2019

An investigation into the relationship between mrub_3013, mrub_1477, and mrub_0224: Are they paralogs?

Melette DeVore
Augustana College, Rock Island Illinois

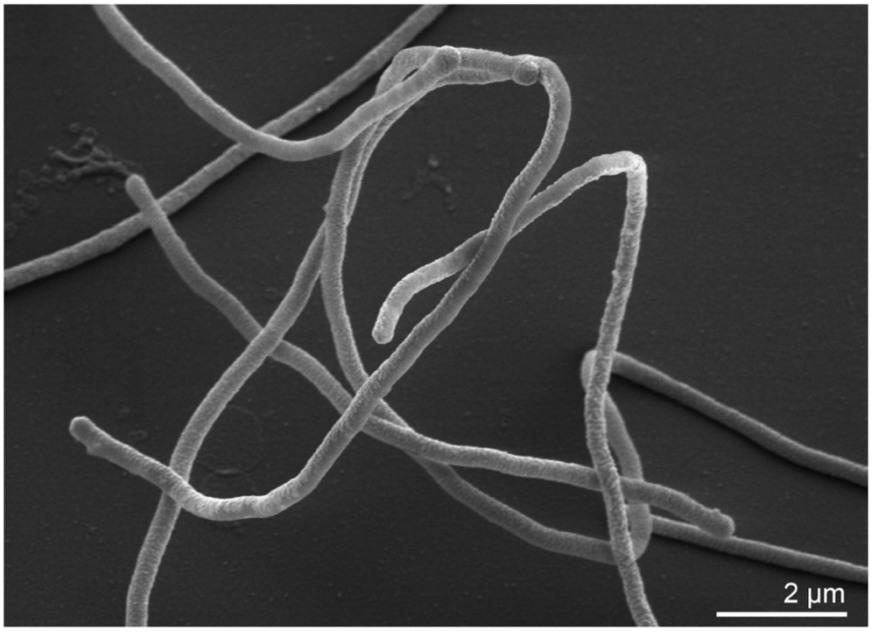
Dr. Lori Scott
Augustana College, Rock Island Illinois

Melette DeVore
Dr. Lori R. Scott Laboratory
Biology Department, Augustana College
639 38th Street, Rock Island, IL 61201

An investigation into the relationship between mrub_3013, mrub_1477, and mrub_0224: Are they paralogs?

INTRODUCTION

Thermophilic organisms live in hot environments that are inhospitable to many species, little is known about how thermophilic bacteria withstand such conditions. Organisms that live in extreme environments are difficult to grow in lab and their natural conditions make research difficult as well (Brininger *et al.* 2018). One goal of the *Meiothermus ruber* genome analysis project is to understand how thermophilic bacteria, such as the microbe *M. ruber*, survive in hot conditions. The name *Meiothermus ruber* comes from "meio," meaning less, "thermus," meaning hot, and "ruber," meaning red. As a whole it means that *M. ruber* is an organism that lives in a less hot environment and produces a red pigment. *M ruber* is typically found in natural hot springs and artificial thermal environments, it can grow in temperatures ranging from 35-70°C, and its optimum growth temperature is 60°C (Tindall *et al.* 2010). *M. ruber* must live in an aerobic environment and is a Gram-negative, rod-shaped bacteria, Figure 1 shows an electron-scanning microscope picture of *M. ruber*. *M. ruber* is an organism of interest because it lives in hot environments and because its genome has been sequenced as part of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) Project (Tindall *et al*. 2010). Previous research has found that the *M. ruber* gene for ProC is orthologous to the *E. coli* gene for ProC, suggesting that there may be other similarities between their genomes (Scott 2018).



Toward the goal of studying how *M. ruber* has adapted to higher temperatures, Dr. Scott is studying proline biosynthesis, with an emphasis on the ProC enzyme, the last enzyme in the biosynthetic pathway of proline. Proline is thought to play a role in stress-management in organisms in harsh environments and understanding its biosynthesis may help in understanding the stress-management of other thermophilic organisms (Scott 2018). The *M. ruber* genome analysis project uses *Escherichia coli* as a model organism due to the well-studied nature of its metabolic pathways and the abundance of data available on the organism. By inserting the *M. ruber proC* gene into *E. coli*, the goal of the project was to show the orthologous nature of the *M. ruber* and *E. coli proC* genes.

Another goal of the *M. ruber* genome analysis project is to predict how *M. ruber* performs its many bioloigcal processes and synthesizes its many cellular components. In this paper, we present evidence that *M. ruber* has the CRISPR-Cas system. CRISPR stands for Clustered Regularly Interspaced Palindromic Repeats, Cas is CRISPR-associated proteins. It is a bacterial defense mechanism against bacteriophages and plasmid invasion that is similar to adaptive immunity in mammals and is found in about 50% of bacteria and 90% of archaea (Wright, Nunez, and Doudna 2016). The CRISPR array component of the CRISPR-Cas system includes a leader sequence followed by repeat sequences separated by spacers that are derived from foreign DNA acquired in previous infections. There are three stages of the CRISPR-Cas defense system: spacer acquisition, CRISPR RNA (crRNA) synthesis, and interference (Jiang and Doudna 2016; Wright *et al*. 2016; Darmon and Leach 2014). Figure 2 Panel A shows a visual representation of these steps. Spacer acquisition involves the identification of foreign DNA and processing it to be inserted in the CRISPR array. New spacers are generally inserted after the leader sequence and a repeat sequence is copied with each spacer acquisition to separate individual spacers. Synthesis of crRNA is the transcription of the CRISPR array and subsequent RNA processing. Mature crRNA consists of one spacer sequence and part or all of repeat sequences on either side of the spacer. Marture crRNA associates with CRISPR-effector complexes, which are composed of Cas proteins, and guides it to foreign DNA. Foreign DNA that is complementary to the crRNA is destroyed, completing CRISPR-Cas defense (Jiang and Doudna 2016; Wright *et al*. 2016; Darmon and Leach 2014).

There are six types of CRISPR-Cas system, types I - III are the best studied mechanisms while types IV - VI have just recently been discovered. Figure 2 Panel B shows each type and their signature protein or effector complex that carries out the actual degradation of foreign DNA. Type I is distinguished by its Cas3 protein, Type II its Cas9 protein, Type III its Cas10 protein. The hallmark of Type IV is Csf1, of Type V is a Cas9-like protein, and of Type VI is C2c2. Types I, III, and IV are considered Class 1 CRISPR-Cas mechanisms as their hallmark effector complex has multiple subunits, the other types are Class 2 because they have a single hallmark protein with multiple domains (Wright *et al*. 2016).

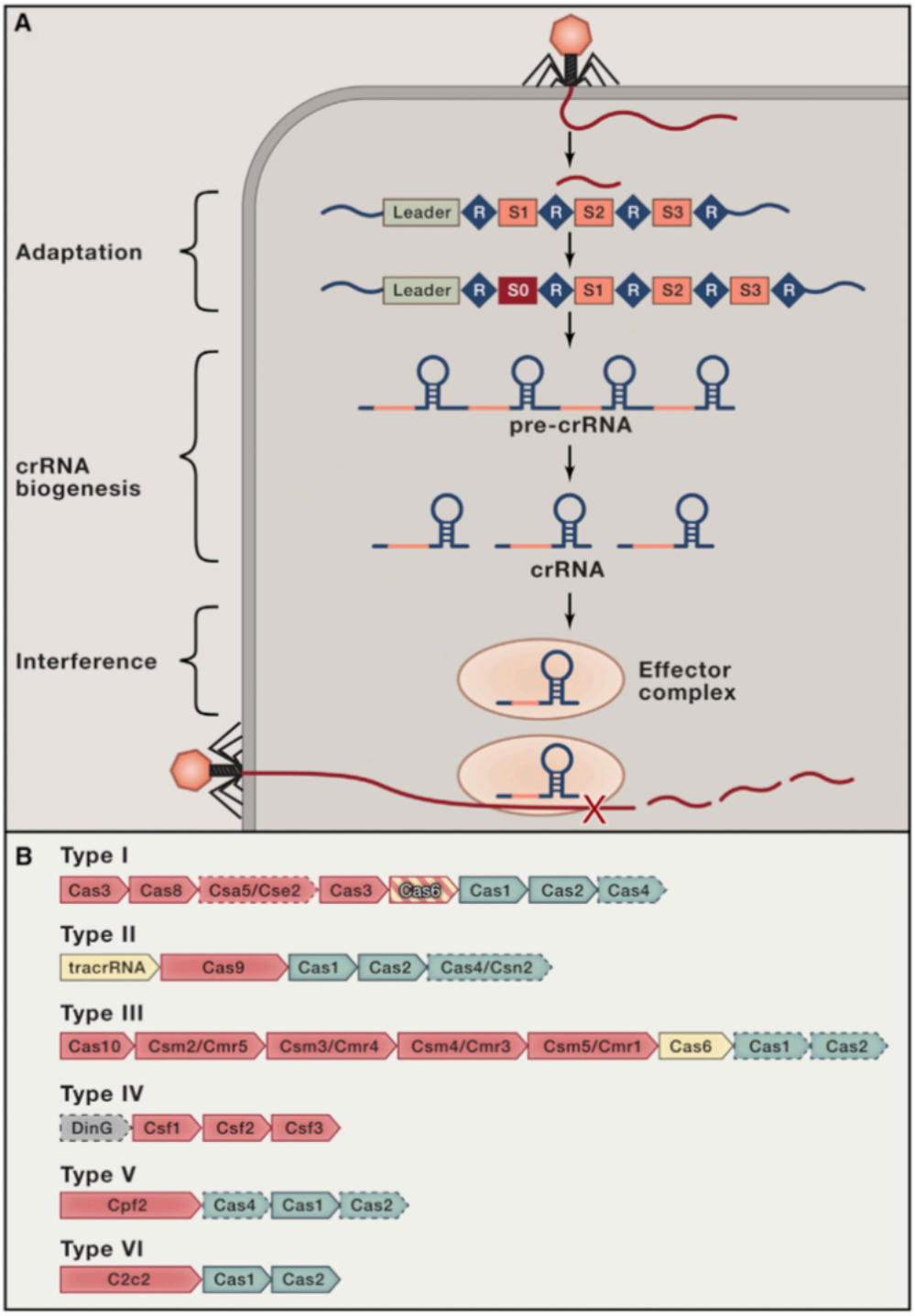


Figure 2. Overview of the CRISPR-Cas system and illustration of the genetic differences between each type of CRISPR-Cas system. Panel A shows the three steps of the CRISPR-Cas defense system: acquisition of spacers, crRNA synthesis, and interference and degradation of foreign DNA. Diamonds with R signify repeat seqeunces. Rectangles with S# indicate spacers, R0 is the most recently added spacer. Panel B shows the various genes that are hallmarks of each type of CRISPR-Cas defense system. Taken from Wright *et al*. (2016)

Type I CRISPR-Cas system are further divided into subtypes A-F. *E. coli* K12 has a Type I-E CRISPR-Cas system which has been well-studied. Its CRISPR array has eight genes for Cas proteins, Cas1 and Cas2 are involved in spacer acquisition, Cas3 is involved in the interference and degradation of foreign DNA. The other proteins, CasA (Cse1), CasB (Cse2), CasC (Cas7), CasD (Cas5e), and CasE (Cas6e) associate to form a Cascade complex that works with crRNA to find and initiate the destruction of invading DNA (Jiang and Doudna 2016). The CRISPR-Cas system of *M. ruber* shows potentially remarkable similarities to that of *E. coli* K12, with a Type 1-E system. It also has genes for a Type II system and some that resemble a Type III system.

The focus of this research is Cas1, which is essential for spacer acquisition. Two Cas1 dimers associate with and effectively "sandwich" a single Cas2 dimer, forming the Cas1-Cas2 complex (Nunez *et al.*2014). How exactly this complex carries out spacer acquisition still requires more research; however, studies have shown that Cas1 is more essential than Cas2. When mutations are induced in the Cas2 active site, there is little to no change in spacer acquisition. On the other hand, when mutations are induced in the Cas1 active site spacer acquisition is effectively shut down almost completely (Nunez *et al.*2014).. Cas1 and Cas2 are universal CRISPR-Cas proteins, and are found in each type of CRISPR-Cas system. What is most interesting about the CRISPR-Cas system of *M. ruber* is that it has three separate genes for Cas1 and Cas2. This begs the question of a paralogous relationship and what, if anything, is different between the three versions.

Paralogs are related genes that arose through gene duplication events, resulting in multiple copies of the same gene. According to Bratlie *et al*. (2010), there are three things that can happen when paralogous genes are kept: one duplicate may evolve a new function, the multiple functions of the original gene may divide between paralogs, or both copies may retain the original function. Paralogs allow for evolution in bacterial genomes, and observation of which paralogs are conserved can indicate which functions are under important selection pressure. Gevers *et al*. (2004) and Sanchez-Perez *et al*. (2008) found that the most conserved paralogs are found in the functional domains of metabolism, transcription, and cellular defense mechanisms. CRISPR-Cas is a cellular defense mechanism against foreign and invading DNA, suggesting that the presence of multiple genes for a single protein in the CRISPR-Cas family is significant.

There is some research into the role of paralogs in adapting to changing environments. Sanchez-Perez *et al*. (2008) suggest the existence of "ecoparalogs" that are different copies of the same protein but with varying functionalities in varying environments. In the halophile (salt-loving) *Salinibacter ruber* there are multiple copies of the same transport protein that operate best at varying salinities (Sanchez-Perez *et al*. 2008). Proteins that are found near the cell surface or are involved in DNA binding were found to have greater numbers of paralogs, suggesting that the environment does play a role in the development of ecoparalogs. Sanchez-Perez *et al*. (2008) predicted that other prokaryotes likely to have ecoparalogs would include other halophilic species and thermophilic species. Through analysis of the three copies of the Cas1 gene in *M.*

ruber, I intend to investigate the relationship between each gene and determine if they are true paralogs with at least 30% similarity over 60% of their sequence.

METHODS

In order to learn more about the CRISPR-Cas system in the model organism, *E. coli* K12 MG1655, I used EcoCyc (Kesler *et al.* 2013), an online database dedicated to *E. coli* K12 MG1655. It contains information on the genome, metabolic processes, and more of *E. coli* K12 MG1655. I specifically focused on the Cas1 protein and *cas1* gene and collected data regarding its structure and function. I then used the KEGG database (Kanehisa *et al.* 2019) and the IMG/M database (Markowitz *et al.* 2012) to collect information on whether CRISPR-Cas systems are present in *M. ruber* and how they are structured. I compared the CRISPR-Cas systems in *E. coli* and in *M. ruber* and chose the *M. ruber* Cas1 genes mrub_3013, mrub_1477, and mrub_0224 for this project.

The IMG/M database and NCBI Blast Multiple Sequence Alignment tool (Madden 2002) were used to confirm the start codon of each *M. ruber* gene.The NCBI Protein BLAST tool was used to compare each *M. ruber* protein to *E. coli* b2755 and produce pairwise alignments of the amino acid sequences. To predict the cellular localization and protein structure of each *M. ruber* protein the bioinformatics tool TMHMM was used to predict the presence of alpha-helices and the bioinformatics tool PRED (Bagos *et al.* 2004) was used to predict the presence of membrane-embedded beta-barrels. PSort-B (Yu *et al.* 2010) was also used to predict the cellular localization of each *M. ruber* protein.

Structural data on each protein was collected using NCBI Protein BLAST and the TIGRFAM (Haft *et al.* 2001), PFAM (Finn *et al.* 2016), and PDB databases (Berman *et al.* 2000). The NCBI Protein Blast tool was used to identify conserved domains in each protein. TIGRFAM, PFAM, and PDB were used to find proteins with similar sequences and domains to the *M. ruber* protein. Using the IMG/M database, the possibility of each gene being in an operon was analyzed. Finally, the website phylogeny.fr was used to evaluate the evolutionary relationships between each *M. ruber* gene. All of these tools were used to determine if mrub_3013, mrub_1477, and mrub_0224 are paralogous genes.

RESULTS

Initial research into the b2755 *cas1* gene in *E. coli* found that *cas1* is part of a CRISPR-Cas Type I-E operon. Cas1 is localized to the cytoplasm and is 305 amino acids long, *cas1* is 918 base pairs long. As part of the operon, it is preceded by *casE* and followed by *cas2*, all proteins are involved in the CRISPR-Cas defense system. Figure 3 shows the *E. coli* K12 MG1655 CRISPR-Cas operon. There are three possible promoters leading to three transcription units, *cas1* is included in two of the three transcription units, though one is unconfirmed. In *E. coli* there is a single *cas1* gene, b2755. In *M. ruber* there are three genes for Cas1, mrub_3013, mrub_1477, and mrub_0224.

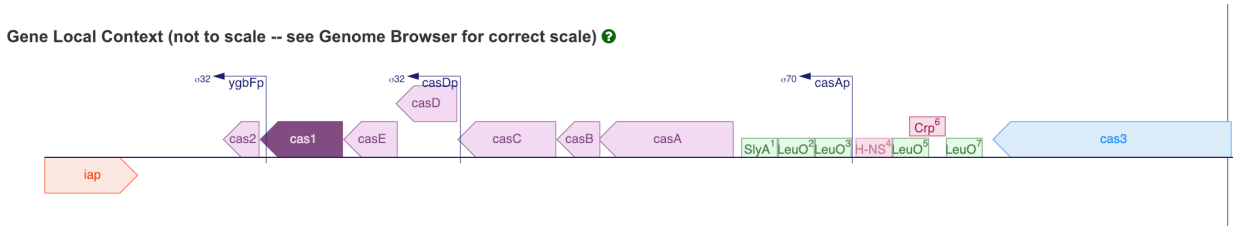


Figure 3. CRISPR-Cas Type I-E operon found in *Escherichia coli* K12 MG1655. The gene of interest is b2755, which codes for Cas1, a CRISPR-associated endonuclease. The *cas1* gene is colored dark purple, the other genes in the operon are a light purple, and the arrows indicate transcription promoters. The green boxes are activators and the red boxes are inhibitors of transcription. Taken from EcoCyc https://ecocyc.org/gene?orgid=ECOLI&id=G7425.

All three *M. ruber* genes are categorized as the CRISPR-associated Cas1 protein. Figure 4 shows the KEGG output for the *M. ruber* CRISPR-Cas system. The map location of mrub_3013 is 3053978-3054940 bp and its protein is 320 amino acids long. The mrub_1477 gene is located at 1504008-1505027 bp and is 339 amino acids long. Finally, the mrub_0224 gene is at 197591-198562 bp and is a 323 amino acid long protein. Each *M. ruber* protein sequence was compared with the *E. coli* Cas1 amino acid sequence using the NCBI Protein Blast tool. Figure 5 shows the pairwise alignments of each comparison. Table 1 contains the E-values and bit scores for each alignment. Mrub_3013 had the best alignment scores with a 40% identity and an E-value of 2e-75, 114 of the 284 aligned amino acids were the same or chemically similar. The next highest percent identity score was for mrub_0224, with an identity score of 34% and an E-value of 5e-10. For mrub_0224, 39 of the 116 aligned amino acids were the same or chemically similar. Finally, mrub_1477 had a percent identity score of 29% and an E-value of 3e-7, 26 of the 89 aligned amino acids were the same or chemically similar.

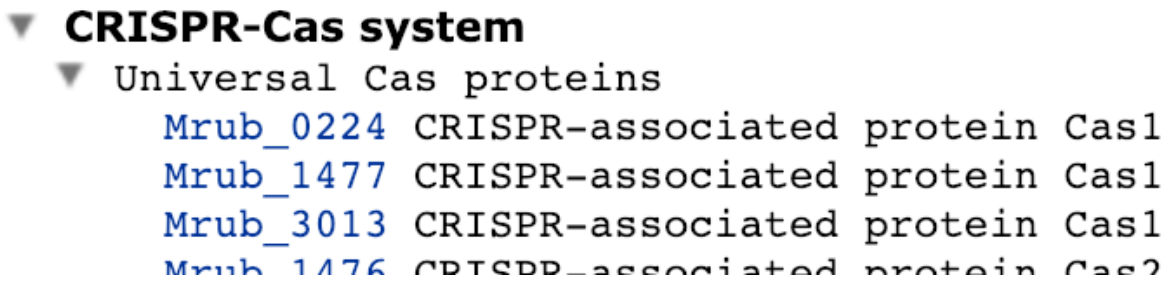


Figure 4. KEGG output for the *M. ruber* CRISPR-Cas system. Mrub_3013, mrub_1477, and mrub_0224 are all identified as CRISPR-associated Cas1 proteins. Taken from KEGG database https://www.kegg.jp/kegg-bin/get_htext.

A: mrub_3013 vs b2755

Range 1: 8 to 290 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
222 bits(565)	2e-75	Compositional matrix adjust.	114/284(40%)	170/284(59%)	2/284(0%)

```
Query  6    LNPIP-LKDRVSMIFLQYGQIDVIDGAFVLIDKTGIRTHIPVGSVACIMLEPGTRVSHAA  64
            L  +P  +D +S ++L++G+++  D A     + G+   IP  ++  +ML PGT ++HAA
Sbjct  8    LQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGV-VAIPAAALGVLMLGPGTSITHAA  66

Query  65   VRLAAQVGTLLVWVGEAGVRVYASGQPGGARSDKLLYQAKLALDEDLRLKVVRKMFELRF  124
            +R  A  G  + WVGE  VR YASG      S  L+ Q +   D +  L+VV++++ LRF
Sbjct  67   IRQLANNGCSVFWVGEEMVRFYASGMGETRSSANLMRQVRAWADPEAHLEVVKRLYRLRF  126

Query  125  GEPAPARRSVEQLRGIEGSRVRATYALLAKQYGVTWNGRRYDPKDWEKGDTINQCISAAT  184
             EP     S+EQ+RG+EG RVR TYA  +++ GV W GR Y   +W   D IN+ ISA
Sbjct  127  PEPLSPELSLEQIRGLEGVRVRETYARWSRETGVEWKGRNYQRGNWAAADPINRAISAGA  186

Query  185  SCLYGVTEAAILAAGYAPAIGFVHTGKPLSFVYDIADIIKFDTVVPKAFEIARRNPGEPD  244
            +CLYG+  AAIL+AGY+PA+GF+HTGK LSFVYD+ADI K +T++P AF +   +    +
Sbjct  187  ACLYGLAHAAILSAGYSPALGFIHTGKQLSFVYDVADIYKAETLIPTAFRVVAESDVGVE  246

Query  245  REVRLACRDIFRSSKTLAKLIPLIEDVLAAGEIQPPAPPEDAQP  288
            R VR   R+  +  K L +++  +  +  A E   P   + A P
Sbjct  247  RRVRHTLREQLKEVKLLERIVSDLHSLFDALETPDPYAADPAAP  290
```

B: mrub_1477 vs b2755

Range 1: 152 to 234 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
37.0 bits(84)	3e-07	Compositional matrix adjust.	26/89(29%)	43/89(48%)	11/89(12%)

```
Query  129  PARRSVEQLRGIEGSRVRATYA-----LLAKQYGVTWNGRRYDPKDWEKGDTINQCISAA  183
            P  RS++++RG+EG    A +A     LL+ ++   ++GR   P      D +N  +S
Sbjct  152  PQARSLDEVRGLEGGAASAYFAAFGDLLLSGEF--RFDGRNKRPPR----DPVNALLSFV  205

Query  184  TSCLYGVTEAAILAAGYAPAIGFVHTGKP  212
             + L     AA+   G  P  GF+H  +P
Sbjct  206  YALLTTQCTAALEGVGLDPQAGFLHALRP  234
```

C: mrub_0224 vs b2755

Range 1: 136 to 245 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
45.4 bits(106)	5e-10	Compositional matrix adjust.	39/116(34%)	56/116(48%)	14/116(12%)

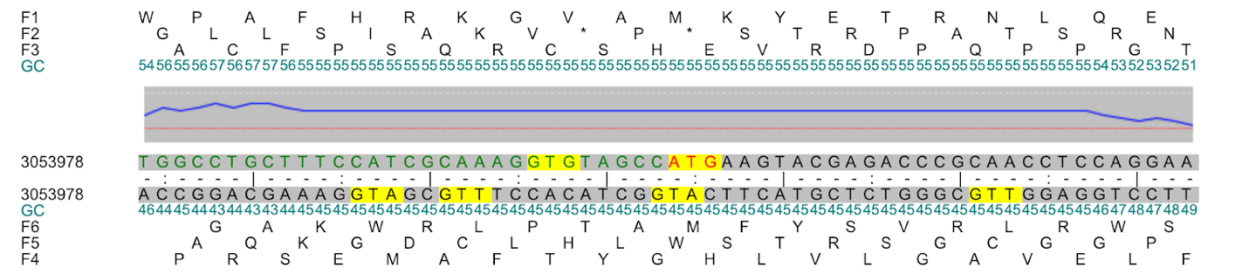
```
Query  126  EPAPARRSVEQLRGIEGSRVRATYALLA---KQYGVTWNGRRYDPKDWEKGDTINQCISA  182
            E  P  RS+E LRGIEG+  RA +A L      YG  ++GR   P      D +N  +S
Sbjct  136  EALPQARSLEALRGIEGNAARAYFAGLQAVLAPYG--FSGRNRRPPT----DAVNAALSY  189

Query  183  ATSCLYGVTEAAILAAGYAPAIGFVHT-GKPL-SFVYDIADIIK---FDTVVPKAF  233
               L G    A+  AG  P +G +HT G+ + +  +D+ +  +    D VV  AF
Sbjct  190  GYMVLLGRVLLALGIAGLHPELGLLHTEGRRVPALAFDLMEEFRVSVVDAVVIAAF  245
```

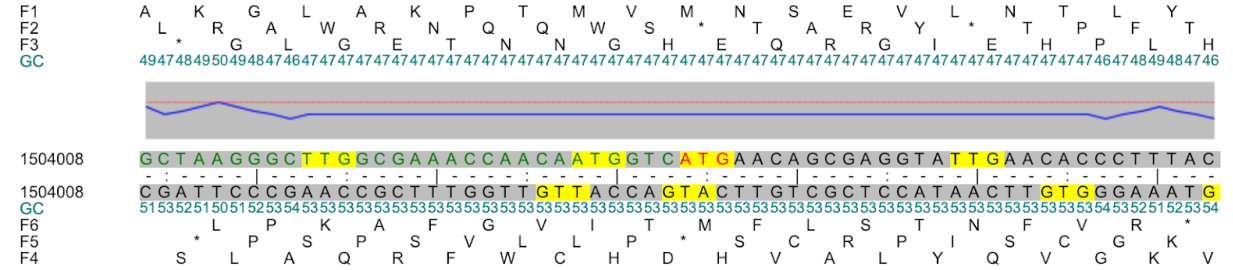
Figure 5. NCBI Protein BLAST alignments of the *M. ruber cas1* genes with *E. coli* b2755. In Panel A the mrub_3013 gene was blasted against b2755 and had an E-value of 2e-75 and an identity score of 40%. In Panel B the mrub_1477 gene was blasted against b2755 and had an E-value of 3e-07 and an identity score of 29%. In Panel C the mrub_0224 gene was blasted against b2755 and had an E-value of 5e-10 and an identity score of 34%.

The start codon for each *M. ruber cas1* was confirmed using both IMG/M and NCBI Blast. Figure 6 shows the IMG/M upstream regions with potential start codons highlighted. The IMG/M tool identified other potential start codons for mrub_3013 and mrub_1477, however with the NCBI data they are not likely to be the actual start codon. There were no other start codons predicted for mrub_0224. Figure 7 shows the NCBI Blast comparison with other, evolutionarily similar organisms. There are no large overhangs in the NCBI Blast Multiple Sequence Alignment for any of the proteins, and there are no other suitable start codons identified in the IMG/M search, so the correct start codons were identified.

A: mrub_3013



B: mrub_1477



C: mrub_0224

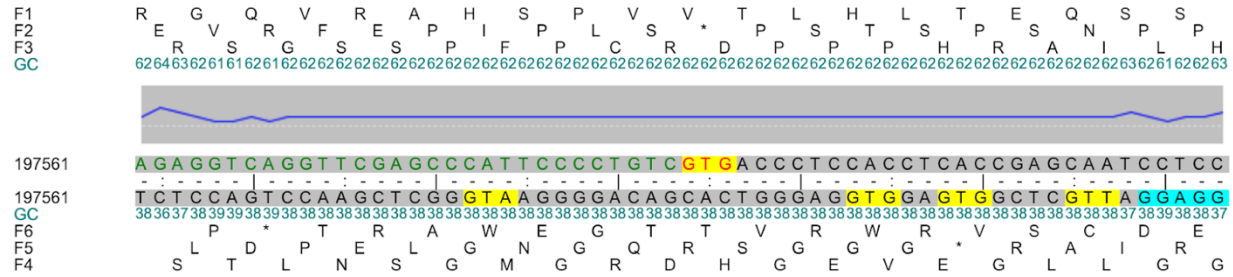


Figure 6. The IMG/M upstream regions and potential start codons. Potential start codons are highlighted in yellow, the start codon typically used for translation is in red font. Panel A shows the upstream region for mrub_3013, there is a potential start codon but it would shift the reading frame. Panel B shows the upstream region for mrub_1477, there are two potential start codons upstream. Panel C shows the upstream region for mrub_0224, there are no other potential start codons.

A: mrub_3013 start codon compared with similar species

```
WP_013015255   1   MK-   --   YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSITHAAIRQLA   71
WP_063843447   1   MK-   --   YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSITHAAIRQLA   71
WP_119361698   1   MR-   --   YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVACYSQEGVVMIPAAALGVLMLGPGTSITHAAIRQLA   71
ADH63133       1   MAE[7]IP[7]YETRNLQELPKFRDGLSYIYLEHGRIEQQDQAVAYYSQEGVVSIPAAALGVLLLGPGTAITHAAIRQLA   88
RIH89453       1   MAE[7]IP[7]YETRNLQELPKFRDGLSYIYLEHGRIEQQDQAVAYYGPDGAVMIPAAALGVLLLGPGTVVTHAAMRQLA   88
WP_119342490   1   MK-   --   YETRNLQELPKFRDGLSYIYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLLLGPGTSITHAAIRQLA   71
WP_119360210   1   MK-   --   YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAFYTQEGVISIPAAALGVLLLGPGTAVTHAAIRQLA   71
WP_018466931   1   MK-   --   YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAFYTQEGVISIPAAALGVLLLGPGTAVTHAAIRQLA   71
WP_051195844   1   MK-   --   YETRNLQELPKFRDGLSYLYLEHGRIEQQDQAVAYYSQDGVVAIPAAALGVLMLGPGTSITHAAIRQLA   71
WP_105317436   1   MPP   VP   -PARNLKELPKFRDGLSYLYVEHAFLEQEAQGIGVYDREGLTLVPVAALGVLFLGPGTRITHAAIRALA   73
WP_053768182   1   MPP   VP   -SARNLKELPKFRDGLSYLYVEHAVVEREAGGIGIYDQEGLTLAPVAGLGVLFLGPGTRITHAAIRLLA   73
WP_018111808   1   MPP   VP   -NTRNLKELPKFRDGLSYLYVEHAFIEQEAQGIGIYTQEGLTLVPVAALGVLFLGPGTRITHAAIRALA   73
WP_015717142   1   MPP   VP   -SARNLKELPKFRDGLTYLYVEHAFIEQEAQGIGIYDQQGLTLVPVAALGVLFLGPGTRITHAAIRTLA   73
WP_114312722   1   MPP   VP   -SARNLKELPKFRDGLSYLYVEHAFIEQEAQGIGIYDQEGLTLVPVAALGVLFLGPGTRITHAAIRALA   73
WP_011229111   1   MPP   VS   -SARNLKELPKFRDGLSYLYVEHAVVEREAGGIGIYDQEGLTLAPVAGLGVLFLGPGTRITHAAVRLLA   73
```

B: mrub_1477 start codon compared with similar species

```
ADD28239       1   MNSEVLNTLYIQTQGVYLRLEGDTLRIEHEDVT-LRNVPLHHLGGLALFGNVLVSPYLLHRCAQDGLEVTWFSESGRFQG   79
AWR86722       1   MNSEVLNTLYIQTQGVYLRLEGDTLRIEHEDVT-LRNVPLHHLGGLALFGNVLVSPYLLHRCAQEGLEVTWFSESGRFQG   79
WP_013159698   1   MTTELLNTLYIQTQGVYLRLESDTLRIQHEDVT-LRHVPLHHLGGLALFGNVLVSPFLLHRCAEDGLEVTWFSESGRFQG   79
WP_119358226   1   MTGELLNTLYVQTQGVYLRLEGDTLRIQHEDVT-LRNVPLHHLGGVAVFGNVLISPFLLHRCAEEGLEVAWFSESGRFQG   79
WP_119339591   1   MTSEILNTLYIQTQGVYLRLEGDTLRIQHENIT-LRNVPMHHLGGVAVFGNVLISPFLLQRCAEEGLEVSWFSESGRFFG   79
WP_018465593   1   MTSELLNTLYIQTQGVYLRLEGDTLRIQHEEVT-LRNVPLHHLGGVAAFGNVLISPFLLHRCAEEGLEVSWFTESGRFQG   79
WP_119359034   1   MTTELLNTLYVQTQGVYLRLEGDTLRIQHEEVT-LRNVPLHHLGGLVMFGNVLISPFLLHRCAEEGLEVAWFTESGRFQG   79
WP_027878459   1   MTSELLNTLYIQTQGVYLRLEGDTLRIQHEDVT-LRNVPLHHLGGLALFGNVLISPFLLARCAEEGLEVSWFSESGRFFG   79
WP_119277187   1   MTSELLNTLYVQTQGVYLRLEGDTLRIQHEDIT-LRNVPLHHLGGLAVFGNVLISPFLLHRCAEEGLEVTWFTESGRFRG   79
WP_027883026   1   MTQELLNTLYVQTQGVYLRLEGDTLRVQHEDVT-LRNVPLHHLGGLAVFGNVLISPFLLARCAEEGLEVSWFSESGRFQG   79
WP_027893398   1   MNTQLLNTLYVQAQGAYLRLQGDTVRVEVEGSL-KRQIPLHHLDGLCLFGNVLVSPFLLHRCAQDGREVAWYGENGRFQG   79
WP_119313943   1   MNTQLLNTLYVQAQGAYLRLQGDTVRVEVEGSL-KCQIPLHHLDGLCLFGNVLVSPFLLHRCAQDGREVAWYGENGRFQG   79
WP_071678275   1   MNRILLNSLFVQTQGAYLRLQGDTVRVEVEGEL-RLQVPLHHLGSLVLFGNVLVSPHLLARCSEDGRSVVWLSEHGRFQG   79
WP_013178158   1   MT-ELLNTLYVQTQGSYLRLEHDTL-KLDIEGKtAAQIPLHHLGGLVVFGNVLLSPFLLHRCAEDGRSVVWLSQNGRFKA   78
WP_013703095   1   MNRVLLNTLFVQTQGAYLHLDHEVLEVKVENEV-RLRVPLHHLGNVAVFGQVLVSPFLIHKLVEDGKELVYYTRSGRFRG   79
```

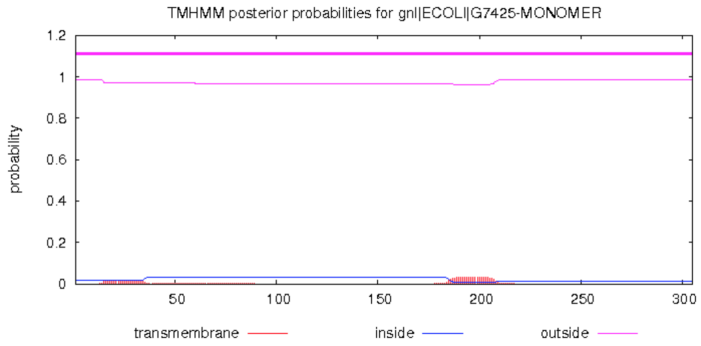
C: mrub_0224 start codon compared with similar species

```
WP_013012523   1   MTLHLTEQSSTLRLSQGRLRVELDEQTLAELPARKVRGVVVWGNVRLTTPALAFLLRQGVPVLYATLEGQLYGQAQAPQS   80
WP_119361385   1   MTLHLTEQSSTLRLSQGRLRVELDEQTLAELPARKVRGVVVWGNVRLTTPALAFLLRQGVPVLYATLEGQLYGQAQAPQS   80
WP_013157212   1   MTLHLTEQSSTLRLRAGRLLVELDEQILAELPARKVRGVVVWGNVRLTTPALAFLLRQGVPVLYATLEGQLYGQAQAPQS   80
WP_027878542   1   MTLHLTEQSSTLRLRQGRLLVELDEQTLAQLPARKVRGVVVWGNVRLTTPALAFLLRQGVPVLYVSLEGQLYGQATALQG   80
WP_119339770   1   MTLHLTEQSSTLRLRQGRLLVELDEQILAELPARKVRGVVVWGNVRLTTPALAFLLRQGVPVLYATLDGQLYGQAIAPLG   80
WP_119275698   1   MTLHLTEQSSTLRLRAGRLLVELDEQILAELPARKVRGVVVWGNVRLTTPALAFLLRQGVPVLYASLEGQLYGQAMAPQG   80
WP_119358090   1   MTLHLTEQSATLRLRQGRLLVELDEQILAQLPARKVRGVVVWGNVRLTTPALAFLLRQEVPVLYTTLEGQLYGQAIAPQG   80
WP_038045136   1   MNLHLTRQGATLRLRQGRLLLEAEGETLASFPARQVRRVAVWGNVRLSTPALTFLLRQGVPVFFFSQDGFLYGVAGAFPE   80
WP_071678023   1   MTLHLAQQGTTLRLREGRLVLEEGGLVLADFPARKVRRVAVWGNVRLSTPALVFLLRQGVPILFLSLEGFLHGVAGAFPE   80
WP_026174948   1   MILHLTHQGAALRLRQGRLLLELEGTSLLSVPARQVRQVAVWGNVRLSNPALGFLLRQGVPILFLSLEGFLYGVAGAFPD   80
WP_053767411   1   MTLHLTHQGATLRLRAGRLLLEKDGITLADFPARQVRRVALWGNVRLSTPALVFLLRQGAPVFFLSLEGFLYGVAGAFPD   80
WP_011229154   1   MTLHLTRQGATLRLRQGRLLLEEEGREVAGFPARQVRSVALWGNVRLSTPALVFLLRQGVPVFFYSLEGFLHGVAGAYPD   80
WP_093006935   1   MTLHLTRQGATLRLRQGRLLLEEEGREVAGFPARQVRSVALWGNVRLSTPALVFLLRQGVPVFFYSLEGFLHGVAGAYPD   80
WP_126190901   1   MVLHLLTQGATLRLRQGRLLLEMEGALLNSYPARQVRQVAVWGNVRLSTPALTFLLRQGIPVLFLSTEGFLYGVAASFPD   80
WP_018461517   1   MHLYLAHQGGTLRLRQGRLLLEGEEGPIASFPARQVRGVALFGNVRLSTPALVFLLRQGAALHFFSLEGALHGSAGAHPD   80
```

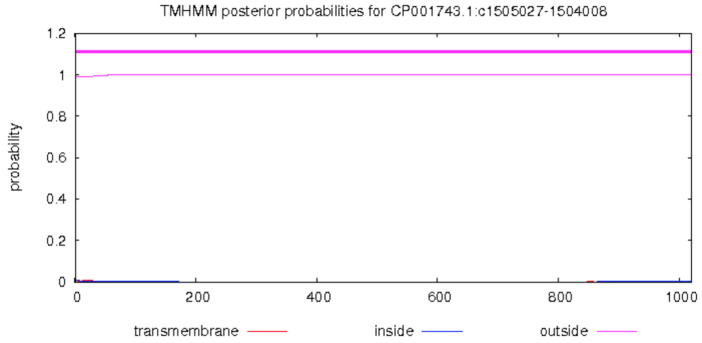
Figure 7. Comparison of start codons for each *M. ruber* gene with evolutionarily similar species. In each panel, the top line is the *M. ruber* Cas1 amino acid sequence. Panel A is mrub_3013, though there are a few gaps, there are no large overhangs and the gaps are found in many of the amino acid sequences, so the start codon appears to have been correctly identified. Panel B is mrub_1477 and shows correct identification of the start codon. Panel C is mrub_0224 and shows correct identification of the start codon.

According to NCBI PubMed Databases, both *M. ruber* and *E. coli* are Gram-negative bacteria. The TMHMM tool predicted zero membrane-embedded alpha-helices in b2755, mrub_3013, mrub_1477, and mrub_0224. Figure 8 shows the transmembrane topology graphs for each protein. Though there are two peaks for transmembrane regions for mrub_3013, the probability of those regions is so low that they are not likely actually crossing a membrane. Protein structure was also predicted and compared using PRED to predict the presence of membrane-embedded beta-barrels. Figure 9 shows the posterior probability plots for b2755 and each *M. ruber* gene. The graph for mrub_3013 is most similar to the one for b2755, and shows that there may be a few beta barrels present in the protein. The posterior probability plots for mrub_1477 and mrub_0224 do not show evidence of any beta-barrels in the proteins.

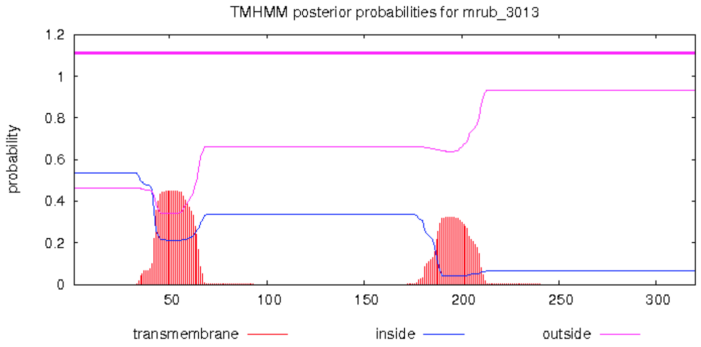
A: *Escherichia coli* b2755



C: *Meiothermus ruber* mrub_1477



B: *Meiothermus ruber* mrub_3013



D: *Meiothermus ruber* mrub_0224

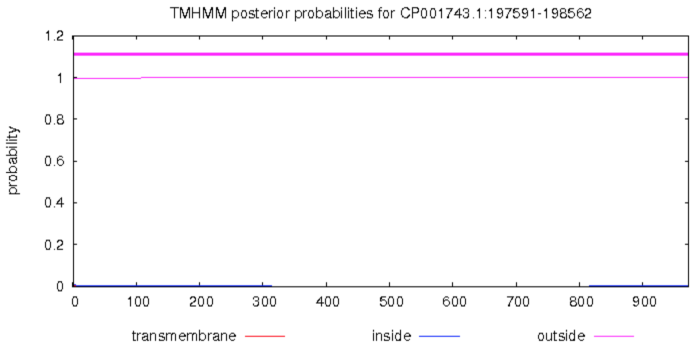
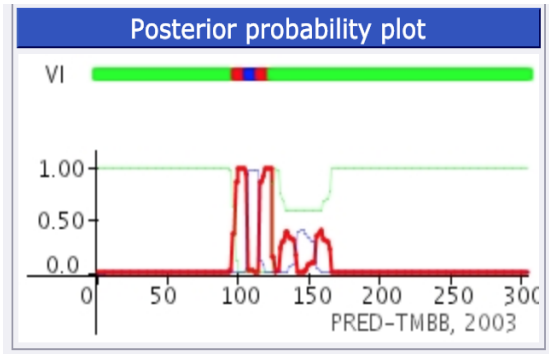
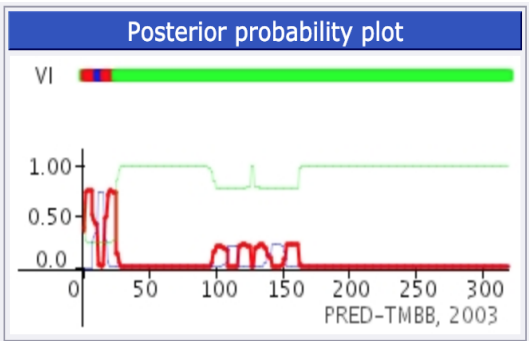


Figure 8. Transmembrane topology graphs for *E. coli* b2755, *M. ruber* mrub_3013, mrub_1477, and mrub_0224. Panel A shows no predicted transmembrane alpha-helices for b2755. Panel B shows two potential transmembrane alpha-helices for mrub_3013, however the probabilities are so low that they are likely not actually transmembrane domains. Panel C and D shows no predicted transmembrane alpha-helices for mrub_1477 and mrub_0224, respectively.

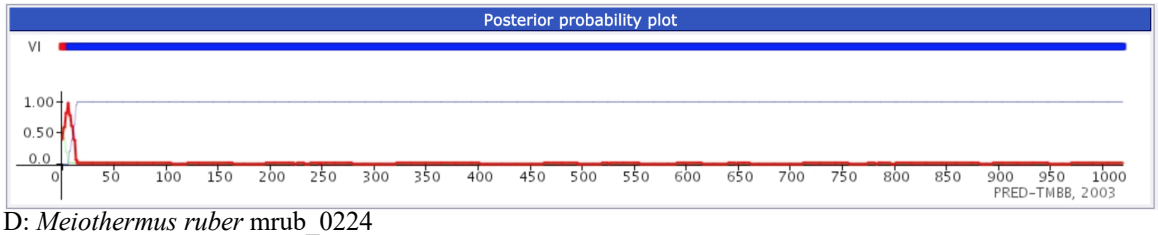
A: *Escherichia coli* b2755



B: *Meiothermus ruber* mrub_3013



C: *Meiothermus ruber* mrub_1477



D: *Meiothermus ruber* mrub_0224

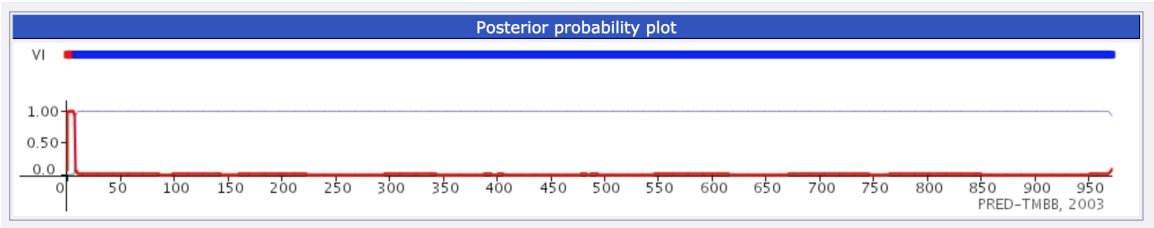


Figure 9. Posterior probability plots for prediction of membrane-embedded beta-barrels for each gene of interest. Panel A is for *E. coli* b2755 and shows a few beta-barrels near the center of the amino acid sequence. Panel B shows a few small peaks for *M. ruber* mrub_3013. Panel C and D show no beta-barrels predicted for mrub_1477 and mrub_0224, respectively.

The PSort-B bioinformatics tool was used to predict the cellular localization of each protein. *E. coli* b2755 is predicted to function in the cytoplasm, as is mrub_3013. The score for mrub_3013 was 8.96 for cytoplasm which is significant for the PSort-B tool, all other scores were too low to be probable areas of function for mrub_3013. *E. coli* b2755 also had a 8.96 score for cytoplasmic localization. The proteins encoded by mrub_1477 and mrub_0224 were not predicted to function anywhere by PSort-B, the data were inconclusive for each amino acid sequence. However, mrub_1477 and mrub_0224 are likely localized to the cytoplasm as well due to their predicted function and predicted protein structure.

The protein structures of b2755, mrub_3013, mrub_1477, and mrub_0224 were further compared using various structural databases. By entering the amino acid sequences of each protein into PFAM, TIGRFAM, and CDD, search hits were collected that were significantly similar to each protein. Table 1 summarizes this data as well as the cellular localization data. B2755, mrub_3013, mrub_1477, and mrub_0224 all pulled the COG group COG1518 from the CDD database. COG1518 is identified as the CRISPR-Cas system-associated endonuclease

Cas1. From the TIGRFAM database both b2755 and mrub_3013 pulled TIGR03638, of the name cas1_ECOLI. This TIGRFAM grouping is labeled as the CRISPR-Cas system-associated endonuclease Cas1 from the CRISPR subtype I-E. Mrub_1477 pulled TIGR0364, which is name cas1_DVULG, and is identified as the CRISPR-associated endonuclease Cas1 of subtype I-C. Finally, mrub_0224 pulled TIGR00287, which is named cas1 and is further identified as CRISPR-associated endonuclease Cas1. Though they were different hits, each gene pulled a CRISPR-associated endonuclease Cas1 from the CDD and TIGRFAM databases.Each gene pulled the same hit from the PFAM database, PFAM01867. PFAM01867 is identified as a CRISPR-associated protein Cas1. To summarize, from all structural protein databases, each gene pulled a hit that was associated with CRISPR-Cas associated protein Cas1.

PDB was also used to pull proteins that were significantly similar to each query protein. *E. coli* b2755 pulled 5VVK, which is the structure of the Cas1-Cas2 complex bound to a full site mimic, the E-value for this hit was 1.28e-168 and it had a bit score of 540.497. Mrub_3013 pulled 3NKD, which is the structure of the CRISP-associated protein Cas1, specifically from *E. coli* K12. The E-value of this match was 3.54e-57 and it had a bit score of 220.32. Mrub_1477 pulled 4WJ0, a CRISPR-associated endonuclease Cas1, with an E-value of 1.13e-25 and a bit score of 115.546. The final query matched mrub_0224 with 4N06, also a CRISPR-associated endonuclease Cas1. This match had an E-value of 4.45e-15 and a bit score of 80.49. Figure 10 shows the alignments between each gene and its respective PDB database match.

Table 1. Summary of data from structural protein databases.

Tool	***E. coli* b2755**	**mrub_3013**	**mrub_1477**	**mrub_0224**
E. coli cas1 BLAST alignment		E-value: 2e-75 Identities: 114/284 (40%)	E-value: 3e-07 Identities: 26/89 (29%)	E-value: 5e-10 Identities: 39/116 (34%)
CDD	COG1518 - Cas1	COG1518 - Cas1	COG1518 - Cas1	COG1518 - Cas1
TIGRFAM	TIGR03638 - Cas1_ECOLI	TIGR03638 - Cas1_ECOLI	TIGR03640 - Cas1_DVULG	TIGR00287 - Cas1
PFAM	PFAM01867 - CRISPR-associated Cas1	PFAM01867 - CRISPR-associated Cas1	PFAM01867 - CRISPR-associated Cas1	PFAM01867 - CRISPR-associated Cas1
PDB	5VVK: Cas1-Cas2 bound to full site mimic	3NKD: Structure of CRISP-associated protein Cas1 from *Escherichia coli* st. K-12	4WJ0: CRISPR-associated endonuclease Cas1	4N06: CRISPR-associated endonuclease Cas1
PSortB	cytoplasm	cytoplasm	unknown	unknown

A: *Escherichia coli* b2755 vs 5VVK

```
       1        10        20        30        40        50        60        70        80        90        10
       |  .     |  .     |  .     |  .     |  .     |  .     |  .     |  .     |  .     |  .     |  .     |
 Query MTWLPLNPIPLKDRVSMIFLQYGQIDVIDGAFVLIDKTGIRTHIPVGSVACIMLEPGTRVSHAAVRLAAQVGTLLVWVGEAGVRVYASGQPGGARSDKLLY
       MTWLPLNPIPLKDRVSMIFLQYGQIDVIDGAFVLIDKTGIRTHIPVGSVACIMLEPGTRVSHAAVRLAAQVGTLLVWVGEAGVRVYASGQPGGARSDKLLY
 Sbjct MTWLPLNPIPLKDRVSMIFLQYGQIDVIDGAFVLIDKTGIRTHIPVGSVACIMLEPGTRVSHAAVRLAAQVGTLLVWVGEAGVRVYASGQPGGARSDKLLY
       |.     |      .     |      .     |      .     |      .     |      .     |      .     |      .     |
       4     10        20        30        40        50        60        70        80        90       100

100       110       120       130       140       150       160       170       180       190       200
|   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .
LYQAKLALDEDLRLKVVRKMFELRFGEPAPARRSVEQLRGIEGSRVRATYALLAKQYGVTWNGRRYDPKDWEKGDTINQCISAATSCLYGVTEXXXXXXXXXXXXXFV
LYQAKLALDEDLRLKVVRKMFELRFGEPAPARRSVEQLRGIEGSRVRATYALLAKQYGVTWNGRRYDPKDWEKGDTINQCISAATSCLYGVTE            FV
LYQAKLALDEDLRLKVVRKMFELRFGEPAPARRSVEQLRGIEGSRVRATYALLAKQYGVTWNGRRYDPKDWEKGDTINQCISAATSCLYGVTEAAILAAGYAPAIGFV
 .     |      .     |      .     |      .     |      .     |      .     |      .     |      .     |      .     |
      110       120       130       140       150       160       170       180       190       200       2

    210       220       230       240       250       260       270       280       290       300  305
    |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |    |
VHTGKPLSFVYDIADIIKFDTVVPKAFEIARRNPGEPDREVRLACRDIFRSSKTLAKLIPLIEDVLAAGEIQPPAPPEDAQPVAIPLPVSLGDAGHRSS
VHTGKPLSFVYDIADIIKFDTVVPKAFEIARRNPGEPDREVRLACRDIFRSSKTLAKLIPLIEDVLAAGEIQPPAPPEDAQPVAIPLPVSLGDAGHRSS
VHTGKPLSFVYDIADIIKFDTVVPKAFEIARRNPGEPDREVRLACRDIFRSSKTLAKLIPLIEDVLAAGEIQPPAPPEDAQPVAIPLPVSLGDAGHRSS
|   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |
210       220       230       240       250       260       270       280       290       300       308
```

B: *Meiothermus ruber* mrub_3013 vs 3NKD

```
       8         20        30        40        50        60        70        80        90        100
       | .     .  |     .   |     .   |     .   |     .   |     .   |     .   |     .   |     .   |     .
 Query LQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGV-VAIPAAALGVLMLGPGTSITHAAIRQLANNGCSVFWVGEEMVRFYASGMGETRSSANLMRQVRA
       L  +P  +D +S ++L++G+++  D A     + G+   IP  ++  +ML PGT ++HAA+R  A  G  + WVGE  VR YASG      S  L+ Q +
 Sbjct LNPIP-LKDRVSMIFLQYGQIDVIDGAFVLIDKTGIRTHIPVGSVACIMLEPGTRVSHAAVRLAAQVGTLLVWVGEAGVRVYASGQPGGARSDKLLYQAKL
       |   |     .   |     .   |     .   |     .   |     .   |     .   |     .   |     .   |     .   |     .
       6  10        20        30        40        50        60        70        80        90       100

  110       120       130       140       150       160       170       180       190       200       210
  |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .
WADPEAHLEVVKRLYRLRFPEPLSPELSLEQIRGLEGVRVRETYARWSRETGVEWKGRNYQRGNWAAADPINRAISAGAACLYGLAHAAILSAGYSPALGFIHTGKQL
 D +  L+VV++++ LRF EP     S+EQ+RG+EG RVR TYA  +++ GV W GR Y   +W   D IN+ ISA  +CLYG+  AAIL+AGY+PA+GF+HTGK L
ALDEDLRLKVVRKMFELRFGEPAPARRSVEQLRGIEGSRVRATYALLAKQYGVTWNGRRYDPKDWEKGDTINQCISAATSCLYGVTEAAILAAGYAPAIGFVHTGKPL
    |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |
    110       120       130       140       150       160       170       180       190       200       210

     220       230       240       250       260       270     278
.    |   .     |   .     |   .     |   .     |   .     |   . |
SFVYDVADIYKAETLIPTAFRVVAESDVGVERRVRHTLREQLKEVKLLERIVSDLHSLFDALE
SFVYD+ADI K +T++P AF +   +    +R VR   R+  +  K L +++  +  +  A E
SFVYDIADIIKFDTVVPKAFEIARRNPGEPDREVRLACRDIFRSSKTLAKLIPLIEDVLAAGE
 .   |   .     |   .     |   .     |   .     |   .     |   . |
     220       230       240       250       260       270  276
```

C: *Meiothermus ruber* mrub_1477 vs 4WJ0

```
       45   50        60        70        80          90        100       110       120       130       140
       |    |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |
 Query LALFGNVLVSPYLLHRCAQDGLEVTWFSESGRFQGRL---AGPVSGNVLLRQAQYRALDNPSSSLYLAQRFVEGKLKNARLVLQRAVRERGXXXXXXXXXX
       + ++G+V ++   LH  AQ G+ + +F+  G + G        +SG++++RQA++    N    L+LA+ FV G  KN    ++R ++  G
 Sbjct IYIYGHVNITSQALHYIAQKGILIHFFNHYGYYDGTFYPRETLLSGDLIIRQAEHYL--NKEKRLFLAKSFVTGGTKN----MERNLKNWGIKAKLSDYLD
       |    |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |
       40   50        60        70        80        90        100       110       120       130

 140       150       160       170       180       190       200       210       220       230       240
 |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .
XXXXXXXXXXXRQLPQARSLDEVRGLEXXXXXXXXXXXXXDLLLSGEFRFDGRNKRPPRDPVNALLSFVYALLTTQCTAALEGVGLDPQAGFLHALRPGRHALALDLLEE
           +L  AR + E+  +E            D  L  EF+   R +RPP++ +NAL+SF+ + L       +    L P   +LH     R +L+LDL E
DYLD-------ELNDARKITEIMNVEARIRQEYYAKW-DENLPEEFKIVKRTRRPPKNEMNALISFLNSRLYATIITEIYNTQLAPTISYLHEPSERRFSLSLDLSEI
           .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |
30            140       150       160       170       180       190       200       210       220       2

     250       260       270       280       290       300       310       320       330     338
.    |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   . |
EEFRAWWADRLALSLLNRKQLAPEHFEVRPGGAVLLNEEGRKAVIVAFQTRRLETVQHPLFKEPVPVGLLPHIQARLLARYLRGDLPQYLPFVG
E F+   ADR+A  L+ +  L  EHF     G VLL EEG K V  A+     ++V+HP     V    L  ++A  L ++L G + +Y P V
EIFKPIIADRVANRLVKKGSLKKEHFREDLNG-VLLTEEGMKIVTKAYNEELQKSVKHPKIGSNVTRQRLIRLEAYKLIKHLVG-VEEYKPLVA
 |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |   .     |
 230       240       250       260       270       280       290       300       310       320
```

D: *Meiothermus ruber* mrub_0224 vs 4N06

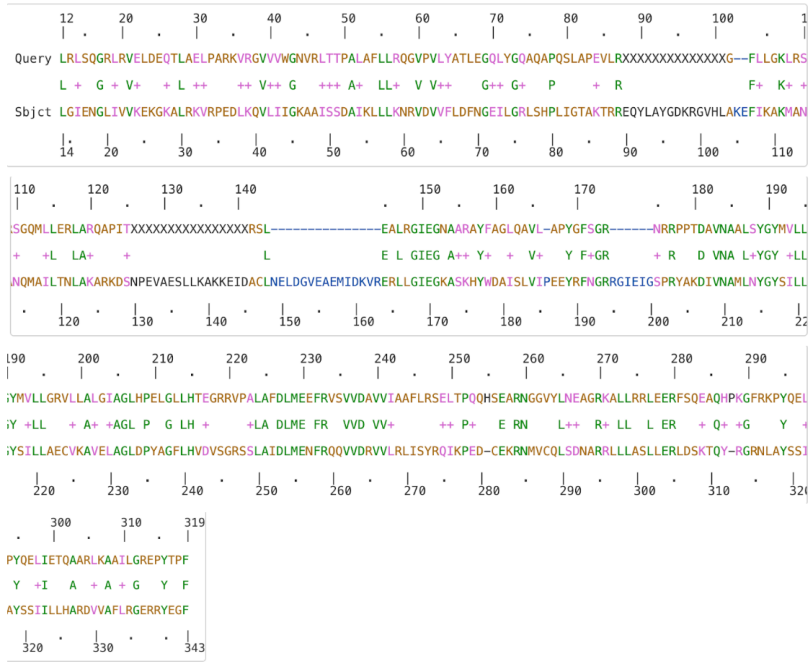
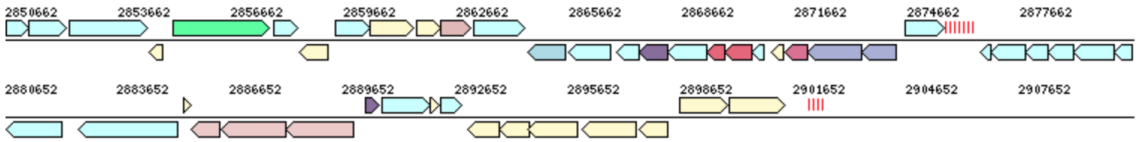


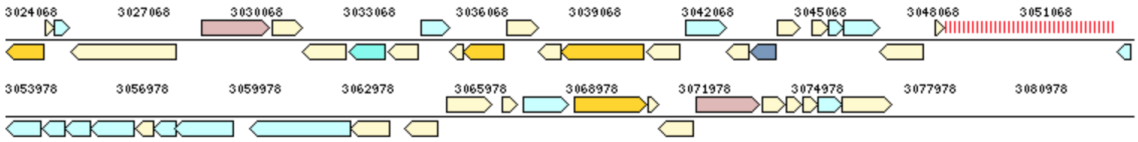
Figure 10. Amino acid sequence alignments of each gene and its top hit from the PDB database. Each gene pulled a significant match with a CRISPR-associated endonuclease Cas1 protein, even though they have different PDB codes. Panel A shows the alignment of *E. coli* b2755 with 5VVK. Panel B shows the alignment of mrub_3013 and 3NKD. Panel C shows the alignment of mrub_1477 and 4WJ0. Panel D shows the alignment of mrub_0224 and 4N06.

IMG/M was again used to investigate the map of the chromosome surrounding each gene and whether they are part of an operon. Figure 11 shows the chromosome maps for each gene of interest and Figure 12 shows the comparison of the operon regions in similar species. To determine if they are part of an operon the genes up and downstream from each gene were noted. For all genes, the gene directly downstream was *cas2*, the upstream gene was *cse3* for b2755 and mrub_3013. The gene upstream from mrub_1477 was *cas4*, and the gene upstream from mrub_0224 was simply called "CRISPR-associated protein." Mrub_3013 appears to be part of an operon similar to that of *E. coli* b2755, it is in a highly conserved region when compared to related species and has similar proteins and gene order to the *E. coli* Type I-E operon. Mrub_1477 also seems to be part of an operon, though not one similar to the *E. coli* operon. Mrub_0224 does not appear to be part of an operon as it is not in conserved area of the chromosome when compared to evolutionarily similar species' chromosome maps.

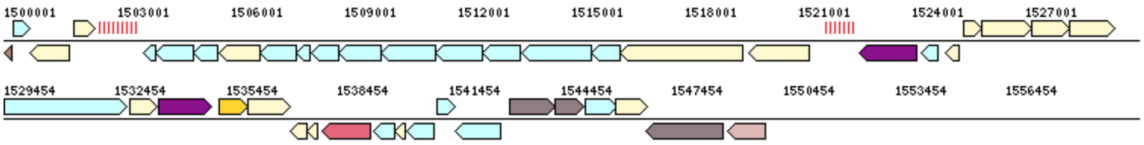
A: Chromosome map of b2755



B: Chromosome map of mrub_3013



C: Chromosome map of mrub_1477



D: Chromosome map of mrub_0224

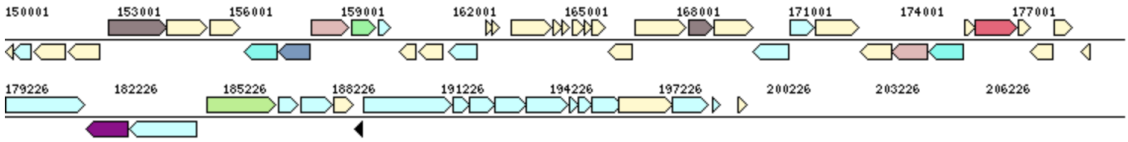
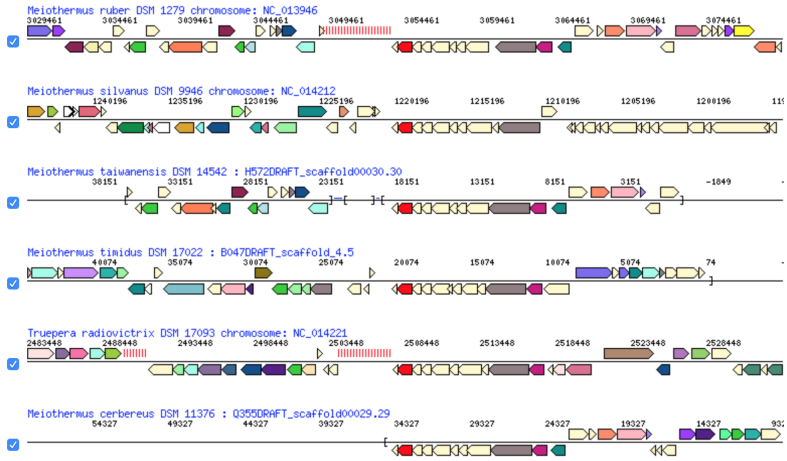
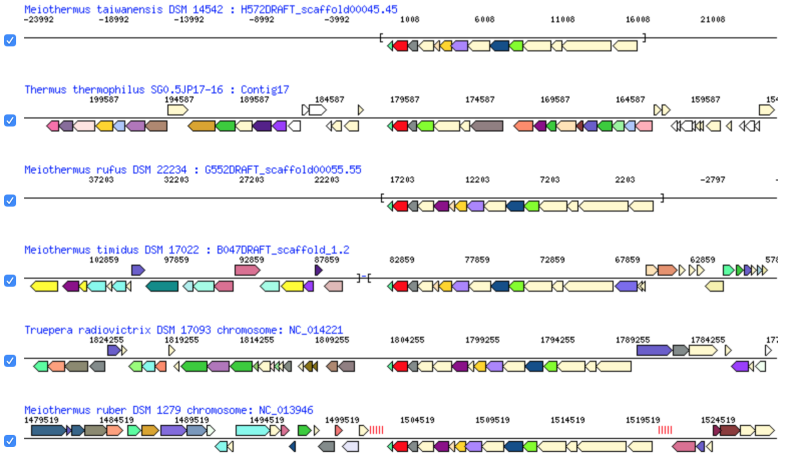


Figure 11. Chromosome maps of each gene of interest. The *cas1* gene in each panel is marked with a red arrow. Panel A is the chromosome map for *E. coli* b2755, Panel B is the chromosome map for mrub_3013, Panel C is the chromosome map for mrub_1477, and Panel D is the chromosome map for mrub_0224. Directly downstream from all *cas1* genes is the *cas2* gene. For mrub_3013 and b2755, the upstream gene is *cse3*. For mrub_1477 the upstream gene is *cas4*, and the gene upstream of mrub_0224 is called "CRISPR-associated protein." The genes appear to be in operons of varying structure.

A: *Meiothermus ruber* mrub_3013



B: *Meiothermus ruber* mrub_1477



C: *Meiothermus ruber* mrub_0224

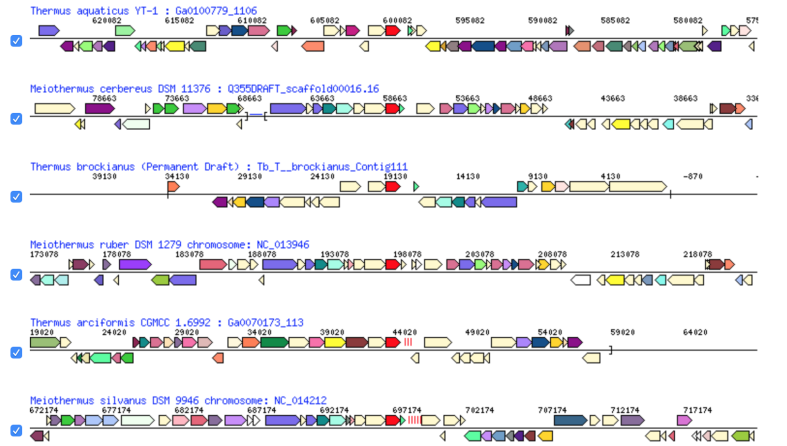
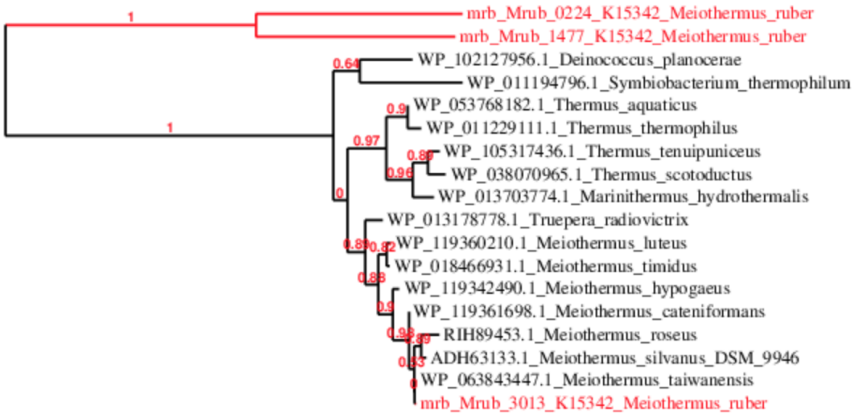


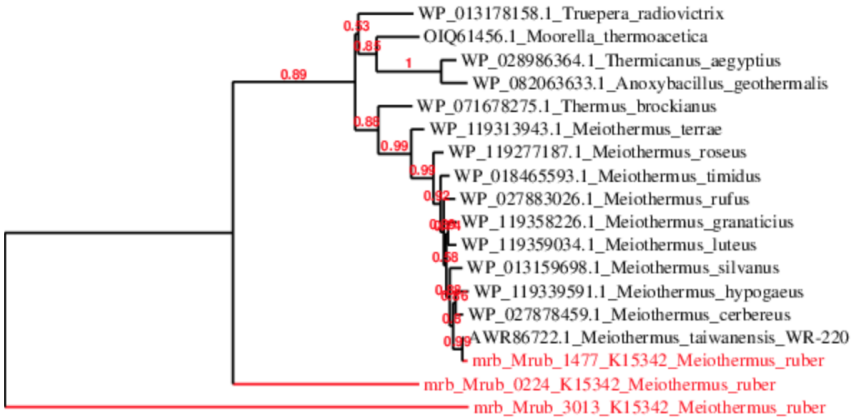
Figure 12. Comparison of operon structures in mrub_3013, mrub_1477, and mrub_0224, respectively, with evolutionarily similar species. Panel A shows relatively strong conservation of the operon structure that mrub_3013 *cas1* is part of with a lot of rearrangement surrounding the operon structure, supporting its role in an operon. Panel B shows some conservation of the operon structure that mrub_1477 *cas1* is part of with a lot of rearrangement surrounding the operon structure, partially supporting its place in an operon. Panel C shows weak conservation of the operon structure that mrub_0224 *cas1* is part of, there is a lot of rearrangement around the gene itself and in the surrounding areas. Mrub_0224 is likely not part of an operon.

Finally, the website phylogeny.fr was used to create phylogenetic trees of each *M. ruber* gene and species with significantly similar sequences. A tree was made based on the NCBI Protein Blast results of similar species, the amino acid sequences of each species' protein was entered and the tree was created based on similarity to show an estimate of the evolutionary relationships between each gene. Figure 13 shows these phylogenetic trees.

A: mrub_3013



B: mrub_1477



C: mrub_0224

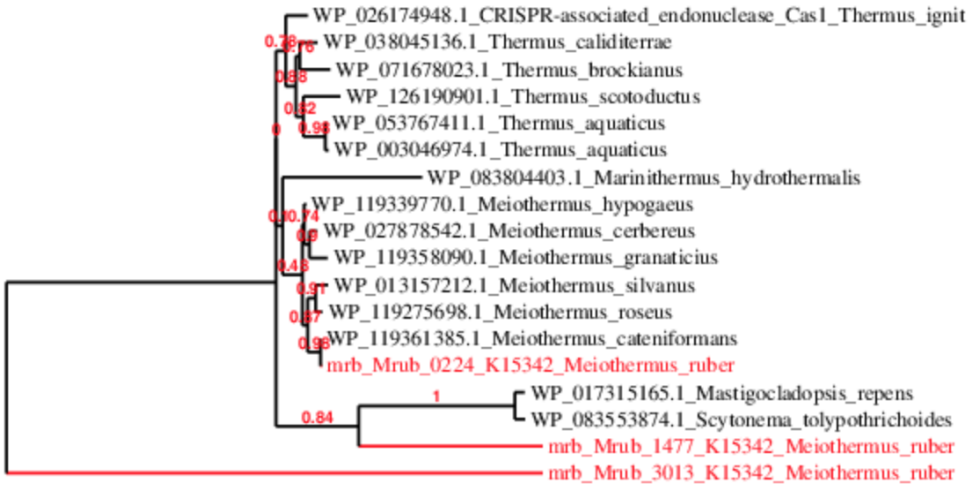


Figure 13. Phylogenetic trees constructed using evolutionarily similar protein sequences. Panel A was constructed using NCBI Protein Blast matches for mrub_3013. Panel B was constructed using NCBI Protein Blast matches for mrub_1477. Panel C was constructed using NCBI Protein Blast matches for mrub_0224. Across all of the phylogenies it is clear that mrub_3013 is on a separate evolutionary branch from mrub_1477 and mrub_0224, who are more similar to each other than they are to mrub_3013.

DISCUSSION

The above results support the orthologous nature of *E. coli* b2755 and *M. ruber* mrub_3013 and raise more questions than answers about the paralogous nature of mrub_3013, mrub_1477, and mrub_0224. Mrub_3013 is most similar to b2755 and the evidence supports that it is part of a Type I-E CRISPR-Cas operon similar to the one found in *E. coli*. *M. ruber* also appears to have Type I-C operon, as evidenced by the similarity between mrub_1477 and other Type I-C Cas1 proteins. While it is clear that each gene codes for CRISPR-associated endonuclease Cas1, the relationship between the genes is unclear. The phylogenies show that there is more similarity between mrub_0224 and mrub_1477, but both are significantly different than mrub_3013. This suggests that if they are paralogs, they either arose through horizontal gene transfer or through gene duplication long enough ago to allow for such divergence.

Future research should investigate the levels of each protein in *M. ruber* to determine which system is most active. Another direction is to replace the *E. coli cas1* gene with one from *M. ruber* to see if is still effective and carries out its function. Further research should also be done into the roles of each *cas1* gene in *M. ruber*, to see if they are paralogs or have some other relationship and how each one evolved to what they are today.

CONCLUSION

Mrub_3013 in *Meiothermus ruber* is an orthologous gene to b2755 in *Escherichia coli*, and it is potentially paralogous to mrub_1477 and mrub_0224. Using structural protein databases, it is clear that each gene is Cas1 and plays a role in the CRISPR-Cas defense system, but what these roles are exactly is unclear for mrub_1477 and mrub_0224. Mrub_1477 and mrub_0224 are more evolutionarily similar to each other than they are to mrub_3013, and may have been the result of horizontal gene transfer or gene duplication, the results are not clear in this respect. Future research should investigate further the relationship between each gene and their roles in *Meiothermus ruber*.

LITERATURE CITED

Bagos PG, Liakopoulos TD, Spyropoulos IC and Hamodrakas SJ.
PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins.
Nucleic Acids Res, 2004 Jul 1;32(Web Server issue):W400-4.

Bagos PG, Liakopoulos TD, Spyropoulos IC and Hamodrakas SJ.
A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins.
BMC Bioinformatics, 2004 Mar 15;5:29.

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/.

Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drablos F. 2010. Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genomics 11(1):588.

Brininger C, Spradlin S, Cobani L, Evilia C. 2018. The more adaptive to change, the more likely you are to survive: Protein adaptation in extremophiles. Seminars in Cell and Developmental Biology 84:158-69.

Darmon E and Leach D. 2014. Bacterial genome instability. Microbiology and Molecular Biology Reviews 78(1):1-39.

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/

Gevers D, Vandepoele K, Simillion C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends in Microbiology 12(4):148-54.

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.

Horvath P and Barrangou R. 2010. CRISPR/cas, the immune system of bacteria and archaea. Science 327(5962):167-70.

Jiang F and Doudna JA. 2015. The structural biology of CRISPR-cas systems. Current Opinion in Structural Biology 30:100-11.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M.; New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47, D590-D595 (2019).

Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.; KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361 (2017).

Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000). Available from: https://www.kegg.jp/kegg/

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.

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

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

Nuñez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. 2014. Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nature Structural & Molecular Biology 21(6):528-34.

Nunez PA, Romero H, Farber MD, Rocha E. 2013. Natural selection for operons depends on genome size. Genome Biology and Evolution 5(11):2242-54.

Sanchez-Perez G, Mira A, Nyrio G, Pasic L, Rodriguez-Valera F. 2008. Adapting to environmental changes using specialized paralogs. Trends in Genetics 24(4):154-8.

Tindall BJ, Sikorski J, Lucas S, Goltsman E, Copeland A, Glavina Del Rio T, Nolan M, Tice H, Cheng JF, Han C, et al. 2010. Complete genome sequence of meiothermus ruber type strain (21). Standards in Genomic Sciences 3(1):26-36.

Wright AV, Nunez JK, Doudna JA. 2016. Biology and applications of CRISPR systems: Harnessing natures toolbox for genome engineering. Cell 164(1-2):29.