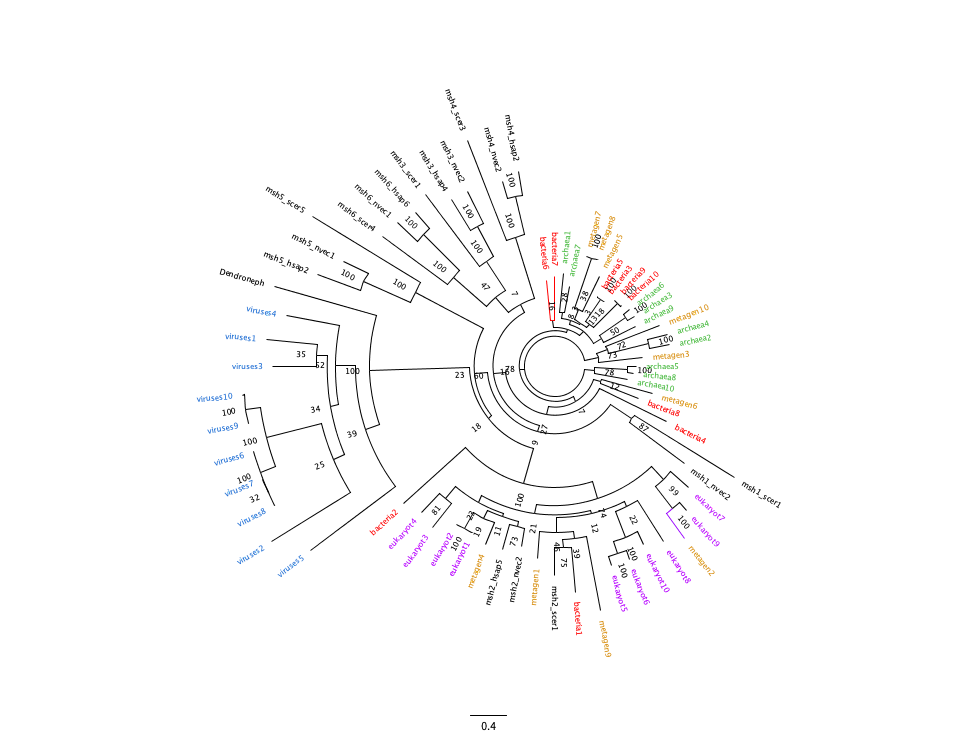
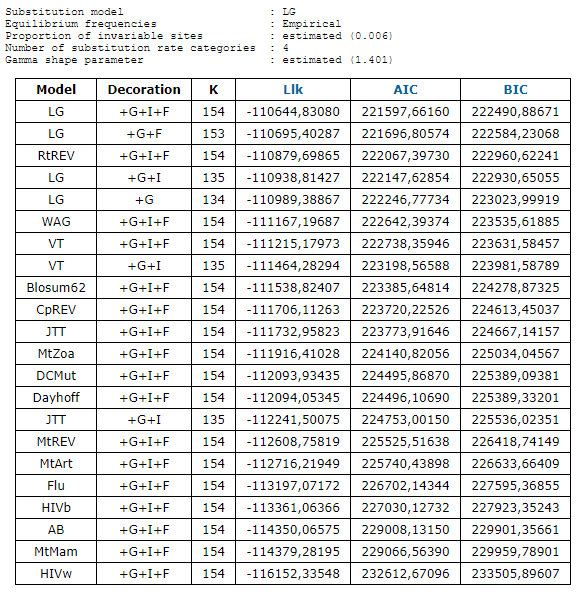
Supplementary Table3: Complete results for MSH4 model selection by SMS smart model selection tool.



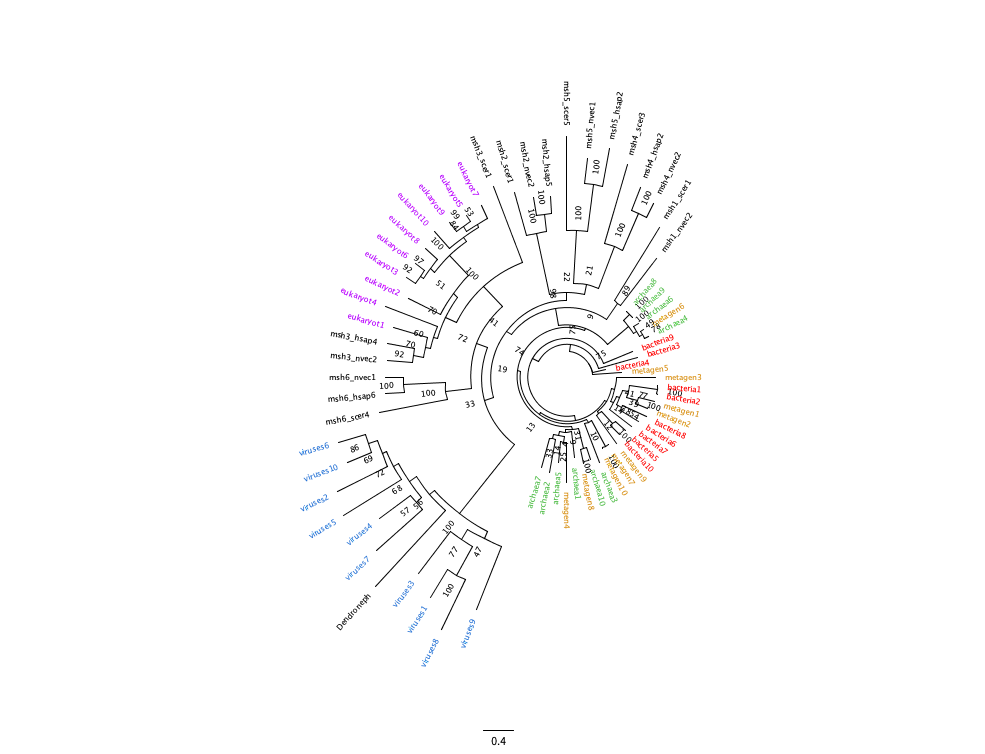
Supplementary Table4: Complete results for MSH5 model selection by SMS smart model selection tool.



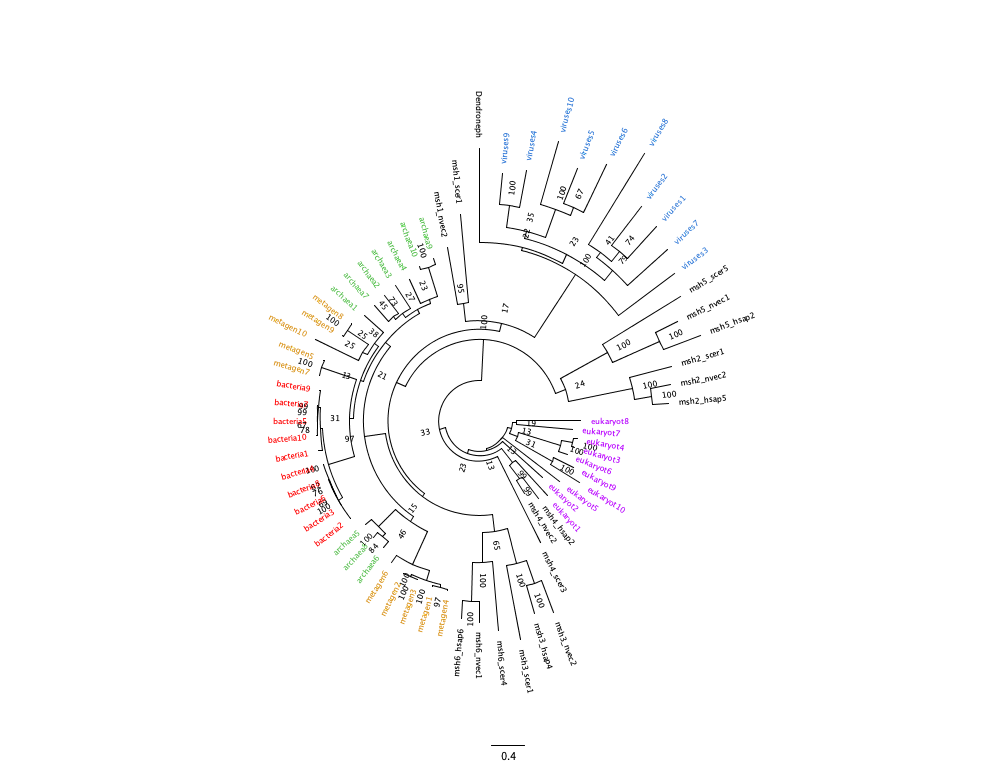
Supplementary Table5: Complete results for MSH6 model selection by SMS smart model selection tool.



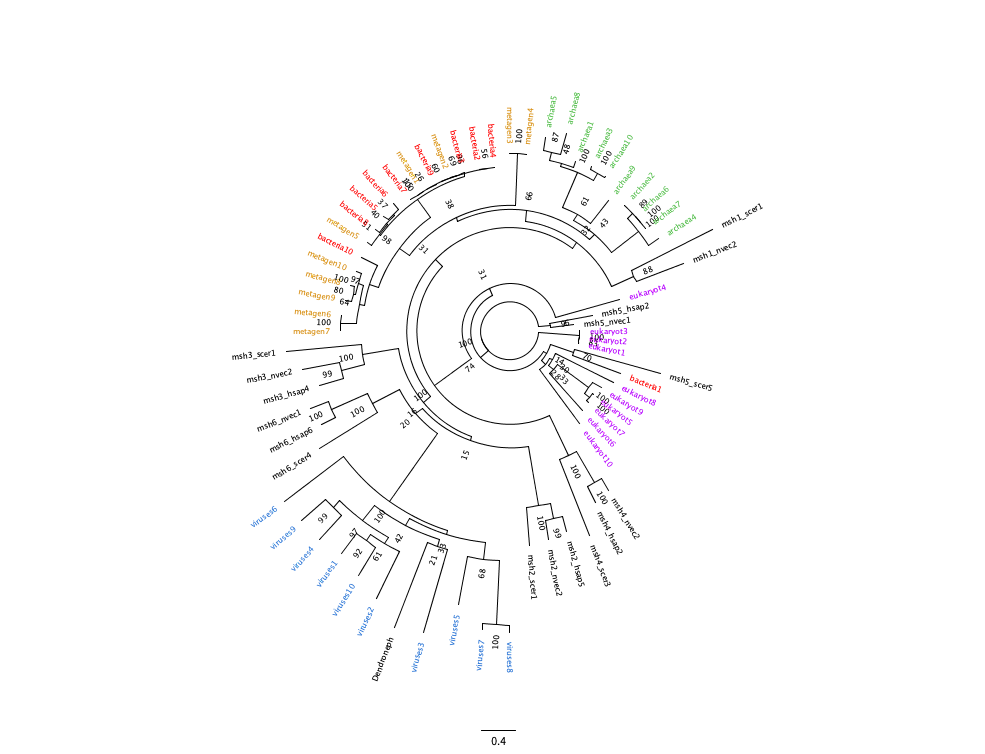
Supplementary Figure 1: Polar phylogenetic tree of MSH2 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.



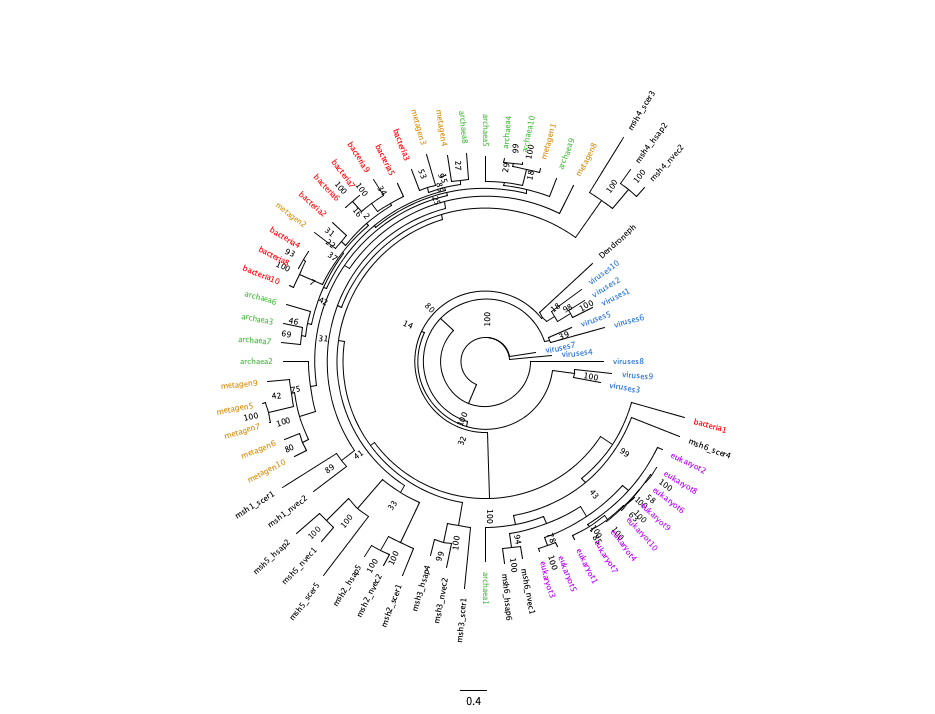
Supplementary Figure 2: Polar phylogenetic tree of MSH3 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.



Supplementary Figure 3: Polar phylogenetic tree of MSH4 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.



Supplementary Figure 4: Polar phylogenetic tree of MSH5 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.



Supplementary Figure 5: Polar phylogenetic tree of MSH6 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.