2017

Mrub_0860, Mrub_0701 and Mrub_2285 are orthologous to *E. coli* b2892, b2562 and b3863 within the RecFOR pathway for Homologous Recombination

Bailey Englund
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

*Mrub*_0860, *Mrub*_0701 and *Mrub*_2285 are orthologous to *E. coli* b2892, b2562 and b3863 within the RecFOR pathway for Homologous Recombination

Bailey Englund
Dr. Lori Scott Laboratory
Biology Department, Augustana College
639 38th Street, Rock Island, IL 61201

INTRODUCTION

Study of *Meiothermus ruber* with *Escherichia coli* as a control

The study of *Meiothermus ruber (M. ruber)* is an important one, as it is an organism that has a lack of research (Scott, 2016b). *M. ruber* is gram-negative bacterium of red pigmentation; it grows in very restricted and extreme habitats, anywhere with a temperature range of 35-70°C (Tindall et al., 2010). It is advantageous to research organisms that are poorly studied to determine identification of proteins and families, to make phylogenetic connections, to understand processes and evolutionary history (JGI, 2017). The JGI's Genomic Encyclopedia of Bacteria and Archaea (GEBA) is program to sequence bacterial and archaeal genomes from poorly studied branches of the Tree of life (JGI, 2017). In this research project, I researched three open reading frames from *M. ruber* for the purpose of predicting their function. *E. coli* was used as a "positive control" because its genome has been sequenced, all of its gene identified and many functionally confirmed. This information is available and searchable through numerous online databases. This project focuses on the database (Keseler *et al.*, 2013), which is devoted to the study of *Escherichia coli* K-12 MG1655 strain. It provides literature-based curation of the entire genome, and of transcriptional regulation, transporters, and metabolic

pathways. Ecocyc, plus other databases housed with the GENI-ACT platform, helped predict the function of the genes within the *M. ruber* genome.

Homologous Recombination

Homologous recombination is an important pathway for DNA double strand break repairs, especially for breaks that can be lethal. Homologous recombination takes place within the S-G2 phase of the cell cycle and adds to the genomic integrity of cells by repair the DNA through strand invasion, Holliday junction, branch migration and DNA synthesis (Kinesha *et al*., 2016b). Specifically, this homologous recombination via the RecFOR pathway can also be called Gap-filling Recombinational Repair, which incorporates the filling of single strand DNA gaps by sister or homologous chromosomes by strand transfer reactions (Persky & Lovett, 2008).

As shown in Figure 1, homologous recombination (in prokaryotes) using the RecFOR pathway begins with one broken strand of DNA. This mechanism uses RecJ, an exonuclease, and SSB, single-stranded binding protein, to bind with its complement on a DNA molecule (Persky & Lovett, 2008). The RecF, RecO, and RecR proteins act together to promote the function of RecA. This allows RecA to have less inhibition by SSB (Kinehisa *et al*., 2016b). RecFOR and RecA are utilized in the second step; RecFOR are gap repair proteins, and RecA is a central strand exchange protein. These proteins are used to form a presynaptic filament between the broken DNA parts. The third step uses DpoI, DNA polymerase I, with DNA strand invasion to form the Holliday junction intermediate. The Holliday junction is a structure formed between two double-stranded DNA to exchange information to bind them together. Branch migration takes place for this information to be exchanged. The fourth step in the process is the use of RuvA,

RuvB and RuvC (RuvABC) or RuvG, which are ATP-dependent helicases used to separate the DNA strands and complete the repair and recombination (Kanehisa *et al*., 2016b).

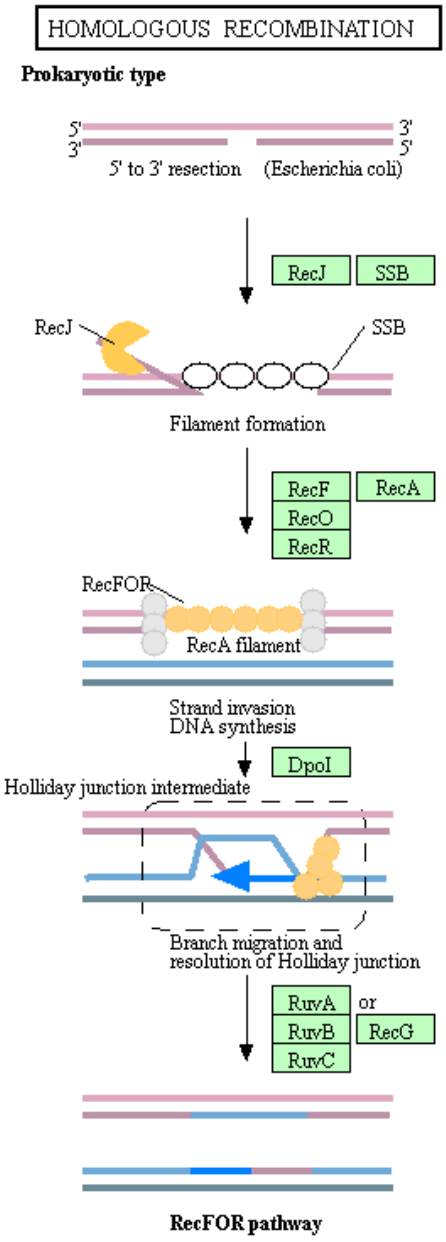


Figure 1. Homologous recombination pathway showing the proteins within each step of the process (Kinehisa *et al.*, 2016b).

Bioinformatics

The use of bioinformatics tools was extremely important to this project and to the study of science as a whole. The main tools consist of computer programs and internet databases. Bioinformatics is serving as a way to identify genes and pathogenic pathways and will continue to be a way to discover cellular networks, complex interactions identifying roles (Debouk &Metcalf, 2000). For those who are knowledgeable about bioinformatics, finding data can be advantageous and easier to do for them to succeed with the field of science (Bayat, 2002).

This project exploited various bioinformatics tools housed within the GENI-ACT platform to determine if *Mrub_0860* and *E. coli b2892, Mrub_0701* and *E. coli b2562,* and *Mrub_2285* and *E. coli b3863* are orthologs, respectively. Databases such as Ecocyc (Keseler *et al*., 2013) are dedicated to the study of the *Escherichia coli* K-12 MG1655 genome. Ecocyc provides literature-based curation of the entire genome, and of transcriptional regulation transporters, and metabolic pathways. Due to these resources, *E. coli* K12 MG1655 is used as a "positive control" because we know much more about *E. coli* as it is the model organism.

By using various bioinformatics tools, I could identify similarities and differences between the amino acid/nucleotide sequences between two proteins/genes of interest. The output of many of these tools is an Expect Value (E-value). A low value between two sequences is indicative of a high sequence similarity, which can be implied that the two sequences have functional similarity. These data can therefore lead to the assumption that two sequences are orthologous to each other and that they have the same function.

Hypothesis

I predict that *Mrub_0860* and *E. coli b2892*, *Mrub_0701* and *E. coli b2562*, and *Mrub_2285* and *E. coli b3863* are orthologs, respectively.

METHODS

I started this research project using the instructions on the GENI-science site (Scott, 2016b). I decided to do research on DNA repair since I have done research on Non-homologous End Joining, another type of DNA repair. When I searched for the *M. ruber* and *E. coli* KEGG pathways to determine what proteins were present, two genes in the *E. coli* pathway were not predicted in the *M. ruber* RecFOR pathway. *T. thermophilus*, an organism in the same phylum as *M. ruber*, had the same set of genes as *E. coli*, however. Next, I did a protein BLAST between the putative *T. thermophilus* orthologs and the *M. ruber* genome. The missing genes were pulled from the *M. ruber* genome. Lab notebook pages were created on GENI-ACT for the three *E. coli* and three putative *M. ruber* orthologs.

Using the different bioinformatics housed within the GENI-ACT platform (Scott, 2016a), the bioinformatics data was complied for each gene. Within the eight modules – Basic Information, Sequence-based Similarity Data, Cellular Localization Data, Alternative Open Reading Frame, Structure-based Evidence, Enzymatic Function, Duplication and Degradation and Horizontal Gene Transfer – I compared the outputs of the various tools between the *E. coli* genes and the *M. ruber* genes to determine if these genes are orthologous to each other. I started off with the basic information module to find the nucleotide and amino acid sequences that must be used throughout the rest of the

modules. BLAST was used in order to find the most similar sequences to that of each gene, whether it is *E. coli* and *M. ruber* (Madden, 2002). Of the top 250 hits of similar sequences for each BLAST, 15 of them were used to make a T-coffee multiple sequence alignment (Notredame *et al*, 2000). The sequences were also used to create a Weblogo, which is an illustration of the most conserved amino acids between the 15 sequences (Crooks *et al*, 2004). TMHMM (Krogh & Rapacki, 2016), SignalP (Petersen *et al*, 2011), LipoP (Juncker *et al*, 2003), PSORT-B (Yu *et al*, 20010) were used to determine the cellular location of the protein within cells. The JGI's platform IMG was used to determine if there was a likely alternative start site for the *M. ruber* genes, which involved looking for other possible start codons in each sequence (Markowitz *et al*, 2012). The Opening Reading Frame module was completely disregarded for the *E. coli* genes. TIGRFAM (Haft *et al*, 2001), Pfam (Finn *et al*, 2016), and Protein Database (PDB) (Berman *et al*, 2000) compared the structural similarities between the different amino acid sequences, and identified the applicable protein families and/or domains, and, in the case of PDB and Pfam, produced a sequence alignment to a consensus sequence. KEGG pathways (Kanehisa *et al*., 2016a). Metacyc/Ecocyc pathways (Keseler *et al.*, 2013) and E.C. numbers (Artimo *et al*., 2012) provided information on the enzymatic function of the proteins. BLAST and/or KEGG were used to determine if paralogs were present in the genome for each gene of interest. The same 15 sequences used for T-coffee were used to make a phylogenetic tree to conclude if horizontal gene transfer could be an option. Within the same module, the use of JGI IMG website was utilized to bring up the ortholog neighborhood for the *M. ruber* genes to determine if the proteins were involved

within an operon. For the *E. coli* genes, colored by KEGG was used instead of ortholog neighborhoods.

RESULTS

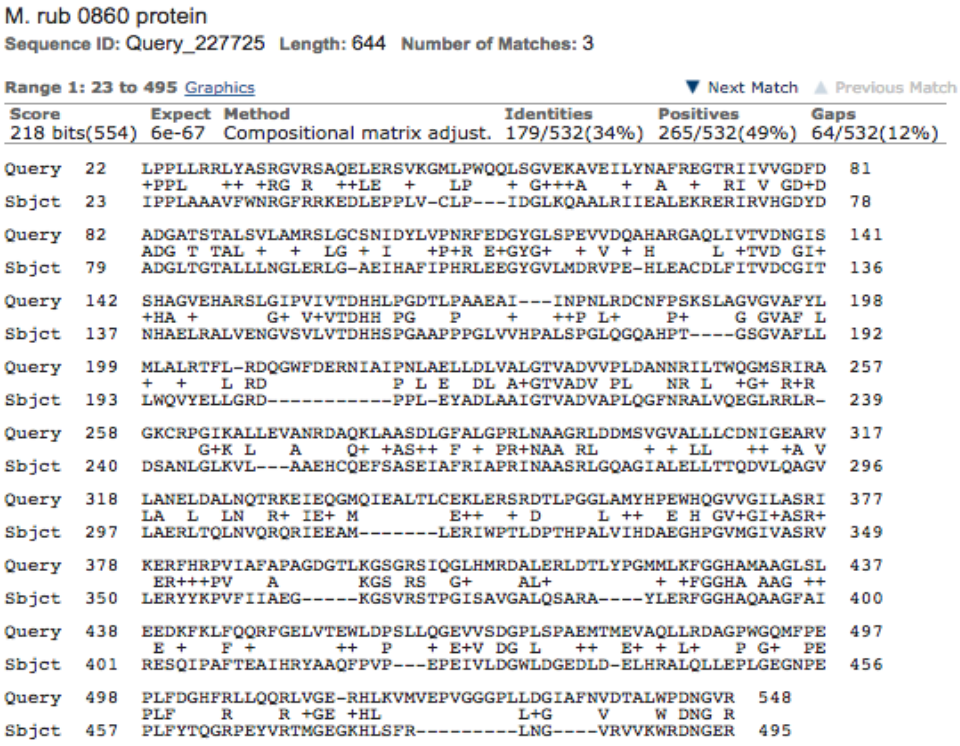
Table 1 summarizes the results from the bioinformatics tools that were used to compare *E. coli b2892* gene to *Mrub_0860*. The BLAST between *Mrub_0860* and *E. coli b2892* resulted in a bit score of 218 with an E-value of 6e-67. The CCD identified the same COG number (COG0608) and name (single-stranded DNA specific exonuclease) for both proteins; the low E-values indicate strong sequence similarity to the COG hit. The cellular localization data from various databases (SignalP, TMH, LipoP and PSORT-B) predicts that both proteins are found within the cytoplasm of the cell, and neither possesses a cleavage site. The COG hit and cellular localization suggests that these two genes are orthologs. The TIGRFAM hits for these two genes are also evidence that they are orthologous with the same hit, TIGR00644, being a single-stranded-DNA specific exonuclease – RecJ protein. The Pfam hit did confirm that both proteins belong to the same protein families: PF01368 (DHH phosphatase family) and PF02272 (DHHA1 domain). The protein database did not pull the same protein domains for each protein; however, it was found that the proteins did have the same enzyme commission number of 3.1.11.6. Both of the genes are a part of the RecFOR pathway within homologous recombination for prokaryotes.

Table 1. *Mrub*_0860 gene orthologous to *E. coli* b2892 gene

Bioinformatics Tool	*M. ruber Mrub_0860* gene	*E. coli* b2892 gene
BLAST	Score: 218 E-value: 6e-67	
CDD Data (COG)	COG0608 – Single stranded DNA-specific exonuclease E-value: 1.04e-106	COG0608 – Single stranded DNA-specific exonuclease (2nd hit) E-value: 9.13e-165
Cellular Localization	Cytoplasm of the cell	
TIGRfam – protein family	TIGR00644 – single-stranded-DNA specific exonuclease RecJ	
	E-value: 1.3e-130	E-value: 3.3e-279
Pfam – protein family	PF01368 (DHH phosphatase family) PF02272 (DHHA1 domain)	
	E-value: 2e-10 E-value: 6e-07	E-value: 6.1e-11 E-value: 1.1e-15
Protein Database	2ZXO Crystal structure of RecJ from *Thermus thermophiles* HB8 E-value: 4.43672e-153	5F54 Structure of RecJ complexed with dTMP E-value: 1.1413e-61
Enzyme commission number	3.1.11.6 – Exodeoxyribonuclease VII	
KEGG Pathway map	Homologous Recombination Prokaryotic Pathway	

Figure 2 is the result of a protein BLAST between two sequences, that of *Mrub_0860* and *E. coli b2892* (Madden, 2002). The two sequences have 34% of the amino acids that are the same between the sequences. Altogether there are 179 amino acids that are exactly the same out of 532 amino acids shown. The E-value is 6e-67, which is very close to being zero. Therefore, these sequences are highly conserved and

less likely to have similarities from random. Similarities shown in Table 1 and Figure 2 are the first piece of evidence that support the hypothesis of *Mrub_0860* and *E. coli b2892* being orthologs.



M. rub 0860 protein
Sequence ID: Query_227725 Length: 644 Number of Matches: 3

Range 1: 23 to 495 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
218 bits(554)	6e-67	Compositional matrix adjust.	179/532(34%)	265/532(49%)	64/532(12%)

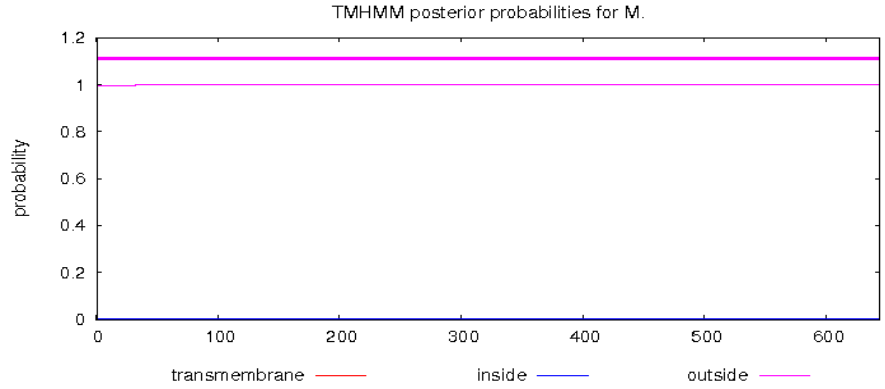
```
Query  22   LPPLLRRLYASRGVRSAQELERSVKGMLPWQQLSGVEKAVEILYNAFREGTRIIVVGDFD  81
            +PPL   ++ +RG R  ++LE  +   LP   + G+++A   +  A  +  RI V GD+D
Sbjct  23   IPPLAAAVFWNRGFRRKEDLEPPLV-CLP---IDGLKQAALRIIEALEKRERIRVHGDYD  78

Query  82   ADGATSTALSVLAMRSLGCSNIDYLVPNRFEDGYGLSPEVVDQAHARGAQLIVTVDNGIS  141
            ADG T TAL +  +  LG + I   +P+R E+GYG+  + V + H      L +TVD GI+
Sbjct  79   ADGLTGTALLLNGLERLG-AEIHAFIPHRLEEGYGVLMDRVPE-HLEACDLFITVDCGIT  136

Query  142  SHAGVEHARSLGIPVIVTDHHLPGDTLPAAEAI---INPNLRDCNFPSKSLAGVGVAFYL  198
            +HA +      G+ V+VTDHH PG   P    +   ++P L+    P+    G GVAF L
Sbjct  137  NHAELRALVENGVSVLVTDHHSPGAAPPPGLVVHPALSPGLQGQAHPT----GSGVAFLL  192

Query  199  MLALRTFL-RDQGWFDERNIAIPNLAELLDLVALGTVADVVPLDANNRILTWQGMSRIRA  257
            +  +   L RD           P L E  DL A+GTVADV PL   NR L  +G+ R+R
Sbjct  193  LWQVYELLGRD-----------PPL-EYADLAAIGTVADVAPLQGFNRALVQEGLRRLR-  239

Query  258  GKCRPGIKALLEVANRDAQKLAASDLGFALGPRLNAAGRLDDMSVGVALLLCDNIGEARV  317
               G+K L   A    Q+ +AS++ F + PR+NAA RL    + + LL   ++ +A V
Sbjct  240  DSANLGLKVL---AAEHCQEFSASEIAFRIAPRINAASRLGQAGIALELLTTQDVLQAGV  296

Query  318  LANELDALNQTRKEIEQGMQIEALTLCEKLERSRDTLPGGLAMYHPEWHQGVVGILASRI  377
            LA  L  LN  R+ IE+ M        E++  + D     L ++  E H GV+GI+ASR+
Sbjct  297  LAERLTQLNVQRQRIEEAM-------LERIWPTLDPTHPALVIHDAEGHPGVMGIVASRV  349

Query  378  KERFHRPVIAFAPAGDGTLKGSGRSIQGLHMRDALERLDTLYPGMMLKFGGHAMAAGLSL  437
             ER+++PV   A       KGS RS  G+    AL+          + +FGGHA AAG ++
Sbjct  350  LERYYKPVFIIAEG-----KGSVRSTPGISAVGALQSARA----YLERFGGHAQAAGFAI  400

Query  438  EEDKFKLFQQRFGELVTEWLDPSLLQGEVVSDGPLSPAEMTMEVAQLLRDAGPWGQMFPE  497
             E +   F +        ++  P   + E+V DG L   ++  E+ + L+   P G+  PE
Sbjct  401  RESQIPAFTEAIHRYAAQFPVP---EPEIVLDGWLDGEDLD-ELHRALQLLEPLGEGNPE  456

Query  498  PLFDGHFRLLQQRLVGE-RHLKVMVEPVGGGPLLDGIAFNVDTALWPDNGVR  548
            PLF   R    R +GE +HL           L+G    V    W DNG R
Sbjct  457  PLFYTQGRPEYVRTMGEGKHLSFR---------LNG----VRVVKWRDNGER  495
```

Figure 2. *Mrub_0860* and *E. coli b2892* similar protein sequence. Sequence alignment was completed using the bioinformatics tool - protein BLAST. The query sequence is *E. coli* and the subject is *M. ruber*. (Madden, 2002)

Figure 3 shows the results for the THM plots for *Mrub_0860* and *E. coli b2892* (Krogh & Rapacki, 2016). Figure 3A does not show any transmembranes helices in *M. ruber*. Figure 3B shows a short red peak, but the height does not reach the cutoff for a TMH. The numbers of predicted transmembrane helices for both genes are predicted to be zero. Therefore, both genes are predicted to be present in the cytoplasm instead of the membrane.

```
# M. Length: 644
# M. Number of predicted TMHs:  0
# M. Exp number of AAs in TMHs: 0.03991
# M. Exp number, first 60 AAs:  0.0374
# M. Total prob of N-in:        0.00266
M.      TMHMM2.0        outside      1   644
```

Figure 3A




```
# E. Length: 577
# E. Number of predicted TMHs:  0
# E. Exp number of AAs in TMHs: 7.32113
# E. Exp number, first 60 AAs:  0
# E. Total prob of N-in:        0.37293
E.      TMHMM2.0        outside      1   577
```

Figure 3B

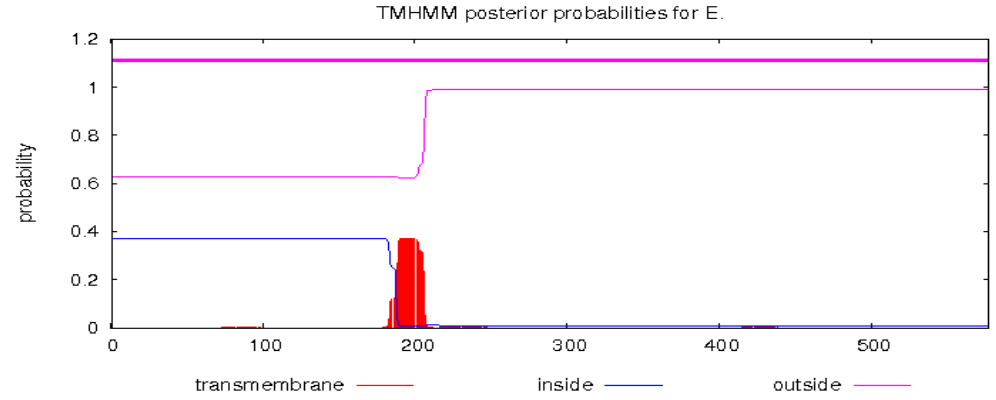


Figure 3. *Mrub_0860* and *E. coli b2892* do not have TMH regions. Figure 3A shows the TMHMM for *Mrub_0860* and Figure 3B shows the TMHMM for *E. coli b2892*. Due to the TMHMM data from TMHMM Server v 2.0, it is predicted that that both proteins have a cytoplasmic location (Krogh & Rapack, 2016).

The figures below, 4A and 4B, are SignalP plots for *Mrub_0860* and *E. coli b2892* (Petersen *et al*. 2011). SignalP is a helpful tool in concluding whether there is a protein cleavage site, which would indicate that the protein is either attached or passes through the cell membrane. This is determined by the D-value, a calculation via S-score

(green line) and Y-score (blue line), and a cutoff value (pink line). Both of the D-values were found to be lower than that of the cutoff value. *Mrub_0860* has a cutoff of 0.570 and a value of 0.130, while *E. coli b2892* has a cutoff of 0.570 and a value of 0.100. Thus, these proteins do not have any cleavage sites, again confirming their cytoplasmic location.

Figure 4A

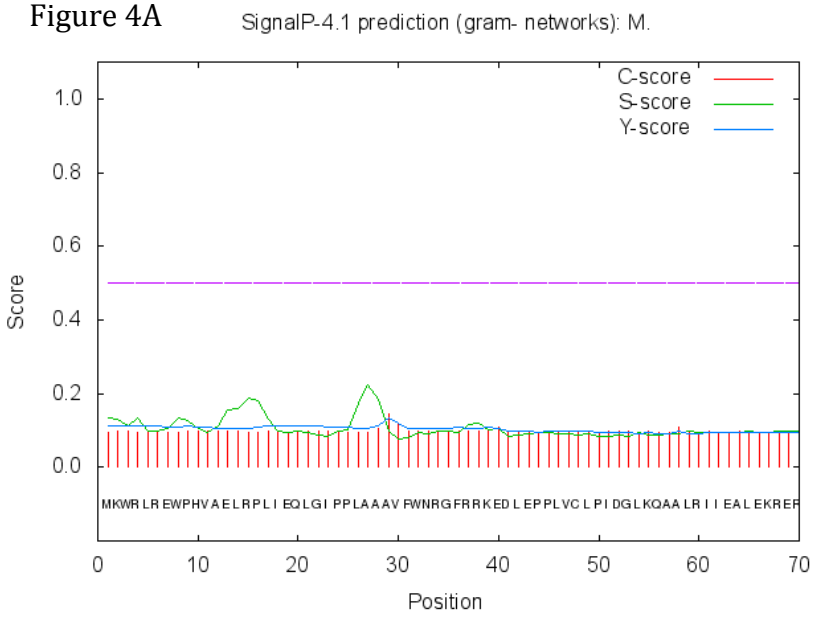


Figure 4B

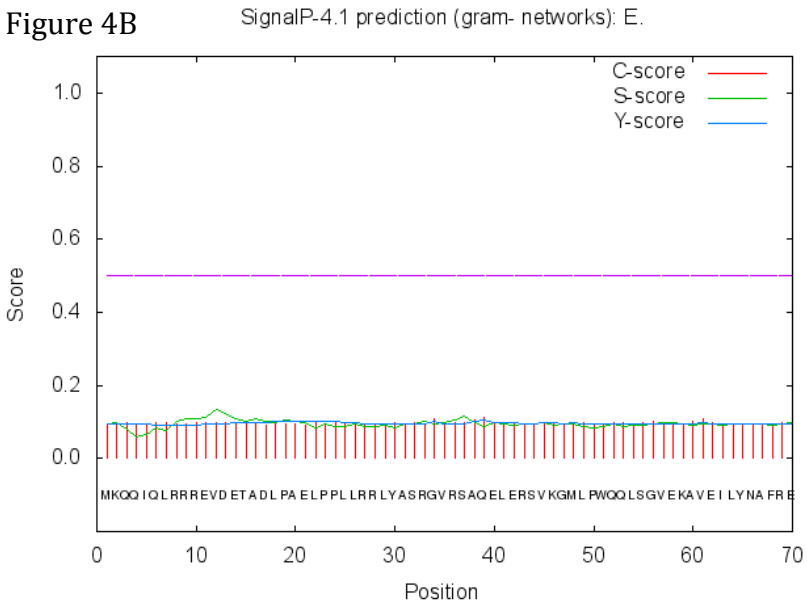


Figure 4. Lack of cleavage sites within *Mrub_0860* and *E. coli b2892*. Figure 4A is *Mrub_0860* and Figure 4B is *E. coli b2892*. The cutoff values were both below the D values and a result of NO to signal peptides. These figures were created via Signal P server 4.1 (Petersen *et al*. 2011).

Figure 5 is the homologous recombination KEGG RecFOR pathway showing the *M. ruber* (Figure 5A) and *E. coli* (Figure 5B) sides. The green color is indicative of enzymes/proteins that are predicted to be present within each organism's genome. Both *M. ruber* and *E. coli* contain the RecJ protein, which is the single stranded DNA specific exonuclease.

Figure 5A

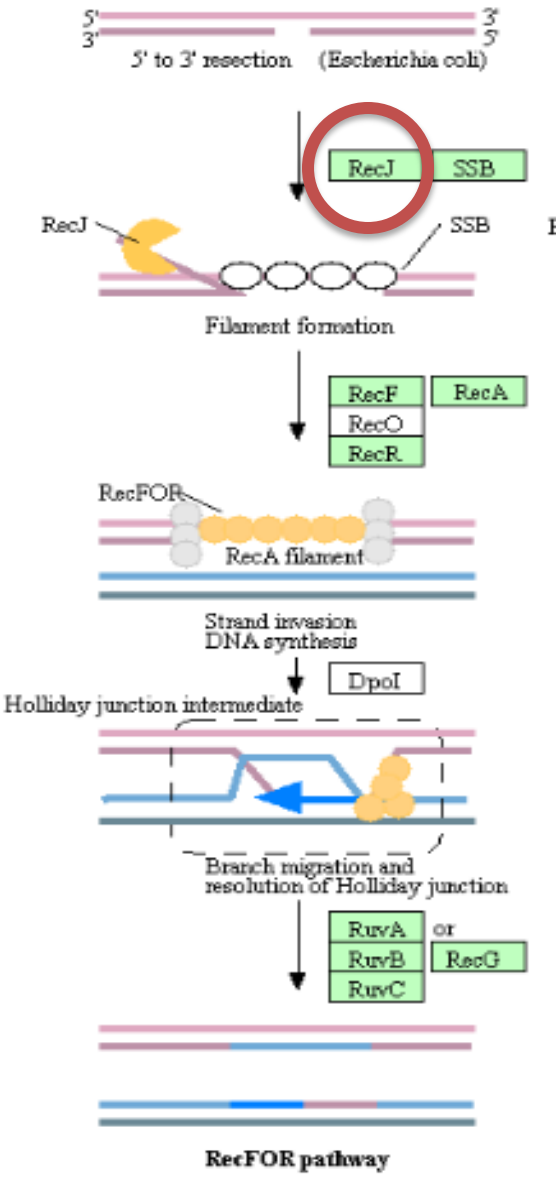


Figure 5B

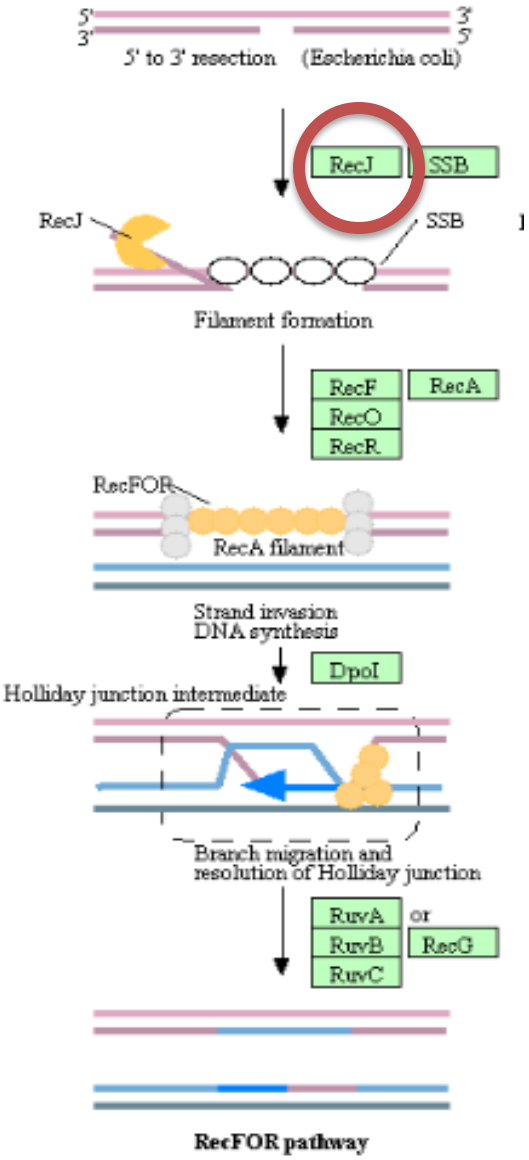


Figure 5. *Mrub_0860* and *E. coli b2892* are in the same recombination pathway - RecFOR pathway. Figure 5A shows the KEGG pathway for *Meiothermus ruber* and Figure 5B shows the KEGG pathway *Escherichia coli.* The KEGG database – The Kyoto Encyclopedia of Genes and Genomes – was utilized to locate these genes within the homologous recombination RecFOR pathway (Kanehisa *et al*., 2016b).

Figure 6 shows the pairwise alignments between the Pfam consensus sequence and *Mrub_0860* or *E. coli b2892*. Figure 6.1 is the alignment for the hit PF01368 – the DHH phosphatase family. The second alignment is shown in Figure 6.2 for the hit PF02272 – the DHHA1 domain. Figure 6.1A and 6.2A are the pairwise alignments for *Mrub_0860* and Figure 6.1B and 6.2B are the pairwise alignments are *E. coli b2892*. Using the Pfam site, it takes the sequences and compares them to hundreds of other different sequences in order to find which sequences are similar throughout genes that code for protein domains. From this comparison between the genes, it is relevant that they have many of the same conserved amino acids (as seen in the first and second rows with capital letters). Therefore, this is another piece of evidence that *Mrub_0860* and *E. coli b2892* are orthologous (Berman *et al*., 2000).

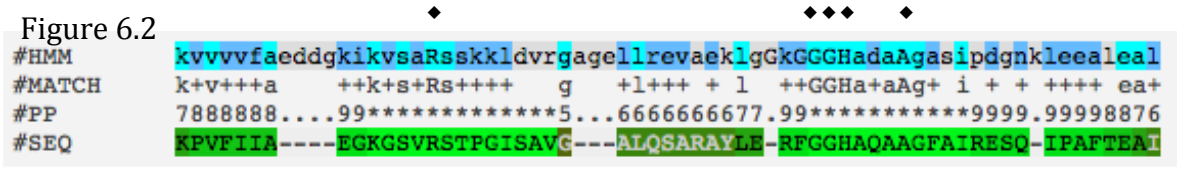
Figure 6.1

A

B

Figure 6.2

A



B

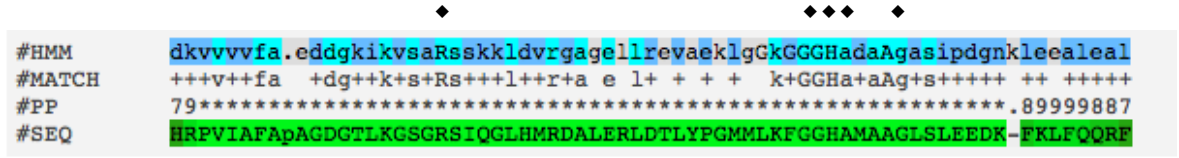


Figure 6. *Mrub_0860* and *E. coli b2892* have the same highly conserved amino acids in two different protein domains as designated by the ◆. Figure 6.1 is the comparison for the PF01368 – the DHH phosphatase family and Figure 6.2 is the comparison for the PF02272 – the DHHA1 domain. These pairwise alignments were constructed via Pfam (Berman *et al*., 2000).

Figure 7 is a map of the chromosome that flanks our gene of interest, colored by their predicted KEGG pathway. The different colors in the figures below are indicative of these genes not being a part of an operon. The *Mrub_0860* is identified by an orange color and there are no other genes near it with the same orange color. The *E. coli b2892* gene is green with no other green genes near it. The red line below or above genes is indicative of the gene of interest. Therefore, Color by KEGG resulted in more evidence that these two genes, *Mrub_0860* and *E. coli b2892* are orthologous (Kanehisa *et al*., 2016a).

7.1

7.2

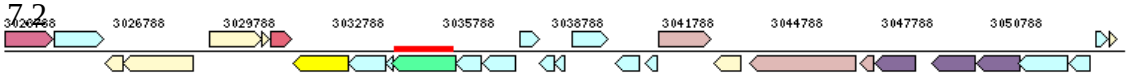


Figure 7. *Mrub_0860* and *E. coli b2892* genes are not a part of an operon. 7.1 is *Mrub_0860* and 7.2 is *E. coli b2892* Colored by KEGG and Ortholog Neighborhoods were used in order to view these images (Kanehisa et al., 2016a).

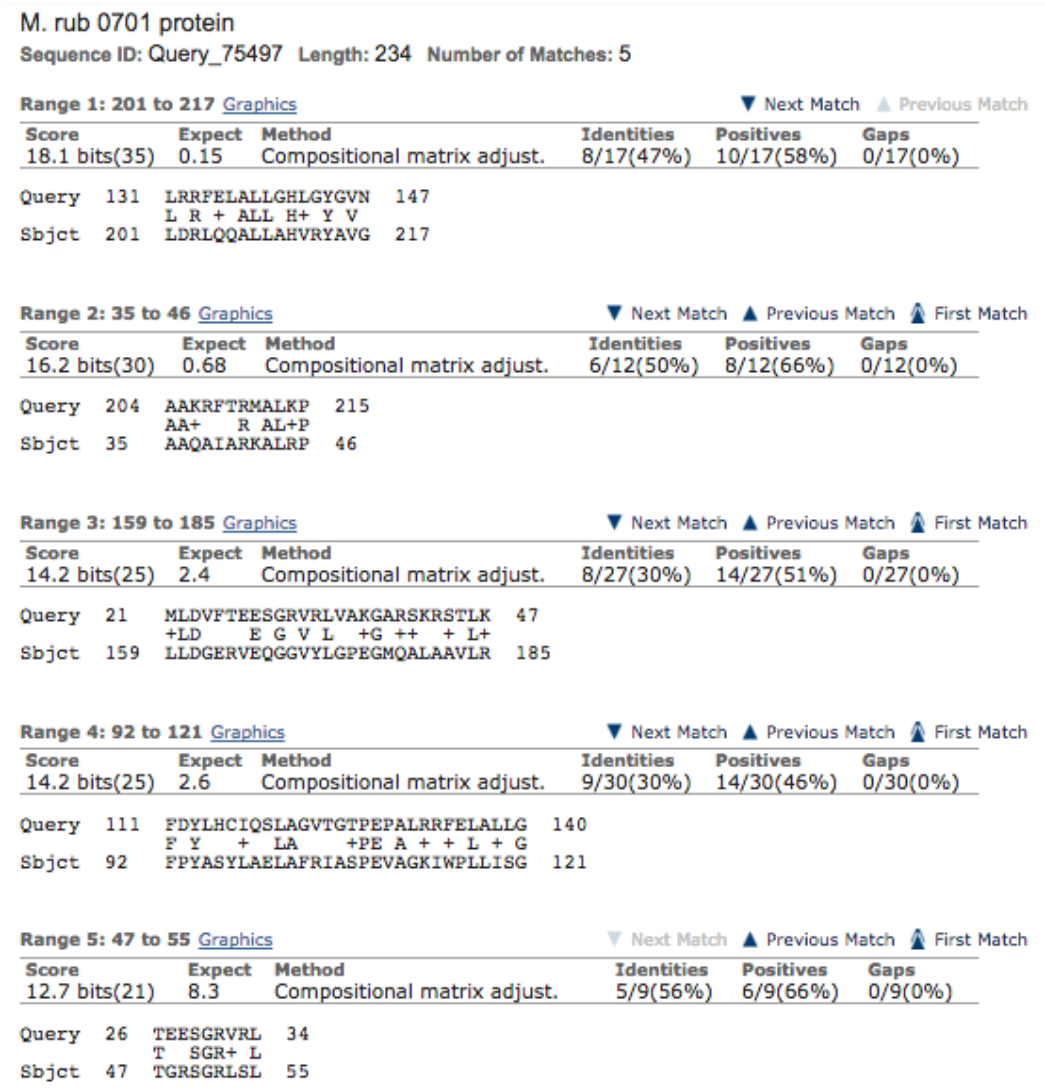
Table 2 summarizes the results from the bioinformatics tools that were used to compare *E. coli b2562* gene to *Mrub_0701*. The BLAST between *Mrub_0701* and *E. coli b2562* resulted in a low bit score and a high E-value. The CCD identified same COG number (COG1381) and name (recombinational DNA repair protein) for both proteins; the low E-values indicate a strong sequence similiarity to the COG hit. However, the E-values for these two genes are much higher than the previous genes. The cellular localization data from various databases (SignalP, TMH, LipoP and PSORT-B) predicts that both proteins are found within the cytoplasm of the cell, and neither possesses cleavage site. The COG hit and cellular localization suggests that these two genes are orthologs. The TIGRFAM hits are hard to compare with these two genes since there was no hit for one of them. The Pfam hit did confirm that both proteins have the same N terminal domain for the recombination protein (PF11967). The protein database pulled two different protein domains but both were portions of the RecO protein. An enzyme commission number was not found for either of proteins. Both of the genes are a part of the RecFOR pathway within homologous recombination for prokaryotes.

Table 2. *Mrub*_0701 gene orthologous to *E. coli* b2562 gene

Bioinformatics Tool	*M. ruber* Mrub_0701 gene	*E. coli* b2562 gene
BLAST	Score: >12 E-value: >0	
CDD Data (COG)	COG1381 – Recombinational DNA repair protein (RecF pathway)	
	E-value: 9.47e-13	E-value: 5.71e-80
Cellular Localization	Cytoplasm of the cell	
TIGRfam – protein family	No TIGRFAM hit	TIGR00613 – RecO: DNA repair protein RecO E-value: 5.5e-43
Pfam – protein family	PF11967 (Recombination protein O N terminal)	
	E-value: 4e-07	E-value: 1.2e-22
Protein Database	4JCV Crystal structure of the RecO complex in an open comformation E-value: 4.50245e-16	3Q8D E. coli RecO complex with SSB C-terminus E-value: 0.0
Enzyme commission number	No EC number	
KEGG Pathway map	Homologous Recombination Prokaryotic Pathway	

Figure 8 is the result of a protein BLAST between two sequences, that of *Mrub_0701* and *E. coli b2562* (Madden, 2002). The two sequences have multiple ranges, however, none of the E-values are significant because they are all larger than zero. The E-value of the portions are fairly large compared to other sequence E-values. Therefore, it is more likely that the sequences share similarities due to random. This is interesting because the same CDD data and Pfam information were determined but the BLAST has many differences. The protein BLAST for all sequences with *Mrub_0701* results in a first hit of DNA recombination protein RecO for *Meiothermus cereberus* with an E-value of 1e-135 and a second hit of DNA recombination protein RecO for *Meiothermus rufus* with

an E-value of 3e-122. This is indicative of *Mrub_0701* being the RecO protein, but it is not similar enough to match with *E. coli*. Since these organisms are phylogenetically different there is a possibility that both Mrub_0701 and E. coli b2562 are RecO, considering the other similarities seen in Table 2.



M. rub 0701 protein

Sequence ID: Query_75497 **Length:** 234 **Number of Matches:** 5

Range 1: 201 to 217 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
18.1 bits(35)	0.15	Compositional matrix adjust.	8/17(47%)	10/17(58%)	0/17(0%)

```
Query  131  LRRFELALLGHLGYGVN  147
            L R + ALL H+ Y V
Sbjct  201  LDRLQQALLAHVRYAVG  217
```

Range 2: 35 to 46 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
16.2 bits(30)	0.68	Compositional matrix adjust.	6/12(50%)	8/12(66%)	0/12(0%)

```
Query  204  AAKRFTRMALKP  215
            AA+    R AL+P
Sbjct  35   AAQAIARKALRP  46
```

Range 3: 159 to 185 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
14.2 bits(25)	2.4	Compositional matrix adjust.	8/27(30%)	14/27(51%)	0/27(0%)

```
Query  21   MLDVFTEESGRVRLVAKGARSKRSTLK  47
            +LD     E G V L   +G ++   + L+
Sbjct  159  LLDGERVEQGGVYLGPEGMQALAAVLR  185
```

Range 4: 92 to 121 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
14.2 bits(25)	2.6	Compositional matrix adjust.	9/30(30%)	14/30(46%)	0/30(0%)

```
Query  111  FDYLHCIQSLAGVTGTPEPALRRFELALLG  140
            F Y    +   LA     +PE A + + L + G
Sbjct  92   FPYASYLAELAFRIASPEVAGKIWPLLISG  121
```

Range 5: 47 to 55 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
12.7 bits(21)	8.3	Compositional matrix adjust.	5/9(56%)	6/9(66%)	0/9(0%)

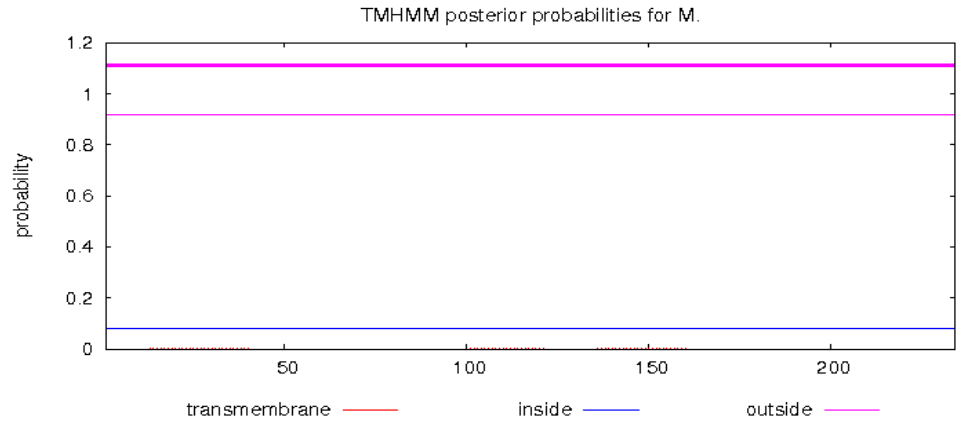
```
Query  26  TEESGRVRL  34
           T  SGR+ L
Sbjct  47  TGRSGRLSL  55
```

Figure 8 (above). *Mrub_0701* and *E. coli b2562* similar protein sequence. Sequence alignment was completed using the bioinformatics tool - protein BLAST. The query sequence is *E. coli* and the subject is *M. ruber* (Madden, 2002).

Figure 9 shows the results for the THM plots for *Mrub_0701* and *E. coli b2562* (Krogh & Rapack, 2016). Figure 9A does not show transmembranes in *M. ruber*. Figure 3B shows a short red peak, but the height does not reach the cutoff for the TMH. The numbers of predicted transmembrane helices for both genes are predicted to be zero. Therefore, both genes are predicted to be present in the cytoplasm instead of the membrane.

```
# M. Length: 234
# M. Number of predicted TMHs:  0
# M. Exp number of AAs in TMHs: 0.05299
# M. Exp number, first 60 AAs:  0.01437
# M. Total prob of N-in:        0.08200
M.      TMHMM2.0        outside      1   234
```

Figure 9A




```
# E. Length: 242
# E. Number of predicted TMHs:  0
# E. Exp number of AAs in TMHs: 0.76504
# E. Exp number, first 60 AAs:  0.00187
# E. Total prob of N-in:        0.05524
E.      TMHMM2.0        outside      1   242
```

Figure 9B

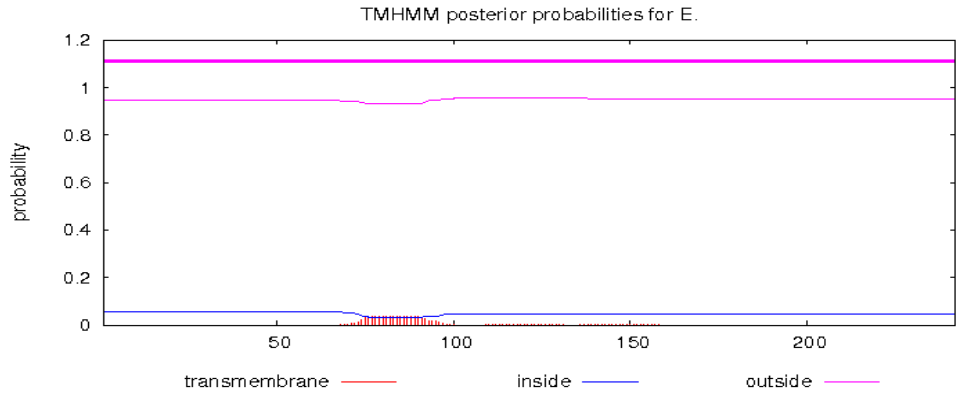


Figure 9. *Mrub_0701* and *E. coli b2562* do not have TMH regions. Figure 9A shows the TMHMM for *Mrub_0701* and Figure 3B shows the TMHMM for *E. coli b2562*. Due to the TMHMM data from TMHMM Server v 2.0, it is predicted that that both proteins have a cytoplasmic location (Krog & Rapack, 2016).

The figures below, 10A and 10B, are SignalP plots for *Mrub_0701* and *E. coli b2562* (Petersen *et al*., 2011). Signal P is helpful in concluding if a protein cleavage site is present, indicating that the protein is either attached to passes through the cell membrane. This is determined by the D-value, a calculation via S-score (green line) and Y-score (blue line), and a cutoff value (pink line), cleavage sites can be determined. Both of the D-values were found to be lower than that of the cutoff value. *Mrub_0860* has a cutoff of 0.570 and a value of 0.123, while *E. coli b2892* has a cutoff of 0.570 and a value of 0.127. Thus, these proteins do not have any cleavage sites, again confirming their cytoplasmic location.

Figure 10A

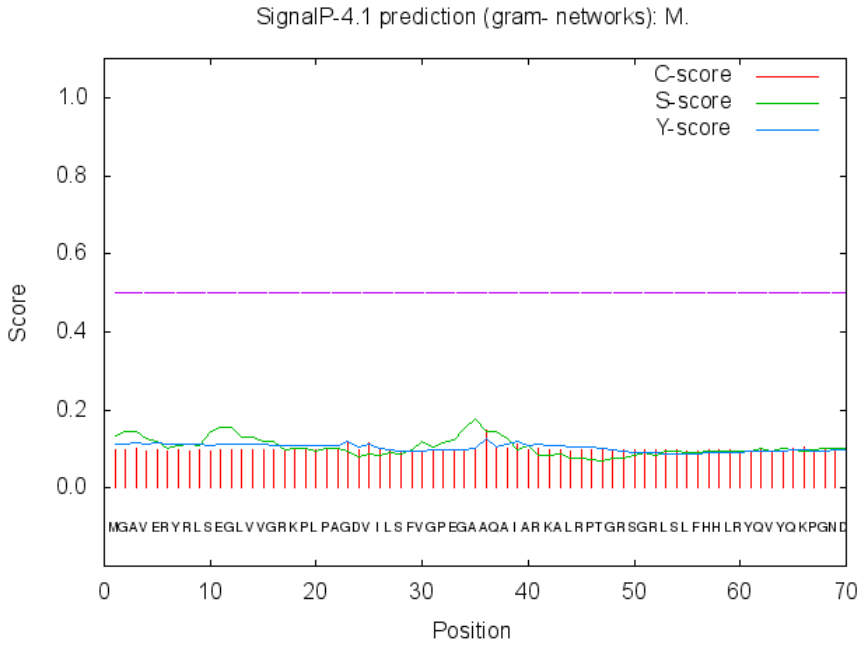


Figure 10B

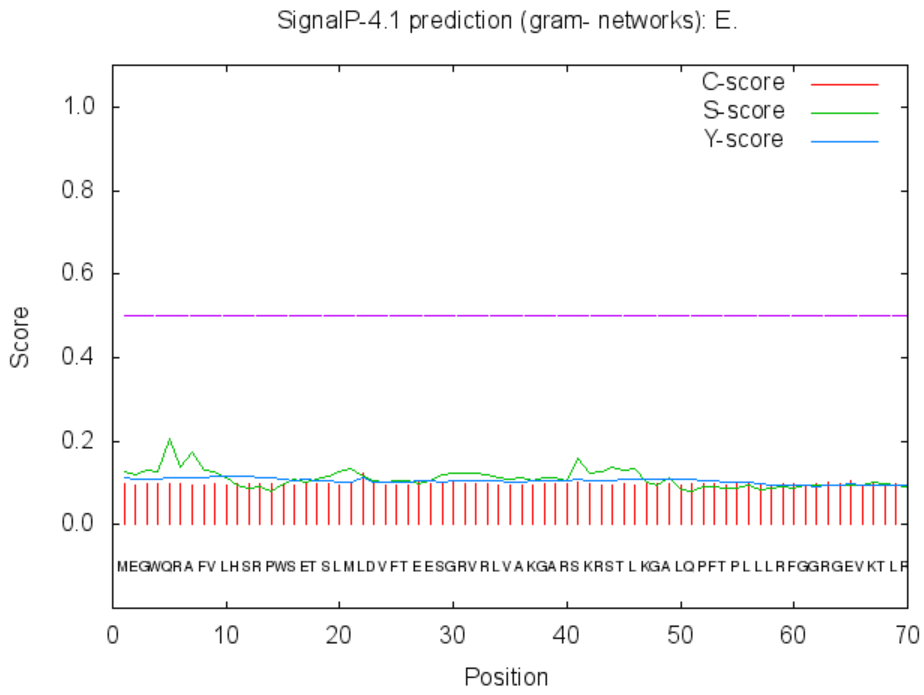


Figure 10. Lack of cleavage sites within *Mrub_0701* and *E. coli b2562*. Figure 10A is *Mrub_0701* and Figure 10B is *E. coli b2562*. The cutoff values were both below the D values and a result of NO to signal peptides. These figures were created via Signal P server 4.1 (Petersen *et al.*, 2011).

Figure 11 is the homologous recombination KEGG RecFOR pathway showing the *M. ruber* (Figure 11A) and *E. coli* (Figure 11B) sides (Kanehisa *et al*., 2016b). The green color is indicative of enzymes/proteins that are present within each organism's genome. The *M. ruber* pathway shows *Mrub_0701* is not present within the RecFOR pathway as the RecO protein is shown in white. However, with the preliminary BLAST between *T. thermophilus* and *M. ruber*, the BLAST indicated that the *M. ruber* does actually have the *Mrub_0701* protein. Therefore, the RecO protein is found within the *M. ruber* RecFOR pathway. *E. coli* contains the gene *b2562* which is the RecO protein, which is a gap repair protein.

Figure 11A

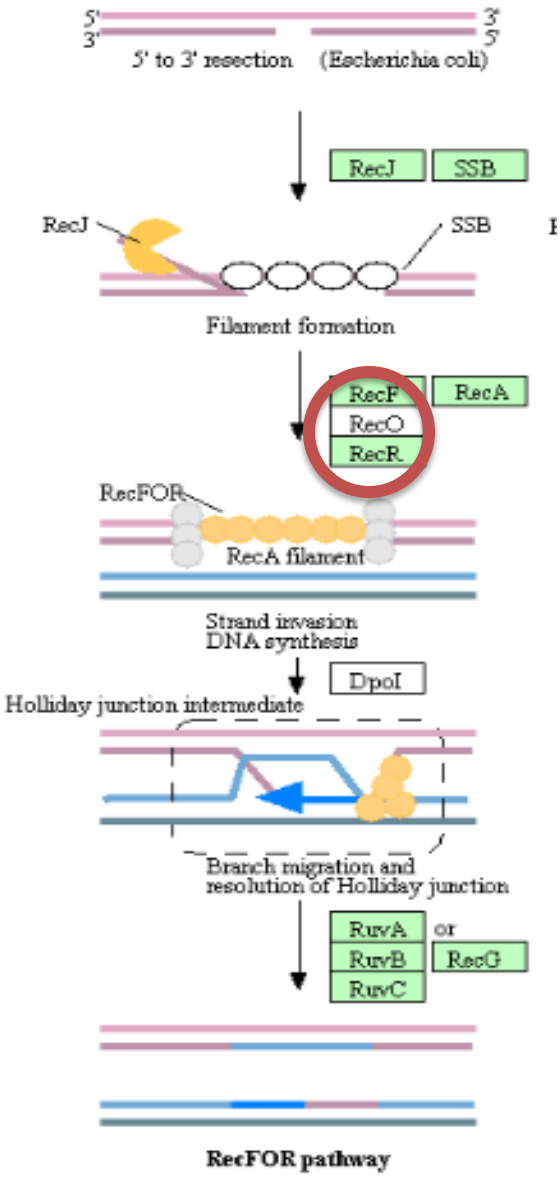


Figure 11B

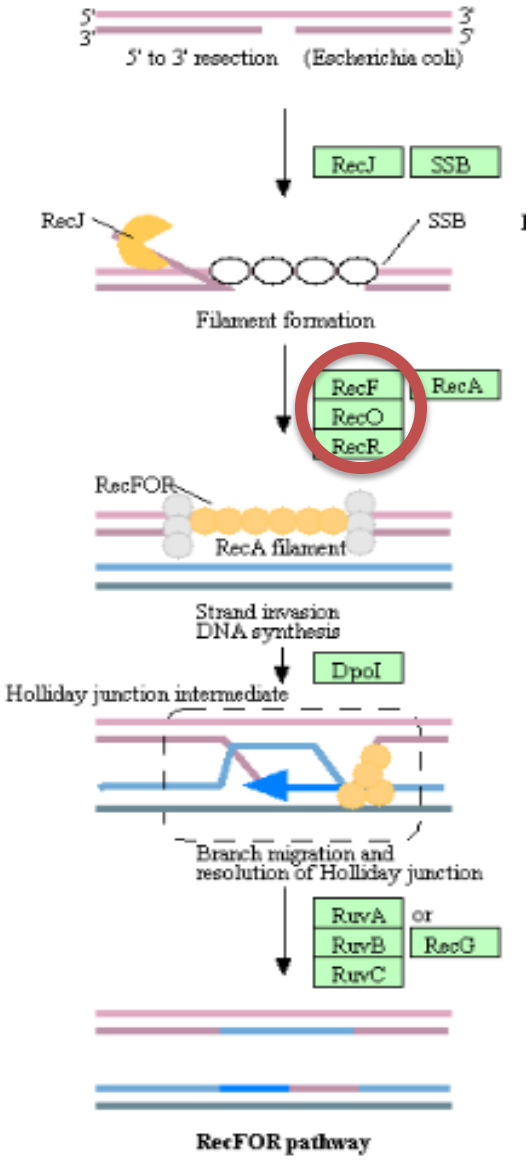


Figure 11. *Mrub_0701* and *E. coli b2562* are in the same recombination pathway - RecFOR pathway. Figure 11A shows the KEGG pathway for *Meiothermus ruber* and Figure 11B shows the KEGG pathway *Escherichia coli.* The KEGG database – The Kyoto Encyclopedia of Genes and Genomes – was utilized to locate these genes within the homologous recombination RecFOR pathway (Kanehisa *et al.*, 2016b).

Figure 12 shows the pairwise alignments between the Pfam consensus sequence and *Mrub_0701* or *E. coli b2562*. Figure 12.1 is the alignment for the hit PF11967 – the Recombination protein O N terminal. The second alignment is shown in Figure 12.2 for the hit PF02565 – the Recombination protein O C terminal. This second hit was only found for *E. coli*. Figure 12.1A is the pairwise alignment for *Mrub_0860* and Figure 12.1B and 12.2B are the pairwise alignments for *E. coli b2892*. There are similar sequences found between the alignments found in Figure 12.1, showing many of the same conserved amino acids (as seen in the first and second rows with capital letters). Therefore, this is another piece of evidence that *Mrub_0701* and *E. coli b2562* are orthologous (Berman *et al*., 2000).

Figure 12.1

A

```
#HMM      teaiVLksrkygEsdlivtlfTrehGkvsalargarksksrlrallqpftllevqlr.kkgrgLktlkqaelik
#MATCH    +e++V  ++++   d+i+++   e G  +a+ar a ++ +r +  l++f++l++q++ k g++L tl+q+el+
#PP       589999999999999*****99999.*********999987766.**************************985
#SEQ      SEGLVVGRKPLPAGDVILSFVGPE-GAAQAIARKALRPTGRSGR-LSLFHHLRYQVYqKPGNDLPTLTQVELVG
```

B

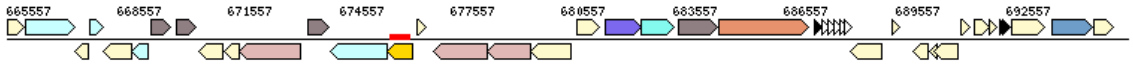
```
#HMM      kteaiVLksrkygEsdlivtlfTrehGkvsalargarksksrlrallqpftllevqlrkkgrgLktlkqaelik
#MATCH     ++a+VL+sr+++E++l++++fT+e+G+v+++a+gar+++s l+++lqpft+l +++  ++++ ktl++ae ++
#PP       689*****************************************************.************987
#SEQ      WQRAFVLHSRPWSETSLMLDVFTEESGRVRLVAKGARSKRSTLKGALQPFTPLLLRFG-GRGEVKTLRSAEAVS
```

Figure 12. *Mrub_0701* and *E. coli b2562* have the some highly conserved amino acids in one protein domain as designated by the ◆. Figure 12.1 is the comparison for the PF11967 - Recombination protein O N terminal and Figure 12.2 is the sequence alignment for E. coli for the Recombination protein O C terminal. These pairwise alignments were constructed via Pfam (Berman *et al.*, 2000).

Figure 13 is a map of the chromosome that flanks our gene of interest, colored by their predicted KEGG pathway. The different colors in the figures are indicative of these genes not being a part of an operon. The *Mrub_0701* gene is shown as an orange color,

and there are no other genes near these that are the same orange color. The *E. coli 2562* gene is shown by green color, with no other green genes near it. The red line that appears above these indicates the gene of interest. Therefore, the Color by KEGG results in more evidence that these two genes, *Mrub_0701* and *E. coli b2562* are orthologous (Kanehisa *et al*., 2016a).

13.1



13.2

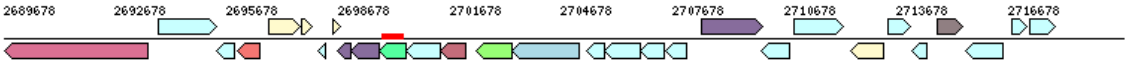


Figure 13. *Mrub_0701* and *E. coli b2562* genes are not a part of an operon. 13.1 is *Mrub_0701* and 13.2 is *E. coli b2562.* Colored by KEGG and Ortholog Neighborhoods were used in order to view these images (Kanehisa *et al.*, 2016a).

Table 3 summarizes the results from the bioinformatics tools that were used to compare *E. coli b3863* gene to *Mrub_2285*. The BLAST between *Mrub_2285* and *E. coli b3863* resulted in a bit score of 234 and an E-value of 1e-67. The CCD identified the same COG number (COG0749) and name (DNA polymerase I – 3'-5', exonuclease and polymerase domains) for both proteins; the low E-values indicate strong sequence similarity to the COD hit. The cellular localization data from various databases (SignalP, TMH, LipoP and PSORT-B) predicts that both proteins are found within the cytoplasm of the cell, and neither possesses a cleavage site. The COG hit and cellular localization suggests that these two genes are orthologs. The TIGRFAM hit was the same for both proteins, TIGF00593 – DNA polymerase I. The Pfam hit did confirm that both proteins

have the same N terminal domain and C terminal domain for the 5'-3' exonuclease with the same hits of PF02739 and PF01367. The protein database did pull two different protein domains; however, the same enzyme commission number for these two genes was 2.7.7.7, DNA directed DNA polymerase. Both of the genes are a part of the RecFOR pathway within homologous recombination for prokaryotes.

Table 3. *Mrub_2285* gene orthologous to *E. coli b3863* gene

Bioinformatics Tool	*M. ruber* Mrub_2285 gene	*E. coli* b3863 gene
BLAST	Score: 234 bits E-value: 1e-67	
CDD Data (COG)	COG0749 – DNA polymerase I – 3' – 5' exonuclease and polymerase domains	
	E-value: 2.92e-117	E-value: 0.00
Cellular Localization	Cytoplasm of the cell	
TIGRfam – protein family	TIGR00593 – DNA polymerase I E-value: 0.0	
Pfam – protein family	PF02739 (5'-3' exonuclease, N-terminal resolvase-like domain) PF01367 (5'-3' exonuclease, C-terminal SAM fold)	
	E-value: 1.4e-49 E-value: 1.3e-27	E-value: 2.7e-55 E-value: 1.8e-30
Protein Database	1BGX TAQ polymerase in complex with TP7, an inhibitory fab E-values: 0.0, 1.849e-64	1D8Y: Entity 2 containing Chain A – Crystal structure of the complex of DNA polymerase I Klenow Fragment with DNA E-value: 0.0
Enzyme commission number	2.7.7.7 – DNA directed DNA polymerase	
KEGG Pathway map	Homologous Recombination Prokaryotic Pathway	

Figure 14 is the result of a protein BLAST between two sequences, that of *Mrub_2285* and *E. coli b3863* (Madden, 2002). The two sequences had multiple ranges, meaning there is multiple parts of the sequence that match up, the first range being the most important. The E-value of the first range is 1e-67, however the other ranges have fairly small E-values also. Therefore, multiple low E-values indicate that there are more conserved amino acids between the sequences than those that are due to random. Similarities shown in Table 3 and Figure 14 are the first piece of evidence that support the hypothesis of *Mrub_2285* and *E. coli b3863* being orthologs.

M. rub 2285 protein

Sequence ID: Query_164977 **Length:** 1273 **Number of Matches:** 5

Range 1: 406 to 679 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
234 bits(597)	1e-67	Compositional matrix adjust.	130/275(47%)	177/275(64%)	6/275(2%)

```
Query  494  AGRYAAEDADVTLQLHLKMWPDLQKHKGPLN----VFENIEMPLVPVLSRIERNGVKIDP  549
            AG + ++  +     H  +W  LQ   G       +++ IE PL  VL+RIE  G+ ++
Sbjct  406  AGDWGSDPGERAQAAH-TLWETLQSRIGENPKIEWLYQQIEKPLSAVLARIEARGISLNA  464

Query  550  KVLHNHSEELTLRLAELEKKAHEIAGEEFNLSSTKQLQTILFEKQGIK-PLKKTPGGAPS  608
            + L   SEEL   ++ LE + H +AG  FN++S  QL+ +L+++  +  P +KT  G  S
Sbjct  465  EYLLQLSEELAKEISYLEAEIHRLAGRSFNVNSRDQLEVVLYDELKLSAPGRKTQTGKRS  524

Query  609  TSEEVLEELALDYPLPKVILEYRGLAKLKSTYTDKLPLMINPKTGRVHTSYHQAVTATGR  668
            T+   LEEL   +P+ + IL YR L KLKSTY D LP +I+PKTGR+HT ++Q  TATGR
Sbjct  525  TAASALEELLGQHPIIERILTYRELTKLKSTYLDPLPKLIHPKTGRLHTRFNQTGTATGR  584

Query  669  LSSTDPNLQNIPVRNEEGRRIRQAFIAPEDYVIVSADYSQIELRIMAHLSRDKGLLTAFA  728
            LSS DPNLQNIPVR E GRRIR+AF A     +V ADYSQIELR++AHLS D+ L+  F
Sbjct  585  LSSVDPNLQNIPVRTEIGRRIRKAFRAAPGMRLVVADYSQIELRVLAHLSGDENLINVFR  644

Query  729  EGKDIHRATAAEVFGLPLETVTSEQRRSAKAINFG  763
            EG+DIH  TAA +FGL  + V +EQRR+AK +  G
Sbjct  645  EGRDIHTQTAAWMFGLEPDKVGAEQRRAAKCLVEG  679
```

Range 2: 14 to 288 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
194 bits(493)	1e-54	Compositional matrix adjust.	116/284(41%)	163/284(57%)	10/284(3%)

```
Query  2    VQIPQ-NPLILVDGSSYLYRAYHAFPPLTNSAGEPTGAMYGVLNMLRSLIMQYKPTHAAV  60
            V++P+ + + LVDG    YR+Y AF  L  S GEPT A++G L  L  L+ +       V
Sbjct  14   VELPRPDRIWLVDGHHLAYRSYFAFEKLATSRGEPTQAIFGFLRTLLKLLKEDGDC-VIV  72

Query  61   VFDAKGKTFRDELFEHYKSHRPPMPDDLRAQIEPLHAMVKAMGLPLLAVSGVEADDVIGT  120
            VFD   +TFR + FE YK+ R   PDD   Q+E +  +V  MGL  L V G EADDVIGT
Sbjct  73   VFDPPTRTFRHDAFEEYKAGRAATPDDFHPQLEKIKELVDLMGLQRLEVPGYEADDVIGT  132

Query  121  LAREAEKAGRPVLISTGDKDMAQLVTPNITLINTMTNTILGPEEVVNKYGVPPELIIDFL  180
            LA++AE+ G PV I TGD+D  QL++  + ++       ++ P+ V  KYGV     +D+
Sbjct  133  LAKKAEQEGYPVRILTGDRDSFQLLSEAVQVM-LPDGRLMHPQAVQEKYGVSVAQWVDYR  191

Query  181  ALMGDSSDNIPGVPGVGEKTAQALLQGLGGLDTLYAEPEKIAGLSFRGAKTMAAKLEQNK  240
            +L+GDSSDN+PG  G+GEKTA  LLQ  G L+ LYA  E +       +  + A LE+++
Sbjct  192  SLVGDSSDNLPGAKGIGEKTAAKLLQEWGSLEGLYANLEAL-------SPKIRASLEESR  244

Query  241  EVAYLSYQLATIKTDVELELTCEQLEVQQPAAEELLGLFKKYEF  284
            +   LS  L+ I TD+ LEL       ++P    L    +K EF
Sbjct  245  DNVQLSRTLSLIHTDLPLELDFRDCHRREPDRGALREALEKLEF  288
```

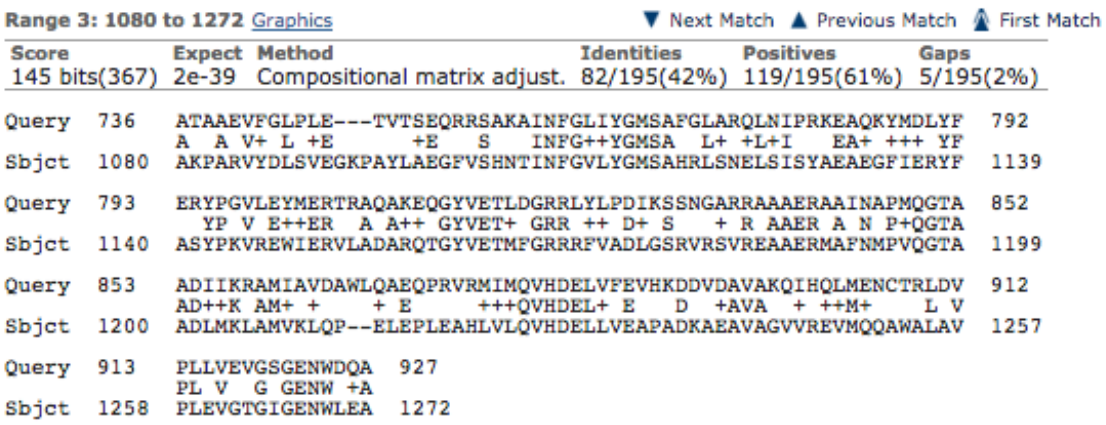


Figure 14. *Mrub_2285* and *E. coli b3863* similar protein sequence. Sequence alignment was completed using the bioinformatics tool - protein BLAST. The query sequence is that of *E. coli* and the subject is *M. ruber* (Madden, 2002).

Figure 15 shows the results for the THM plots for *Mrub_2285* and *E. coli b3863* (Krog & Rapack, 2016). Figure 15A does not show transmembranes helices for *M. ruber*. Figure 15B shows a very short red peak, but the height does not reach the cutoff for the TMH. The numbers of transmembrane helices for both genes are predicted to be zero. Therefore, both genes are predicted to be present in the cytoplasm instead of the membrane.

Figure 15A

```
# M. Length: 1273
# M. Number of predicted TMHs:  0
# M. Exp number of AAs in TMHs: 0.01028
# M. Exp number, first 60 AAs:  0
# M. Total prob of N-in:        0.00001
M.      TMHMM2.0        outside      1   1273
```

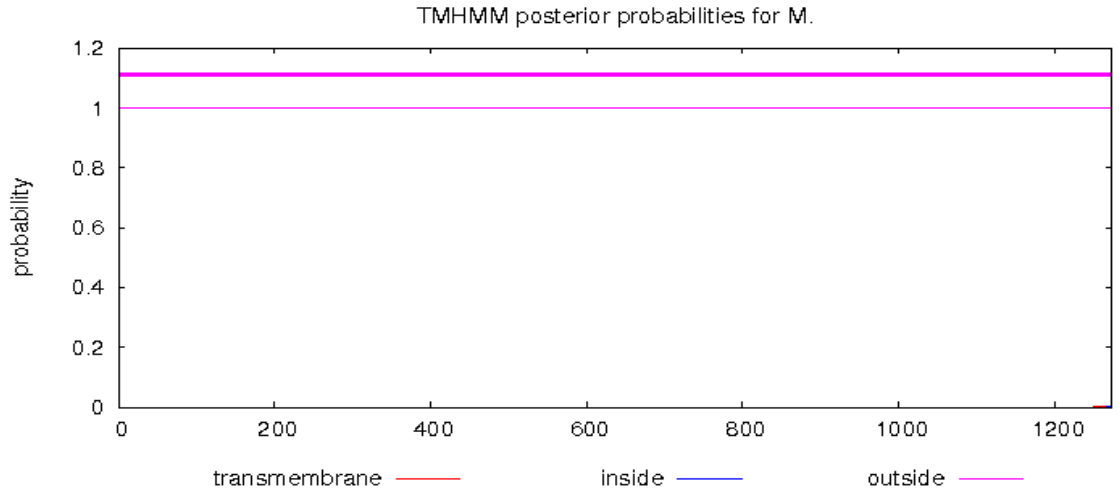


Figure 15B

```
# E. Length: 928
# E. Number of predicted TMHs:  0
# E. Exp number of AAs in TMHs: 0.17868
# E. Exp number, first 60 AAs:  0.04058
# E. Total prob of N-in:        0.00192
E.      TMHMM2.0        outside      1   928
```

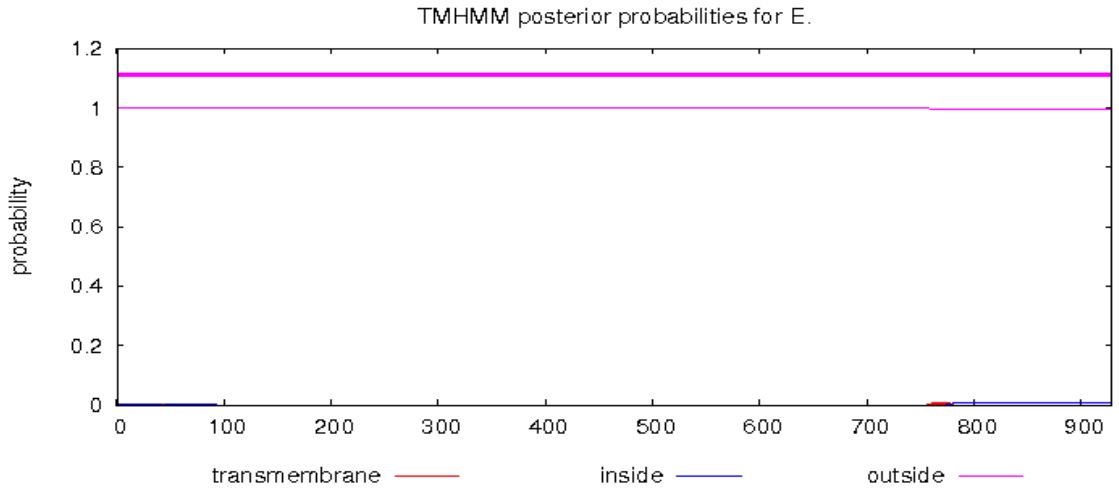


Figure 15. *Mrub_2285* and *E. coli b3863* do not have TMH regions. Figure 15A shows the TMHMM for *Mrub_2285* and Figure 15B shows the TMHMM for *E. coli b3863*. Due to the TMHMM data from TMHMM Server v 2.0, it is predicted that that both proteins have a cytoplasmic location (Krogh & Rapack, 2016).

The figures below, 16A and 16B, are SignalP plots for *Mrub_2285* and *E. coli b3863* (Petersen *et al*. 2011). SignalP plot determines whether there is a protein cleavage site indicative of a protein being either attached or passes through the cell membrane. The D-value determines this and is calculated via S-score (green line) and Y-score (blue line), and a cutoff value (pink line). Both of the D-values were found to be lower than that of the cutoff value. *Mrub_2285* has a cutoff of 0.570 and a value of 0.097, while *E. coli b3863* has a cutoff of 0.570 and a value of 0.159. Thus, these proteins do not have any cleavage sites, again confirming their cytoplasmic location.

Figure 16A

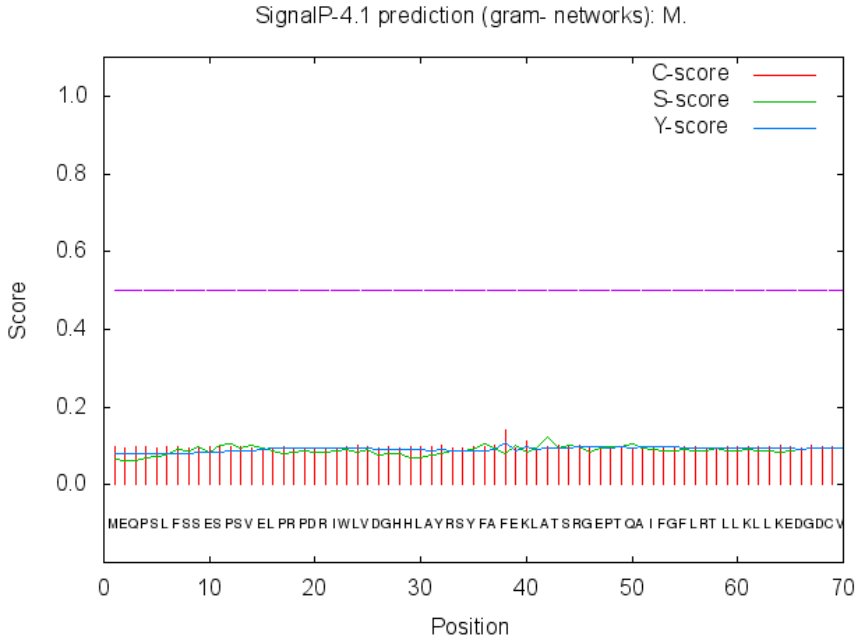


Figure 16B

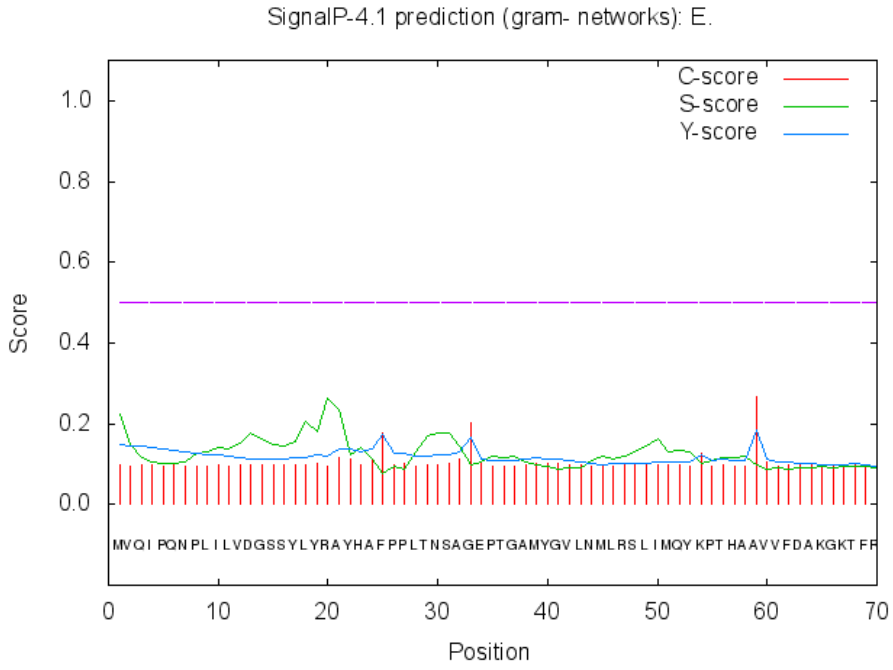


Figure 16. Lack of cleavage sites within *Mrub_2285* and *E. coli b3863*. Figure 16A is *Mrub_2285* and Figure 16B is *E. coli b3863*. The cutoff values were both below the D values and a result of NO to signal peptides. These figures were created via Signal P server 4.1 (Petersen *et al.*, 2011).

Figure 17 is the homologous recombination KEGG RecFOR pathway showing the *M. ruber* (Figure 17A) and *E. coli* (Figure 17B) sides (Kinehisa *et al.*, 2016b). The green color is indicative of enzymes/proteins that are predicted to be present within each organism's genome. The *M. ruber* pathway shows *Mrub_2285* is not present within the RecFOR pathway as the DpoI protein is shown in white. However, the preliminary BLAST between *T. thermophilus* and *M. ruber* indicated that *M. ruber* does have the *Mrub_2285* protein. Therefore, the DpoI protein is found within the *M. ruber* RecFOR pathway. *E. coli* contains the gene *b3863*, the DpoI protein, which is DNA polymerase I.

Figure 17A

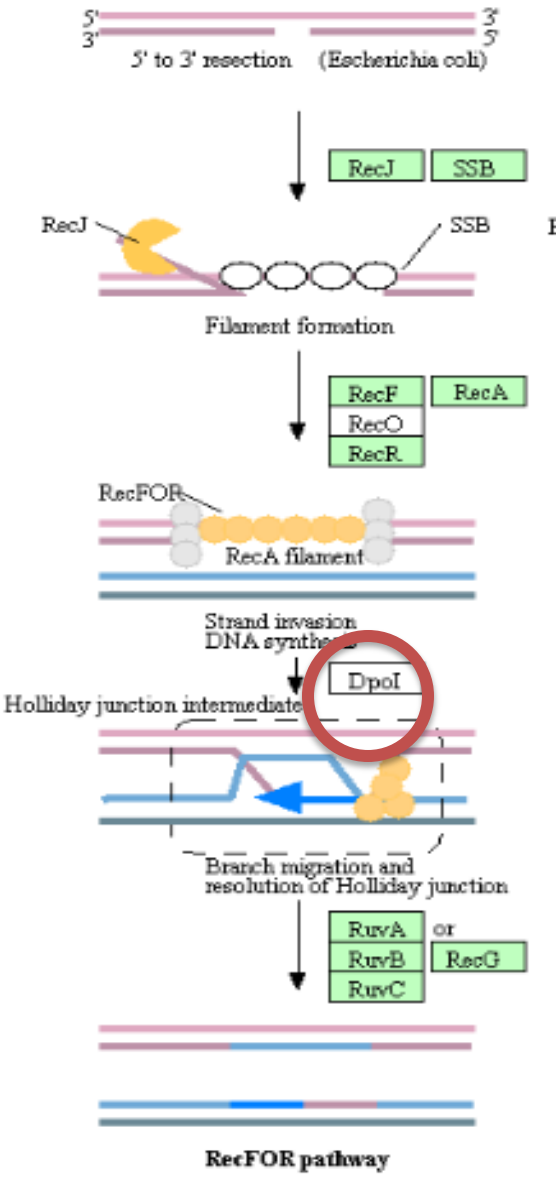


Figure 17B

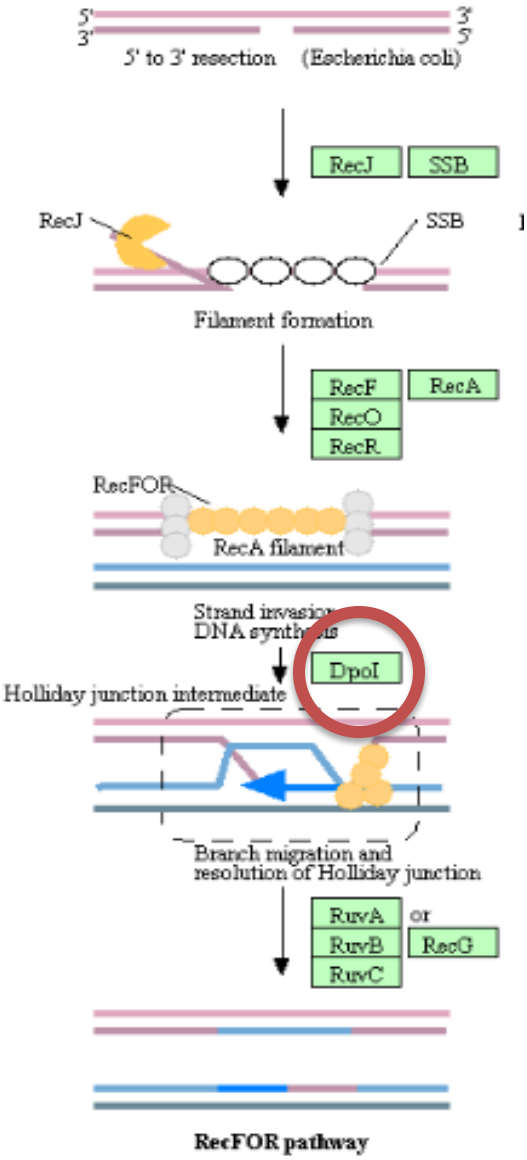
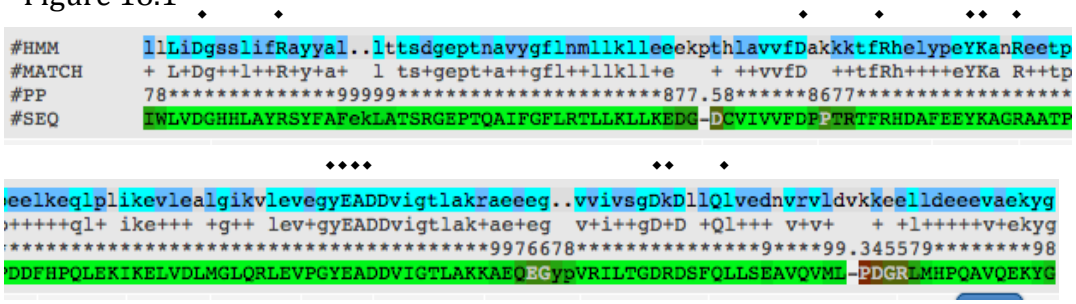


Figure 17. *Mrub_2285* and *E. coli b3863* are in the same recombination pathway - RecFOR pathway. Figure 17A shows the KEGG pathway for *Meiothermus ruber* and Figure 17B shows the KEGG pathway *Escherichia coli*. The KEGG database – The Kyoto Encyclopedia of Genes and Genomes – was utilized to locate these genes within the homologous recombination RecFOR pathway (Kanehisa *et al.*, 2016b).

Figure 18 shows the pairwise alignments between the Pfam consensus sequence and *Mrub_2285* or *E. coli b3863*. Figure 18.1 is the alignment for the hit PF02739 – the 5'-3' exonuclease, N-terminal resolvase-like domain. The second hit is shown in Figure 18.2 for the hit PF01367 – the 5'-3' exonuclease, C-terminal SAM fold. Figure 18.1A and 18.2A are the pairwise alignments for *Mrub_2285* and Figure 18.1B and 18.2B are the pairwise alignments are *E. coli b3863*. From this comparison between the genes, it is relevant that they have many of the same conserved amino acids (as seen in the first and second rows with capital letters). Therefore, this is another piece of evidence that *Mrub_2285* and *E. coli b3863* are orthologous (Berman *et al*., 2000).

Figure 18.1

A



B

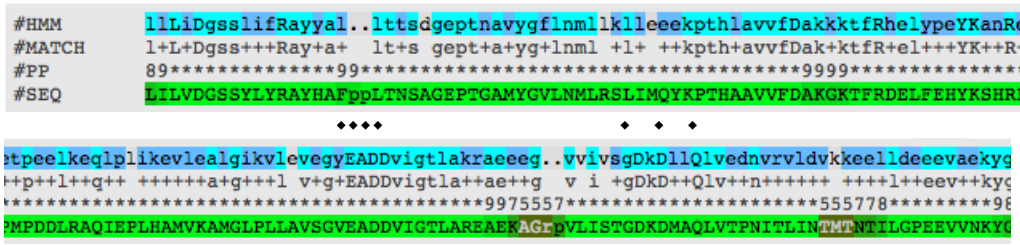
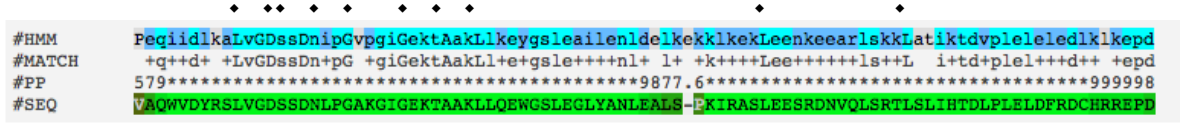


Figure 18.2

A



B

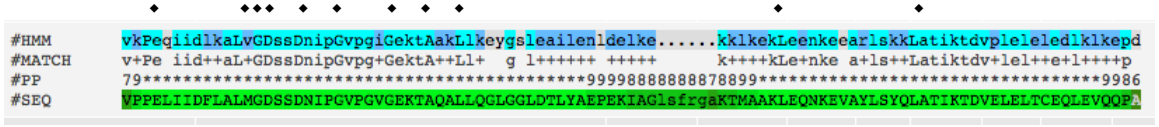
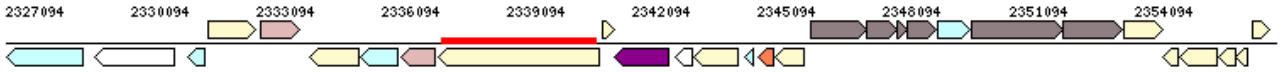


Figure 18. *Mrub_2285* and *E. coli b3863* have the same highly conserved amino acids in two different protein domains. Figure 18.1 is the comparison for the PF02739 - 5'-3' exonuclease, N-terminal resolvase-like domain and Figure 18.2 is the comparison for the PF01367 - 5'-3' exonuclease, C-terminal SAM fold. These pairwise alignments were constructed via Pfam (Berman *et al.*, 2000).

Figure 19 is a map of the chromosome that flanks our gene of interest, colored by their predicted KEGG pathway. The different colors in the figures are indicative of these genes not being a part of an operon. *Mrub_2285* is cream in color with no other cream colored genes around it. *E. coli b3863* is pink in color with no other pink genes near it. The red line above or below the gene is indicative of the gene of interest. Therefore, the Color by KEGG results in more evidence that these two genes, *Mrub_2285* and *E. coli b3863* are orthologous (Kanehisa *et al*., 2016a).

19.1



19.2

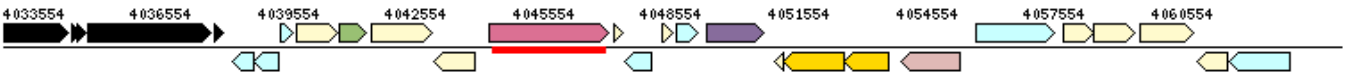


Figure 19. *Mrub_2285* and *E. coli b3863* genes are not a part of an operon. 19.1 is *Mrub_2885* and 19.2 is *E. coli b3863* Colored by KEGG and Ortholog Neighborhoods were used in order to view these images (Kanehisa *et al.*, 2016a).

CONCLUSIONS

The results from this research have shown that *Mrub*_0860 and *E. coli b2892*, *Mrub*_0701 and *E. coli b2562*, and *Mrub*_2285 and *E. coli* b3863 are orthologous. This is indicative of the genes relating to species that come from a common ancestor. These results have stemmed from first, a BLAST analysis between the *E. coli* sequences and *M. ruber* sequences. Other bioinformatics tools were used including: cellular localization of the proteins, TMH, SignalP, LipoP, PSORT-B showing the same location of all these proteins, within the cytoplasm. Also utilized were TIGRFAM and Pfam to determine a

match of what each protein was, that consisted of single-stranded-DNA specific exonuclease (RecJ protein), DNA repair protein (RecO), and DNA polymerase I (DpoI) and in what protein domains they are apart of. Based on these and more bioinformatics tools that were used throughout this project, there were many similarities between the different genes and little to no deviations. It is concluded through the multiple bioinformatics tools that *Mrub_0860* is orthologous to *E. coli b2892, Mrub_0701* is orthologous to *E. coli b2562* and *Mrub_2285* is orthologous to *E. coli b3863*.

Site-directed mutagenesis (SDM) is an *in vitro* procedure performed to create a desired mutation in a double-stranded DNA plasmid (NEB, 2017). SDM may be utilized for many reasons including: studying changes in protein activity, screening for mutations with desired properties and to introduce or remove restriction endonuclease sites. Figure 20.3 shows an alanine mutation to *Mrub_0860.* This mutation is a missense mutation by changing one amino acid to another. The substitution is alanine for a histidine, creating a change of CAC to GCC at positions 466 and 468.

Replacing the histidine with an alanine will affect the protein function of the enzyme due to the fact that histidine substituted with another amino acid does not work well (Betts & Russell, 2003). Histidine is charged, polar amino acid that common in active or binding sites. Replacing this with an alanine, which is a dull, non-reactive, nonpolar amino acid and usually substituted for small amino acids, can lead to function inhibition. *Mrub_0860*, single stranded DNA-specific exonuclease (RecJ), may not be able to function properly, leading to DNA repairs via gap-filling to not take place if RecJ can't make the gap wider in order for the single-stranded binding protein and other

proteins to bind. This results in damaging affects on the RecFOR homologous recombination pathway.

Figure 20.1

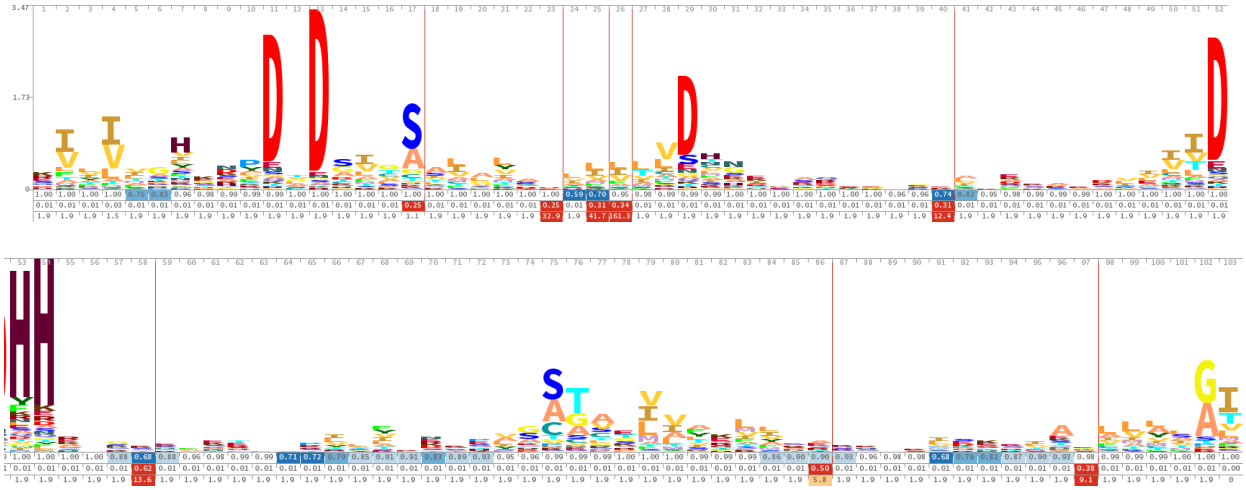


Figure 20.2

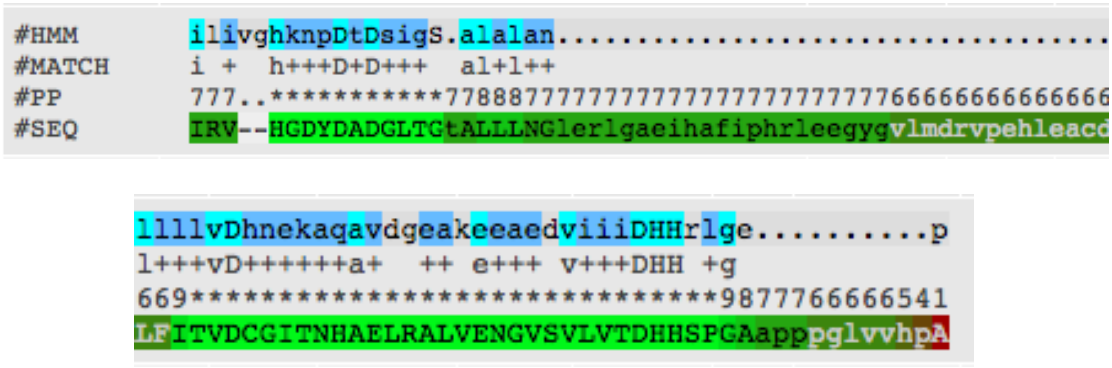


Figure 20.3

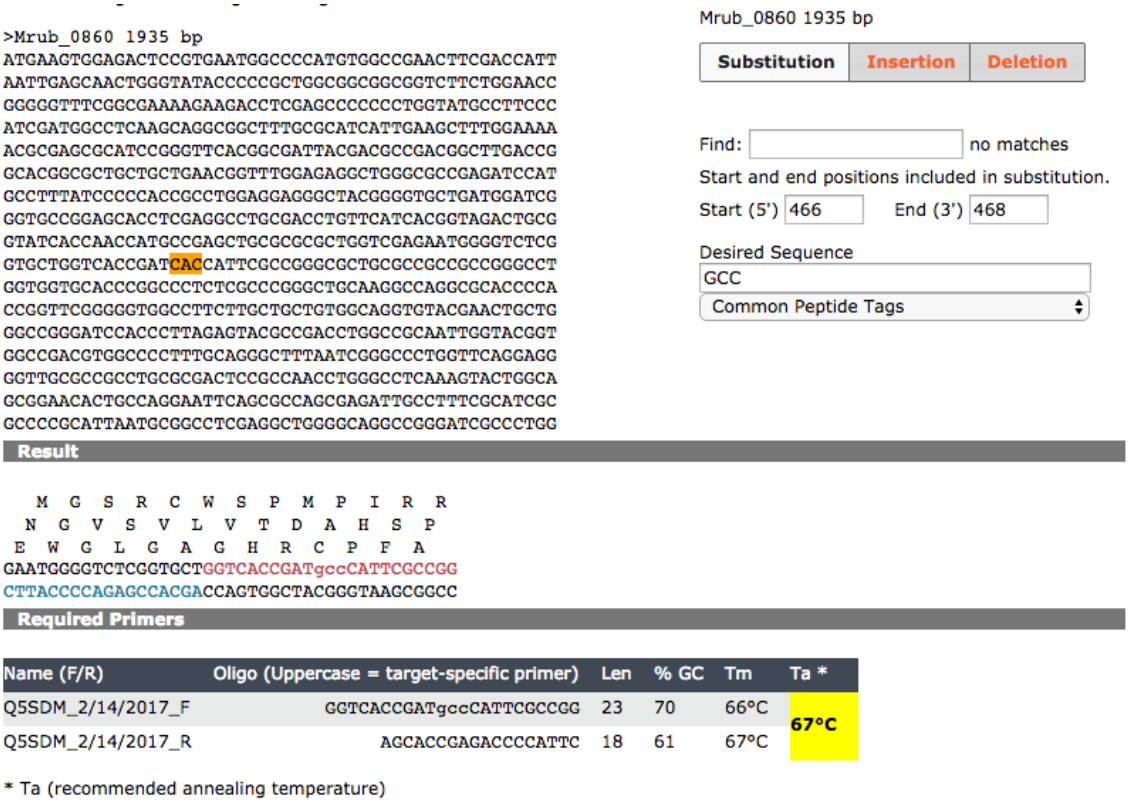


Figure 20. Missense mutation of *Mrub_0860* by substitution of an alanine for a histidine. The missense mutation was created at positions 466 and 468 by changing CAC to GCC. Figure 20.1 shows the HMM logo for *Mrub_0860* gene, this shows the most conserved amino acids in the sequence. The most conserved is indicated by the larger the letter is and the least conserved is the smaller the letter is. Histidine is one of the largest letters in the HMM logo, the histidine that was changed is the first histidine of the second panel of Figure 20.1. Figure 20.2 is a confirmation that this histidine is highly conserved in M. ruber and many other organism sequences found through Pfam. Figure 20.3 shows the missense mutation with the primers that would be needed to make this DNA site-directed mutation in the laboratory. The missense SDM was created via NEB, http://nebasechanger.neb.com/.

Works Cited

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., Stockinger, H. ExPASy: SIB bioinformatics resource portals, Nucleic Acids Res, 40(W1):W597-W603, 2012. Available from: http://enzyme.expasy.org/enzyme-search-ec.html

A. S. Juncker, H. Willenbrock, G. von Heijne, H. Nielsen, S. Brunak and A. Krogh. Protein Sci. 12(8):1652-62, 2003; [2016 Dec 6]. Available at: http://www.cbs.dtu.dk/services/LipoP/

Bayat, A. (2002) Bioinformatics. 324:1018-1022. [2017, Feb. 6]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1122955/

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E.. [Internet]. 2000. The Protein Data Bank. [2016 Dec 6]. Available from: http://www.rcsb.org/.

Betts M.J., Russell R.B. 2003. Amino acid properties and consequences of substitutions. Amino Acid Properties and Consequences of Substitutions. Bioinformatics for Geneticists. John Wiley & Sons 14:290-314 [2017 Feb 12]. Available from: http://moodle.augustana.edu/pluginfile.php/294318/mod_resource/content/1/bettsRussell2003.pdf

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, 2004; [2016 Dec 6]. Available at: http://weblogo.berkeley.edu/

Debouk C, Metcalf B. The impact of genomics on drug discovery. Annu Rev Pharmacol Toxicol.2000;40:193–208. [PubMed]

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A. 2016. The Pfam protein families database: towards a more sustainable future: Nucleic Acids Res., 44:D279-D285; [2016, Dec. 6]. Available from: http://pfam.xfam.org/

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29(1):41-3.

JGI - Phylogenetic Diversity. U.S. Department of Energy Joint Genome Institute. [2017, Feb 2]. Available from: http://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/

Kanehisa M, Sato Y, Kawashima M, Furumichi M. and Tanabe M. (2016a) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res., 44, D457–D462; [2016 Dec 6]. Available from: http://www.genome.jp/kegg/

Kanehisa M, Sato Y, Kawashima M, Furumichi M. and Tanabe M. (2016b) KEGG Pathway: Homologous recombination. [2016 Dec 6]. Available from: http://www.genome.jp/kegg/-bin/show_pathway?map=ko03440&show_description=show

Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martinez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Schroder, I., Shearer, A., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R.P., Paulsen, I., and Karp, P.D. 2013. EcoCyc: fusing model organism databases with systems biology Nucleic Acids Research 41:D605-612.

Krogh A, Rapacki K. TMHMM Server, v. 2.0. Cbs.dtu.dk. 2016 [accessed 2016 Dec 6]. http://www.cbs.dtu.dk/services/TMHMM/

Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097/

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. CDD: NCBI's conserved domain database. Nucleic Acids Res.28(43): D222-2: [2016 Dec 6]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25414356?dopt=AbstractPlus

Markowitz VM, Chen IA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, *et al*. 2012. IMG: The integrated microbial genomes database and comparative analysis system. Nucleic Acids Research 40(D1):D115-22. Available from: http://nar.oxfordjournals.org/content/40/D1/D115.full

New England BioLabs (NEB). 2017. Applications: Site Directed Mutagenesis. [2017 Feb 12]. Available from: https://www.neb.com/applications/cloning-and-synthetic-biology/site-directed-mutagenesis

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302 (1):205-17 Available from: http://www.ebi.ac.uk/Tools/msa/tcoffee/

Persky, N.S., Lovett, S.T. Mechanism of Recombination: Lessons from *E. coli*. Critical Reviews in Biochemistry and Molecular Biolog, 43:347-370, 2008. Available from: http://spetses.med.harvard.edu/pages/Micro201/Lovett_08_review.pdf

Petersen T.N., Brunak S., von Heijne, G., Nielsen, H.Discriminating signal peptides from transmembrane regions. Nature Methods, 8:785-786, 2011. Available from: http://www.cbs.dtu.dk/services/SignalP

Scott, L. 2016a. Meiothermus ruber Genome Analysis Project. GENI-ACT. [2017. Feb 10]. Available from: http://www.geni-act.org/

Scott, L. 2016b. Meiothermus ruber Genome Analysis Project. GENI-SCIENCE. [2017, Feb. 10]. Available from: http://www.geni-science.org/secure/projects/view/

Tindall *et al*. 2010. Complete genome sequcne of Meiothermus ruber type strain. Stand Genomic Sci 3(1): 26-36.

Yu N.Y., Wagner J.R., Laird M.R., Melli G., Rey S., Lo R., Dao P., Sahinalp S.C., Ester M., Foster L.J., Brinkman F.S.L. (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics **26(13):1608-1615**