

Augustana College Augustana Digital Commons

Meiothermus ruber Genome Analysis Project

Biology

Winter 2-13-2019

Mrub_3018 is Orthologous to *E. coli* B2759 (casB)


Kyle Parker

Augustana College, Rock Island Illinois

Dr. Lori Scott

Augustana College, Rock Island Illinois

Follow this and additional works at: <https://digitalcommons.augustana.edu/biolmruber>

 Part of the [Bioinformatics Commons](#), [Biology Commons](#), [Computational Biology Commons](#), [Genomics Commons](#), and the [Molecular Genetics Commons](#)

Augustana Digital Commons Citation

Parker, Kyle and Scott, Dr. Lori. "Mrub_3018 is Orthologous to *E. coli* B2759 (casB)" (2019). *Meiothermus ruber Genome Analysis Project*.

<https://digitalcommons.augustana.edu/biolmruber/44>

This Student Paper is brought to you for free and open access by the Biology at Augustana Digital Commons. It has been accepted for inclusion in Meiothermus ruber Genome Analysis Project by an authorized administrator of Augustana Digital Commons. For more information, please contact digitalcommons@augustana.edu.

Kyle Parker

What is the *Meiothermus Ruber* Genome analysis project?

The *Meiothermus ruber* Genome analysis project is the study of a bacteria called *Meiothermus Ruber* (*M. ruber*). The studies goal is to find out as much as possible about this bacterium and how it's biological process compares to other well-known bacteria like *Escherichia Coli* (*E. coli*). The current goal of the project is to study the similarities and differences between the system called Clusters of Regularly Interspersed Short Palindromic Repeats or CRISPR-Cas system in *E. coli* and *M. ruber*. The reason for using *E. coli* in our study is that thousands of articles have been published on *E. coli*, which demonstrates that it is a well -studied organism. On the other hand, *M. ruber* is poorly studied, as evidenced by having fewer than 100 cited articles in PubMed (Scott, personal communication). *E. coli* serves as the model organism for our study because its CRISPR-Cas system is well-documented (*e.g.*, see recent review articles by Jiang and Doudna, 2015; Wright *et al.*, 2016). CRISPR-Cas is an important defense mechanism for many bacteria and most Archaea. It is one of several systems that fight off foreign invading DNA that destroys the cell as a consequence of the infection process. Recently CRISPR-Cas has been studied more significantly because of its gene editing properties. (Jiang *et. al*, 2015) This means that CRISPR-Cas is an important starting place for the genome analysis project as it could show a difference in function that could change the way the world thinks about CRISPR and could lead to more studies being done on *M. ruber*.

What is *M. ruber*?

Meiothermus Ruber was initially discovered in 1975 by Loginova and Egorova (Loginova and Ergova, 1975). The name for *M. ruber* essentially means less hot red bacteria from its Greek and Latin roots. *M. ruber* is a gram-negative rod-shaped bacterium that is red in color and is found mostly in warm environments (35-75°C) (Tindall *et. al*, 2010). The optimal temperature for growth is around 60°C. *M. ruber* also has no reported pathogenicity (Field *et. al*, 2008)

Meiothermus as a genus only has a few bacteria in it most of which are understudied bacteria. However, several important key features have been discovered as a result of sequencing *M. ruber*'s genome (Tindall *et. al*, 2010), such as its base pair length, an initial estimate on the number of coding regions, and a preliminary identification of the components of its CRISPR-Cas system.

What is the CRISPR-Cas system?

CRISPR is a part of DNA that is involved with protecting the cell against foreign DNA. The Cas of CRISPR- Cas stands for CRISPR associated proteins (Jiang *et. al*, 2015). The model system *E. coli* has a single type of CRISPR- Cas system, called Type I-E, as part of its genome and it is used to fight off bacteriophages from putting its DNA into the *E. coli* cells.

In *general*, the CRISPR-cas system works in three steps: spacer acquisition, CRISPR-Cas expression and DNA interference (Jiang *et. al*, 2015). The first step, called spacer acquisition, is one in which the CRISPR- Cas system will obtain a spacer from invading DNA. Spacers are derived from sections of invader bacteriophages' DNA called protospacers that allows the cell to identify the foreign DNA's presence. The protospacers are recognized by the CRISPR CASCADE using a PAM sequence. PAM sequence or protospacer adjacent motif is a sequence of DNA that is upstream of the protospacer region and allows the host cells' CRISPR system to recognize the foreign DNA's arrival (Redding *et. al* 2015). The PAM sequence distinguished the foreign DNA from the CRISPR RNA(crRNA). During spacer acquisition the Cas1 and Cas2 proteins identify the PAM sequence and then recruit Cas3 for cleavage of the protospacer from the invading DNA (Jiang *et. al*, 2015). The cell will then start the CRISPR expression stage (*aka*, CRISPR RNA biogenesis stage) in which the spacer is transcribed into a large RNA strand called the pre-crRNA. The pre-crRNA is processed during the expression stage into mature crRNA and has a formation containing three parts the leader sequence, the repeats and a spacer. The leader sequence occurs before the first repeat in the crRNA and its function is to signal the beginning of the CRISPR RNA. Brault *et al.* (2012) proposes that the leader sequence contains the promoter for the whole crRNA region. After the leader sequence a repeat is present. The repeats are palindromic repeats and occur in between, before and after the spacers and form hairpins. The final portion is the acquired spacer (Jiang *et. al*, 2015). Once the mature crRNA has been made, the final step called DNA interference begins. The newly created crRNA will now form a complex with Cas proteins to fend of invading DNA that is recognized using the spacer regions as reference points. When the system recognizes a foreign invading DNA, with a protospacer and PAM sequence, the crRNA will bind to the protospacer on the invading DNA using the integrated spacer in the crRNA, forming an R-loop (Westra *et.al*, 2012). The CasA protein will then recognize the R-loop and recruit Cas3 to degrade the invading DNA. The whole system is then called CRISPR-associated complex for antiviral defense or CASCADE

Additional Cas proteins also play a role in *E. coli* CRISPR-Cas defense. *E. coli* contains 8 cas genes which are *cas3*, *cse1*, *cse2*, *cas7*, *cas5e*, *cas6e*, *cas1* and *cas2*. The proteins unite into the S formation (Zhao *et. al*, 2014). CasE forms the head of the structure while CasD, CasA and CasC form the tail of the S-formation. CasC and CasB will then connect the head and tail of the S-formation. Zhao mentions that CasC and CasB will then form a backbone which is important for binding the foreign DNA to the CASCADE complex. The most important of the proteins are *Cas 1-3*. *Cas1* and 2 are important for spacer acquisition and *Cas3* is the helicase used to cut the spacer out (Jiang *et. al*, 2015). A picture of the complex is shown below in Figure 1.

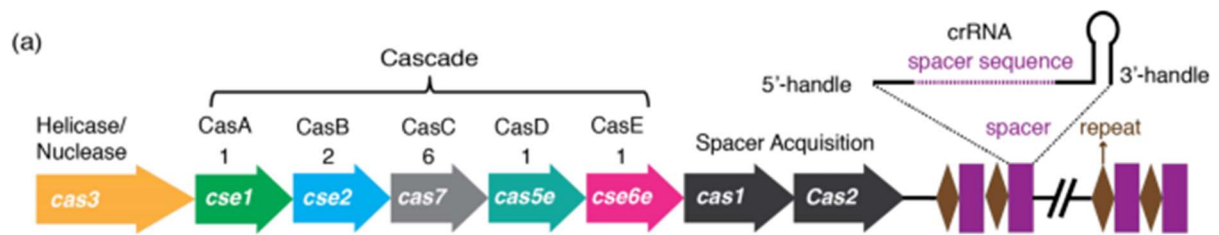


Figure 1. Above is a picture of the CRISPR-cas CASCADE system. To the left is the Cas proteins CASCADE in *E. coli* including all 8 essential proteins, to the right is a sample crRNA from *E. coli*. *CasB* is light blue in the picture above. (Jiang *et. al*, 2015).

The CRISPR CASCADE is highly versatile in every bacterium so much so that scientists have found ways to use it outside out cells for gene editing which is one of the main applications of CRISPR. CRISPR has been highly studied in recent years specifically because of this reason. Scientist were able to remove and purify an entire CRISPR-Cas CASCADE (Beloglazova *et. al*, 2015). The hope by those scientists is to use it for gene editing in humans eventually, barring the approval by an ethics board since this is a highly debated topic across the globe.

What is *CasB* or *cse2* in *E. coli*?

Cse2 also known as *CasB* as mentioned above is part of the CRISPR- Cas CASCADE and will help function to fight off invading DNA into the cell. *CasB* so far has only been found in class 1-E type CRISPR systems (Nam *et. al*, 2012). *CasB* is 160 aa and 609 nt long and its gene is located at 2,882,155 <- 2,882,637 in the cell's DNA (Jackson *et. al*, 2014). *E. coli* has been noted to have an average number of 5-10 *casB* genes per cell (Twiss *et. al*, 2005). *Cse2*'s main function is as a structural protein (Nam *et. al*, 2012). *Cse2* will connect the head and tail of *the S* formed in the CASCADE. *Cse2* does not make any direct connection with the crRNA in the complex (Jackson *et. al* 2014). It has been noted that both ends of the *cse2* protein are positively charged which some say has indicated is used to stabilize displaced DNA (Jackson *et. al*, 2014). A function that helps with the complex besides stabilizing the structure has not yet been determined for *E. Coli* It has been noted however that a removal of *casB* has shown either a significant decrease in function of CRISPR or a loss of function, which indicated that as a protein *casB* is important to CRISPR function (Nam *et. al*, 2012).

Purpose question of study.

Is Mrub_3018 orthologous to *CasB* in *E. coli* (B2759)

Materials and Methods

Gathering data for this project on *M. ruber* and *E. coli* involved using a collection of free online bioinformatics tools chosen by Dr. Scott. The first tool that was used is EcoCyc, from

which we collected background information on CRISPR-Cas system, especially *casB*/CasB, in the model organism. EcoCyc is a website dedicated to *E. coli* K-12 MG1655. (Kessler *et. al*, 2013). It helped to find the location, amino acid sequence of the protein, chromosomal location, and the operon structure.

The next bioinformatics tool used in this project was KEGG (Kanehisa *et. al*, 2019). KEGG mines the GenBank database (Benson *et. al*, 2008) for genes predicted to be part of metabolic pathways and cellular structures. We used KEGG to predict if *M. ruber* had CRISPR-Cas genes and to compare it to the known *E. coli* system. (Kanehisa *et. al*, 2019).

Because most of the bioinformatics tools in this project used the amino acid sequence as the query, it was necessary to confirm that we were using the correct CasB sequence. Consequently, we used the Department of Energy's database called IMG/M (Markowitz *et. al*, 2012), specifically the "Sequence Viewer For Alternate ORF Search" program, to analyze the 5' upstream region of *casB* for a potential Shine-Delgarno sequence and an alternative start codon. Subsequently, we created a multiple sequence alignment using NCBI BLAST (Madden *et. al*, 2002) to confirm that similar proteins have comparable amino termini. NCBI BLAST is a bioinformatics tool that creates both a pairwise sequence alignment between two proteins or aligns many sequences to each other. Once we confirmed that we likely have the correct amino acid sequence, we aligned the *E. coli* CasB and putative *M. ruber* CasB using NCBI BLAST.

The next step was studying where *CasB* functions in the cell. The three tools we used to determine this are TMHMM, PRED and PSORT-B. TMHMM is a bioinformatics program used to determine if there are any transmembrane helices present, by reading if any hydrophobic amino acids are present in the amino acid sequence (Krogh *et. al*, 2001). PRED is used to determine if a protein has any beta-barrels present (Bagos *et. al*, 20-4). PSORT-B is a bioinformatics tool that will analyze a sequence for its most likely location in the cell (Yu *et. al*, 2010).

The next group of studies were chosen to find the protein domain and family. The tools used were COG, TIGRFam, Pfam, and PDB. COG is a bioinformatics tool that is used to determine the proteins domain (Marchler-Bauer *et. al*, 2016). TIGRFam is a bioinformatics tool used to find the proteins family by comparison from its own database (Haft *et. al*, 2001). Pfam is another bioinformatics tool that will compare the domain of the protein to similar domains it has in its database and form a consensus sequence as well as a logo for the protein (Finn *et. al* 2016). Finally, PDB is used to compare the proteins to proteins from its database and to find a crystalized picture of the protein (Berman *et. al*, 2016). The tool will also provide the user with data on family and domain as well as articles that discuss function of the analyzed protein.

The next step was to determine if the gene of interest (GOI) is part of an operon. The two tools used for this were KEGG and IMG. KEGG is a tool used to see if an operon is present for Mrub_30183018 (Kanehisa *et. al*, 2019). KEGG will show a picture of the chromosome near the GOI. After KEGG IMG's chromosome map was used and compared to the chromosome maps of other similar bacteria (Markowitz *et. al*, 2012).

The final step was to determine if the proteins are paralogs. Using the Mrub_3018/CasB amino acid sequence as the query, we used NCBI BLAST to search the *M. ruber* genome for a possible paralog (Madden *et. al*, 2002).

Results

Table 1 summarizes the results of many different bioinformatics tools used to assess the similarity between Mrub_3018 and *E. coli* B2759. The first row of data is the KEGG output used to find the locus tags of each gene as well as their protein product, and chromosomal coordinates (Kanehisa *et. al*, 2019). According to KEGG the product of both genes is the same protein. However, they do have different locus tags which is to be expected by genes from different bacteria. The KEGG data also showed the amino acid sequences for both which was 202 and 160 for Mrub_3018 and *E. coli* B2759, respectively. The KEGG data also shows the proteins location on the chromosome. The two genes are at different locations. The next row is the pairwise sequence alignment BLAST data when comparing Mrub_3018 and B2759 (Madden *et. al*, 2002). The percent identity and the E-value from this tool do not support that these two genes are related. The E-value of 1.5 shows that the genes are matched up by chance and the percent identity is 50% indicating that the two genes only match up 50% of the time. However, the two genes have varying amino acid lengths which makes them hard to compare in a pairwise alignment. The next row is COG data (Marchler-Bauer *et. al*, 2016). The COG hits for both locus tags produce the same product but a different number. This is significant in that the genes will create the same protein but may vary due to amino acid sequence. The score is high and the E-value is close to zero indicating that the sequence was not matched by random chance and that the genes make a protein from the same family. The next row is PSORT-B data (Yu *et. al*, 2010). As mentioned above PSORT-B will give a localization score for every location in a cell and the highest is where the protein is predicted to be found. In this case for Mrub_3018 it is found in cytoplasm and for B2759 it was unknown. The next row is TMHMM data which shows no transmembrane helices for both Mrub_3018 or B2759 indicating they are not part of the membrane of the cell (Krogh *et. al*, 2001). PRED is the data in the next row and shows that no beta barrels are present indicating it is not a transmembrane protein (Bagos *et. al*, 2004). TIGRfam is the data from the next row (Haft *et. al*, 2001). Both Mrub_3018 and *E. coli* had the same protein product and same TIGRfam number. The score for both genes is high and the E-value is low indicating the two sequences were not matched by random chance, showing that there is a significant chance these two proteins are in the same family. The next row is Pfam data (Finn *et. al*, 2016). The Pfam number and product, like TIGRfam, were the same. This indicates they are in the same domain. The low E-value and high bit score indicate that the sequences were not compared with random chance. The final row is the data from PDB. The PDB numbers were different in this case but the genes were found to be a part of the same protein family. The bit score for both sequences was higher than all other tools as well as the E-value being low indicates that the sequences were not matched by random chance with the found family, confirming the TIGRfam data that the two genes are part of the same family. PDB also shows a picture of the crystalized protein. The proteins look different but based on previous data are part of the same family and domain and produce the same product indicating they perform the same function in their respective bacteria (Berman *et. al*, 2016).

Table of Bioinformatics data

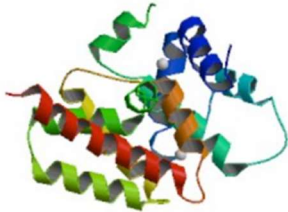
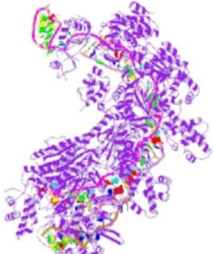
Bioinformatics tool	<i>M. ruber</i> Mrub_3018 gene	<i>E. coli</i> B2759 gene
KEGG(locus tags)	Mrub_3018 Product: CasB Amino Acid number: 202 Coordinates: 3057916... 3058524	B2759 Product: CasB Amino acid number: 160 Coordinates: 2882155... 2882637
Blast Mrub_3018 vs B2759	E-value 1.5 Percent Identity 50 percent	
COG hits	Cog number: <i>cl09719</i> Cog name: Cse2 I-E family	Cog Number: cd09670 Cog name: Cse2_I-E family
	Score 110.72 E-value 7.21e-31	Score 120.25 E-value 1.80e-35
PSORT-B	Location: cytoplasm	Location: unknown
TMHMM	Transmembrane helices:0	Transmembrane helices:0
PRED	Beta barrels: 0	Beta Barrels:0
TiGRfam- protein family	TiGRfam number: TIGR02548 TiGRfam name: <i>CRISPR-associated protein Cse2</i>	
	Score 118.2 and E-value 3.4e-32	Score 177.8 and E-value 4e-50
Pfam	Pfam number: PF09485 Pfam name: CRISPR_Cse2	
	E-value 1.84e-34	E-value: 1.1e-09
PDB hit	PDB code- 3WA8 PDB name <i>CRISPR-associated protein, Cse2</i> Score 410.223 E-value 9.66927e-115	PDB code 5H9E PDB name Crystal structure of <i>E. coli</i> CASCADE Score 331.643 E-value 3.29589E-91
	Cellular picture 	Cellular picture 

Table 1. data from various bioinformatics tools. In descending order: KEGG, NCBI blast, GenBank, COG, PSORT-B, TiGRfam, Pfam, and PDB. (Kanehisa *et. al*, 2019), (Madden *et. al*, 2002), (Marchler-Bauer *et. al*, 2016), (Yu *et. al*, 2010), (Krogh *et. al*, 2001). (Bagos *et. al*, 2004). (Haft *et. al*, 2001), (Finn *et. al*, 2016), (Berman *et. al*, 2016)

Once the correct start codon was found the aforementioned BLAST pairwise sequence alignment was done between Mrub_3018 and *E. coli* B2759. The pairwise alignment is shown below in Figure 3 (Madden *et. al*, 2002). The alignment only had a percent identity of 50% and only a bit score of 14.2. the E-value was also high at 1.5. The high E-value and low bit score indicates that the sequences were aligned by chance and are not directly the same sequence. The two sequences are different in size which could cause the bit score to be low and the E-value to be high. Despite the fact the two genes produce the same product they will have evolutionary differences that have caused a poor sequence alignment.

Range 1: 112 to 121 Graphics				▼ Next Match	▲ Previous Match
Score	Expect	Method	Identities	Positives	Gaps
14.2 bits(25)	1.5	Compositional matrix adjust.	5/10(50%)	8/10(80%)	0/10(0%)
Query	22	ADLARMRRGL	31		
		AD+ ++RR L			
Sbjct	112	ADMVQLRRLL	121		

Range 2: 27 to 46 Graphics				▼ Next Match	▲ Previous Match	▲ First Match
Score	Expect	Method	Identities	Positives	Gaps	
13.1 bits(22)	3.0	Compositional matrix adjust.	7/20(35%)	12/20(60%)	0/20(0%)	
Query	156	RIYWSELLRDLLAWNRRERKP	175			
		R+ + LRD+ A+ R +P				
Sbjct	27	RVSEPDELRDIPAFYRLVQP	46			

Figure 3. an NCBI BLAST pairwise sequence alignment of *Mrub_3018* and *E. coli* B275. The *Mrub_3018* sequence is above while the B275 sequence is below. (Madden *et. al*,2002)

After the pairwise alignment the location of both proteins in the cell were determined using several bioinformatics tools. The first tool was TMHMM which is used transmembrane helicies (Haft *et. al*, 2001). The tool uses a graph to determine if the protein has a transmembrane helix and if spikes are present on the graph then the protein is transmembrane. A flat line indicates the protein does not have a transmembrane helix. Shown below, in Figure 4, is the topographical map output from *E. coli* B2759 and Mrub_3018. Both topographical maps have the same result. Both graphs are flat and have no transmembrane proteins present. The TMHMM data rules out both proteins being present in or on the cell

membrane.

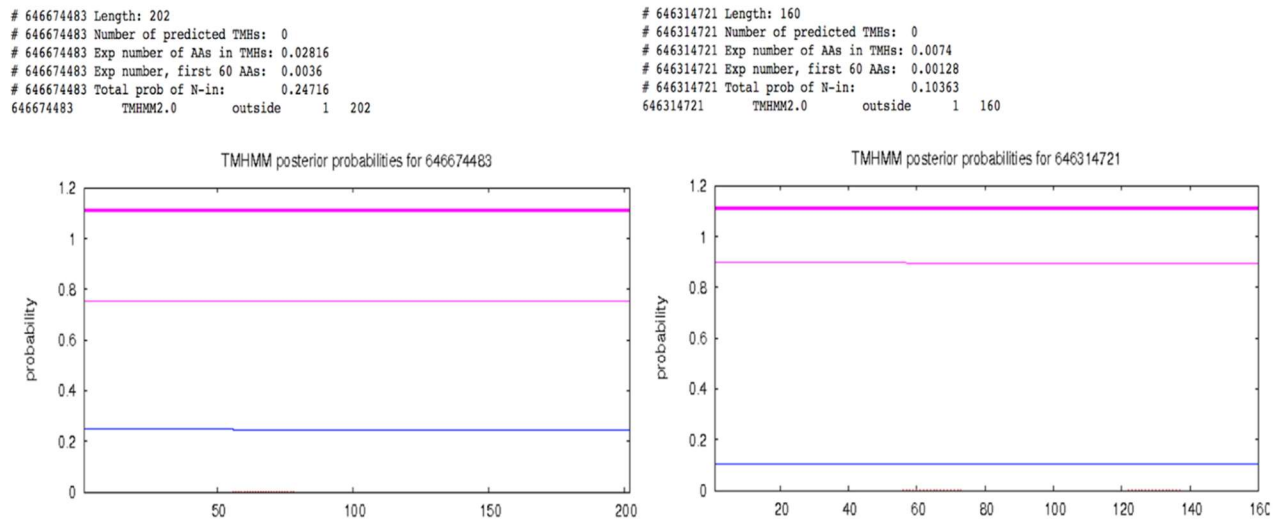


Figure 4. TMHMM output for *E. coli* B2759(right) and Mrub_3018(Left). The outputs show no transmembrane helices, indicating that this protein is not in or on the cell membrane of the cell. (Haft *et. al*, 2001)

The next bioinformatics tool used to find location was PRED (Bagos *et. al* 2004). PRED tests for membrane imbedded beta-barrels. Figure 5 is the graph output of PRED. In PRED the graph will go up and down consistently throughout the graph if a beta-barrel present. Mrub_3018 and *E. coli* B2759 both are not membrane imbedded beta-barrels according to the graph because while there are several spikes in the data it is not consistent across the graph. Since neither Mrub_3018 and *E. coli* B2759 are not membrane bound proteins it is most likely that they are in the cytoplasm.

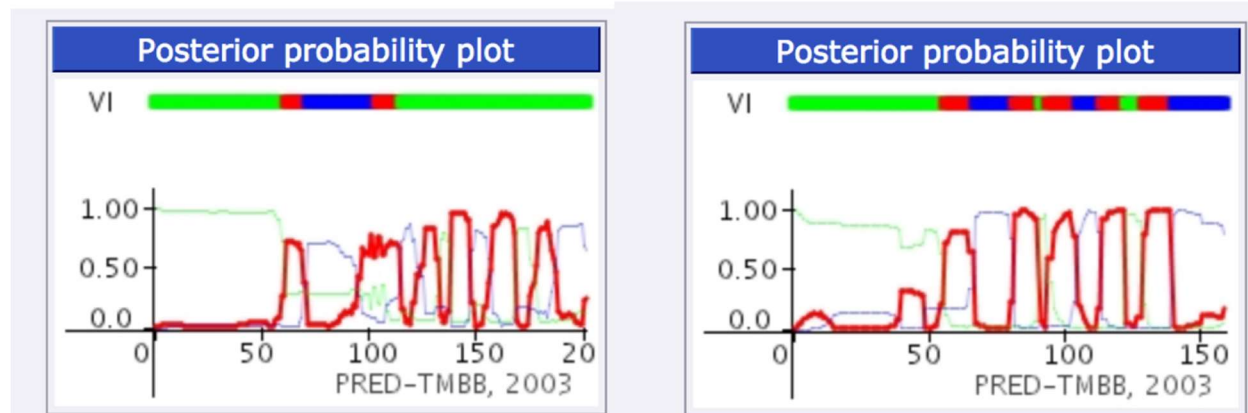


Figure 5. PRED data of Mrub_3018(left) and *E. coli* B275(right). The graph for both genes shows that a beta-barrel is not present due to the inconsistency of the peaks. (Bagos *et. al* 2004)

The final tool used for location is PSORT-B which will give a score to indicate what location is most likely for the protein (Yu *et. al*, 2010). Table 2 shows the scores for each

location for both Mrub_3018 and *E. coli* B2759. In the scores for Mrub_3018 the highest score is cytoplasm which indicates that it is most likely in the cytoplasm. The scores for *E. coli* B2759 are all the same so the data is indeterminate for *E. coli*. From previous data, indicating it is not a membrane bound protein, it is most likely in the cytoplasm as well.

Table for PSORT-B data

Cellular location	Scores	
	<i>Mrub_3018</i>	<i>E. coli B2759</i>
Cytoplasmic score	8.96	2
Cytoplasmic Membrane score	.51	2
Cell wall score	None	None
Periplasmic score	.26	2
Outer Membrane score	.01	2
Extracellular score	.26	2

Table 2. Table of data from PSORT-B that shows the scores for each location of the cell. The data shown in the table suggests cytoplasm for *Mrub_3018* and is indeterminate for *E. coli*. (Yu *et. al*, 2010)

Once the location was determined tests were performed to see if the genes were part of an operon. The first tool used to determine if the genes are in an operon is KEGG (Kanehisa *et. al*, 2019). KEGG has a picture of the genes flanking either side of the GOI. In Mrub_3018 the gene is flanked by Mrub_3017 and Mrub_3019. Both of these genes are part of the *M. ruber* CRISPR CASCADE indicating that it is part of an operon. In *E. coli CasB* or *B2759* is flanked by *CasA* and *CasC*. Both *CasA* and *CasC* are part of the *E. coli* CRISPR CASCADE. This indicates that the GOI in *E. coli* is also part of an operon. Shown in Figure 6 is the KEGG operon picture for *E. coli* and *Mrub_3018*.

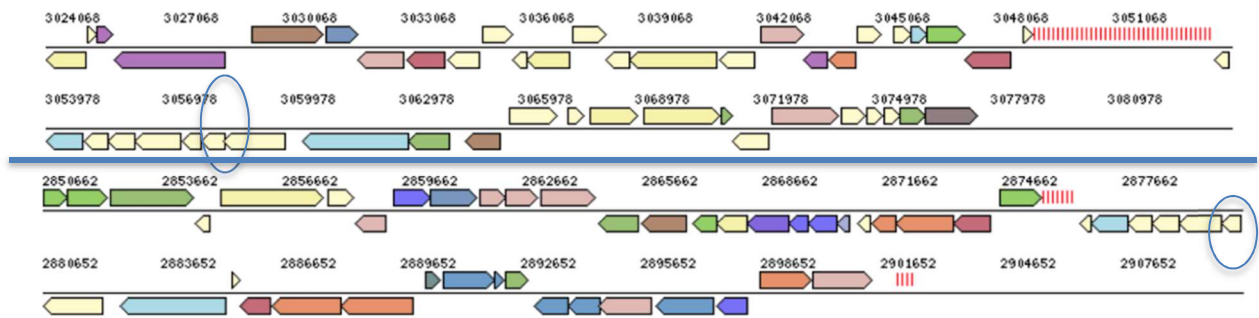


Figure 6. an image of the genes flanking the GOI from KEGG in both *E. coli* and *M. ruber* are shown above. *M. ruber* is the top picture and *E. coli* is the bottom picture. The two GOIs are circled above. the Two pictures are separated by the line in the center. (Kanehisa *et. al*, 2019),

In order to confirm that the *Mrub_3018* gene is part of an operon it was compared to other similar bacteria in the same area of the chromosome. The other bacteria should have the

same genes flanking the GOI and the flanking genes should be a part of the same function as the GOI. Figure 7 shows other bacteria compared to *M. ruber* with the GOI highlighted in red. The GOI is flanked by the same gene in each of the similar bacteria and the flanking genes have the same function. This data confirms that Mrub_3018 is part of an operon, which makes it similar to *E. coli* B2759 as it is also part of an operon.

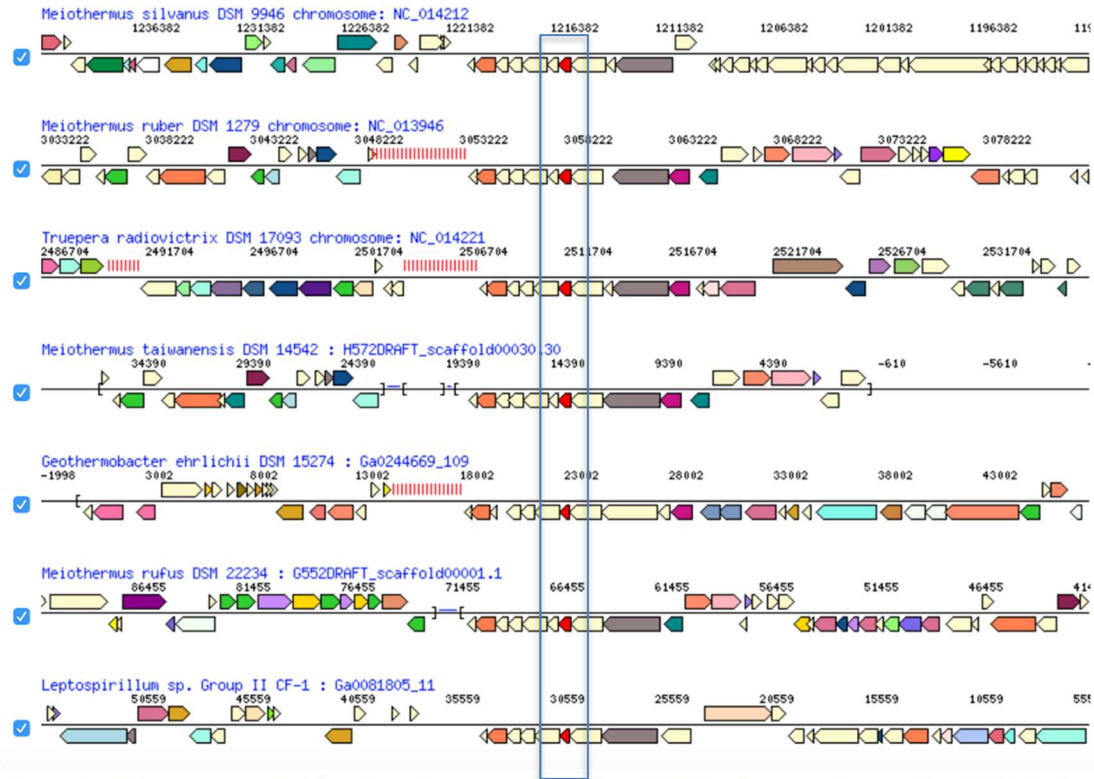


Figure 7. The gene neighborhood comparison by IMG is shown above. the GOI is shown in red in each bacteria inside of the box. The GOI is in the same place in each bacteria and is flanked by a similar gene in each bacterium. This indicates that the GOI is part of an operon. (Markowitz *et. al*, 2012)

The final test was to see if *M. ruber* has a paralog in within its genome for Mrub_3018. The tool used for this was NCBI BLAST in the case that Mrub_3018 was run against its own genome (Madden *et. al*, 2002). One paralog was found by NCBI which was glycosyltransferase. A pairwise alignment was run and is shown below in Figure 8. The percent identity is 28 percent and the E-value is 7.3. The high E-value and the low percent identity indicates these sequences were matched up by chance and are not paralogs. *E. coli* B2759 was also tested for a paralog and it was compared to a different version of *CasB* from itself. The low E-value and high bit score indicate it is a paralog.

Score	Expect	Method	Identities	Positives	Gaps
23.5 bits(49)	7.3	Compositional matrix adjust.	14/50(28%)	27/50(54%)	3/50(6%)
Query 112	QLYLARDQSKSIEQR---FIALLDADDEEQLPYRLRQMVQLIESQDDIRIY			158	
	L +A ++++I Q +IA LDAD+ P +LR+ + L + ++				
Sbjct 81	NLGVALSRNRRAISQASGDWIAFLDADDVWHPNKLREQLHLARLNSSLVVF			130	

Score	Expect	Method	Identities	Positives	Gaps
329 bits(844)	6e-116	Compositional matrix adjust.	159/160(99%)	159/160(99%)	0/160(0%)
Query 1	MADEIDAMALYRAWQQLDNGSCAQIRRVSEPDELRDIPAFYRLVQPPFGWENPRHQALLR			60	
Sbjct 1	MADEIDAMALYRAWQQLDNGSCAQIRRVSEPDELRDIPAFYRLVQPPFGWENPRHQALLR			60	
Query 61	MVFCLSAGKNVIRHQDKKSEQTTGISLGRALANSGRINERRIFQLIRADRTADMVQLRRL			120	
Sbjct 61	MVFCLSAGKNVIRHQDKKSEQTTGISLGRALANSGRINERRIFQLIRADRTADMAQLRRL			120	
Query 121	LTHAEPVLDWPLMARMLTWWGKRERQQLLEDFVLT TNKNA		160		
Sbjct 121	LTHAEPVLDWPLMARMLTWWGKRERQQLLEDFVLT TNKNA		160		

Figure 8. a pairwise alignment between glycosyltransferase and Mrub_3018. The sequences do not align well as indicated by the high E-value and low percent identity. This means this is not a paralog. *E. coli* B2759 was compared to a protein from the same family as itself from within *E. coli* and had a low E-value with a high bit score. This means the other protein is a paralog of *E. coli* B2759. The pairwise alignment of the two are shown in the second alignment. (Madden *et al* 2002)

Conclusion.

The results from the bioinformatics tools indicate that *E. coli CasB* and Mrub_3018 have an orthologous relationship. This means that these two genes are related by evolution. The evidence for this starts with their structural relationship. The TIGRfam data shows that they have the same name and number from TIGRfam. This would indicate that the two proteins are in the same family. Next domain was tested with Pfam. Pfam gives the same PFam name and number for both genes which indicates they are part of the same domain. COG confirmed that they are part of the same domain by giving the same product for the two orthologs. The COG number was different for both proteins however. The two genes are therefore in the same protein family and protein domain pointing towards an orthologous relationship. The sequences were then aligned using NCBI BLAST which did not support orthologs but the sequences could have evolved from each other but kept the same function over time. Before the pairwise alignment the start codon was made to be in the correct place. The start codon was found to be GTG instead of ATG which is different from the usual start codon. Next the location and structure was compared for both GOIs. The TMHMM data and PRED data both support that the genes are not in the membrane. The PSORT-B data states that the *M. ruber* gene is in the cytoplasm but it is unsure for *E. coli*. Judging from the first two tests on, TMHMM and PRED, the *E. coli* gene is also most likely found in the cytoplasm. Once the location was determined the genes were tested to see if they were in an operon. From the data on IMG the genes were flanked by the same genes which were both part of the same function of CRISPR CASCADE indicating an operon relationship. Mrub_3018 was also compared to other bacteria that were similar to it and the GOI was flanked by the same genes with the same functions in

their respective CRISPR CASCADE, confirming the operon structure. The final test was to see if Mrub_3018 had a paralog. Mrub_3018 did not have a paralog because the only sequence it could be aligned with had a low bit score and high E-value. B2759 however did have a paralog from the same protein family in *E. coli* on NCBI blast with a low E-value and high bit score.

Each bioinformatics tool essentially showed that the two genes produce the same product in CasB, are in the same protein family and domain as well as are in the same place in the protein. The results then point to the fact that Mrub_3018 and B2759 have an orthologous relationship

Citations

A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer.
Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.
Journal of Molecular Biology, 305(3):567-580, January 2001.
(PDF, 959503 bytes)

Bagos PG, Liakopoulos TD, Spyropoulos IC and Hamodrakas SJ.
[PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins.](http://bioinformatics.biol.uoa.gr/PRED-TMBB/)
<http://bioinformatics.biol.uoa.gr/PRED-TMBB/input.jsp>

[Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. \(2008\). GenBank. Nucleic Acids Research, 36\(suppl 1\), D30. doi:10.1093/nar/gkm929](https://doi.org/10.1093/nar/gkm929)

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/>.

Brault, A. C., Savage, H. M., Duggal, N. K., Eisen, R. J., & Staples, J. E. (2012). Viruses.4(10), 2291-2311. doi:10.3390/v10090498

Brian J Tindall¹ , Johannes Sikorski¹ , Susan Lucas² , Eugene Goltsman² , Alex Copeland² , Tijana Glavina Del Rio² , Matt Nolan² , Hope Tice² , Jan-Fang Cheng² , Cliff Han^{2,3} , Sam Pitluck² , Konstantinos Liolios² , Natalia Ivanova² , Konstantinos Mavromatis² , Galina Ovchinnikova² , Amrita Pati² , Regine Föhnrich¹ , Lynne Goodwin^{2,3} , Amy Chen⁴ , Krishna Palaniappan⁴ , Miriam Land^{2,5} , Loren Hauser^{2,5} , Yun-Juan Chang^{2,5} , Cynthia D. Jeffries^{2,5} , Manfred Rohde⁶ , Markus Göker¹ , Tanja Woyke² , James Bristow² , Jonathan A. Eisen^{2,7} , Victor Markowitz⁴ , Philip Hugenholtz² , Nikos C. Kyrpides² , Hans-Peter Klenk¹ , and Alla Lapidus^{2*} . (2010). Complete genome sequence of *meiothermus ruber* type strain; Standards in Genomic Sciences,

Finn, R.D., Coghill, 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/>

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 2008; 26:541-547. PubMed doi:10.1038/nbt1360

Fuguo Jiang¹ and Jennifer A. Doudna¹. The structural biology of CRISPR-cas systems. *Curr Opin Struct Biol*, , 100-111.

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.

Jackson, R. N., Golden, S. M., Erp, P. B., Carter, J., Westra, E. R., Brouns, S. J. J., . . . Wiedenheft, B. (2014). Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science*, 345(6203), 1473-1479. doi:10.1126/science.1256328

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).

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.

Loginova LG, Egorova LA. Obligate thermophilic bacterium *Thermus ruber* in hot springs of Kamchatka. *Mikrobiologiya* 1975; 44:661-665.

Lori scott, Personal communication

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>

Nam, K. H., Huang, Q., & Ke, A. (2012). Nucleic acid binding surface and dimer interface revealed by CRISPR-associated CasB protein structures. *FEBS Letters*, 586(22), 3956-3961. doi:10.1016/j.febslet.2012.09.041

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

Redding, S., Sternberg, S., Marshall, M., Gibb, B., Bhat, P., Guegler, C., . . . Greene, E. (2015). Surveillance and processing of foreign DNA by the escherichia coli CRISPR-cas system doi://doi.org/10.1016/j.cell.2015.10.003

Twiss, E., Coros, A. M., Tavakoli, N. P., & Derbyshire, K. M. (2005). Transposition is modulated by a diverse set of host factors in escherichia coli and is stimulated by nutritional stress. *Molecular Microbiology*, 57(6), 1593-1607. doi:10.1111/j.1365-2958.2005.04794.x

Westra, E. R., van Erp, Paul B. G., Künne, T., Wong, S. P., Staals, R. H. J., Seegers, C. L. C., . . . van der Oost, J. (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by CASCADE and Cas3. *Molecular Cell*, 46(5), 595-605. doi:10.1016/j.molcel.2012.03.018

Westra, E., van Erp, P. G., Künne, T., Wong, S., Staals, R. J., Seegers, C. C., . . . van der Oost, J. (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by CASCADE and Cas3. *Molecular Cell*, 46(5), 595-605. doi:10.1016/j.molcel.2012.03.018

Wright, A. V., Nuñez, J. K., & Doudna, J. A. (2016). Biology and applications of CRISPR systems: Harnessing nature's toolbox for genome engineering. *Cell*, 164(1-2), 29. doi:10.1016/j.cell.2015.12.035

Zhao, H., Sheng, G., Wang, J., Wang, M., Bunkoczi, G., Gong, W., . . . Wang, Y. (2014). Crystal structure of the RNA-guided immune surveillance CASCADE complex in escherichia coli. *Nature*, 515(7525), 147-150. doi:10.1038/nature13733