View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Augustana College, Illinois: Augustana Digital Commons

Augustana College
Augustana Digital Commons

Meiothermus ruber Genome Analysis Project | Biology

Winter 2-13-2019

Mrub_3015 is orthologous to the b2757 gene found in *Escherichia coli* coding for casD

Ramona Collins
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

Collins, Ramona and Scott, Dr. Lori. "Mrub_3015 is orthologous to the b2757 gene found in *Escherichia coli* coding for casD" (2019). *Meiothermus ruber Genome Analysis Project.*
https://digitalcommons.augustana.edu/biolmruber/49

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.

Mrub_3015 is orthologous to the b2757 gene found in *Escherichia coli* coding for *casD.*

Ramona A. Collins
Dr. Lori R. Scott Laboratory
Biology Department, Augustana College
639 38th Street, Rock Island, IL 61201

Introduction

The *Meiothermus ruber* (*M. ruber*) genome analysis project is to aid in the understanding of less studied bacterial phyla with the collaboration of the U.S Joint Genome Institute and the Leibniz-Institut DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH) to form the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project. Through a better understanding of novel biosynthetic pathways and processes we can gain insights into the great diversity within microbial cells and likely discover genes not present in the few well-studied organisms (Scott, 2018). *Meiothermus ruber* is a non-pathogenic bacteria chosen for this study because of its unstudied potential and the likely discovery of novel genes and processes.

Meiothermus ruber is a Gram negative thermophilic rod-shaped bacteria thriving at temperatures ranging from 35 ◦ C to 70 ◦ C and producing a red pigmentation (Tindall *et al.*, 2010). It has been found in a variety of places around the world including, but not limited to, thermal hot springs, spent nuclear fuel storage systems, hot fermenters fed with wastewater, and throughout paper mills as a biofilm (Ekman, *et al.*, 2007). *M. ruber* was first described as a member of the genus *Thermus* by Loginova and Egorova in 1975 and received its current name in 1996 when it was transferred to the newer genus of *Meiothermus* by Nobre *et al.* (Nobre *et al.*, 1996). Being non-pathogenic and not hindering economic growth in any major way *M. ruber* is a lesser studied organism compared to *E. coli* and *Salmonella* whose genomes have been completely analyzed. Therefore, *M. ruber* may contain biosynthetic pathways or alternate genes not found in largely studied organisms.

To better understand the biological interactions and functions of a novel organism, a well-studied model organism found to share similar functionalities with the experimental organism may be employed as a reference point. *Escherichia coli* (*E. coli*) has been extensively studied and has a database called Ecocyc detailing its biochemical processes, structures, and genome (Keseler *et al.,* 2012). A high-profile process originally discovered in *E. coli*, but now known to exist in less than half of bacteria and Archaea, are CRISPR-*Cas* systems, which give the cell adaptive and hereditary immunity to invading DNA (Darmon and Leach, 2014).

CRISPR-Cas immune response is carried out in three steps: adaptation, crRNA expression, and interference (Darmon and Leach, 2014). Adaptation begins with the cell recognizing foreign genomic DNA and assimilating a portion of the invading DNA into the CRISPR system. This is done with the help of cas proteins targeting a specific area of the

invading DNA, called the proto-spacer which includes the proto-spacer adjacent motif (PAM) that is recognized by *cas1* and begins the assembly of the CRISPR-cas system (Wright *et al.*, 2016). The addition of a new sequence to the CRISPR array, called spacers, happens on the 5' end directly followed by the duplication of a new repeat sequence; even though, sometimes more than one proto-spacer can be added from the same foreign DNA, each will be separated by a spacer (Darmon and Leach, 2014). As each new spacer is added onto the 5' end of the CRISPR array, an immunological history can be gleaned from their organization from newest to oldest. The expression of crRNA happens with the reinfecting of a host cell or the heredity history of a cell, triggered by the existence of proto-spacers on the invading DNA. The entire CRISPR array is transcribed as pre-crRNA, also known as premature CRISPR RNA, which is then processed by the system's cas proteins and/or cellular ribonucleases and the third step interference occurs either via silencing of the proto-spacer or the degradation of foreign DNA (Darmon and Leac, 2016).

The CRISPR-*Cas* system is usually arranged in the genome as an operon, a series of coordinately expressed genes coding for enzymes or proteins of consecutive steps in a process (Darmon and Leach, 2014 and Nuñez *et al.,* 2012). Clustered regularly interspaced short palindromic repeats, or CRISPR, is integral to providing defense against a genetic intrusion into the cell via a phage (Darmon and Leach, 2014). The CRISPR array is made up of a leader sequence rich in the bases Adenine and Thymine, followed by identical repeats separated by spacers. The spacers are identical to a short segment of foreign DNA from previous infections. A full CRISPR-*Cas* system also includes a set of CRISPR-associated (*cas*) genes, which produce proteins involved in the cell's immune response *(i.e. nuclease, helicase, polymerase, and polynucleotide-binding proteins)* (Darmon and Leach, 2014; Horvath and Barrangou, 2010).

Individual CRISPR-Cas systems, of which there are multiple types, are commonly arranged as an operon. Figure 1 shows the organization of the *E. coli* CRISPR-*cas* operon with its various *cas* genes labeled, taken from the website at https://biocyc.org/gene?orgid=ECOLI&id=G7427#tab=TU. The combination of various *cas* genes determines the functionality of the CRISPR/cas system; therefore, the number of *cas* genes and the nature of the proteins synthesized varies (Darmon and Leach, 2014). Despite the variability of CRISPR/*cas* systems there are certain combinations of *cas* genes, one of which is usually a signature gene, that repeatedly occur, which has allowed for a classification of types.

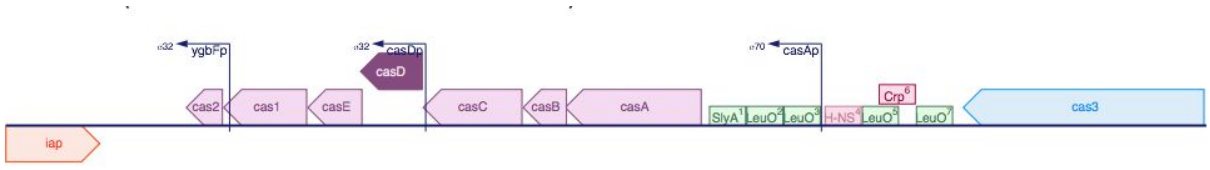


Figure 1. CRISPR operon of *E. coli* showing *cas* genes with the target *casD* (locus tag 2757) gene darkened. Image taken from the website at https://biocyc.org/gene?orgid=ECOLI&id=G7427#tab=TU

To date, CRISPR/*cas* systems can be classified into six types labeled with their respective roman numerals I-VI, though I-III are the most studied. *E. coli* CRISPR/*cas* is classified as type I-E, with the signature gene *cas3* (labeled in Figure 1 in light blue). For this project, we chose the gene *casD*, which is responsible for the maturation of pre-crRNA to crRNA (Jiang, 2015; Wright *et al.,* 2016). *CasD* has multiple synonyms frequently used in literature as the gene was found at various times by various researchers. On the Ecocyc site (Keseler *et al*., 2012) the synonyms of *casD* are listed as *ygcl*, *cas5e*, and *casD*. However, for this paper the *E. coli* locus tag of b2757 will be used to differentiate it from the putative *M. ruber* ortholog (mrub_3015).

Taking a closer look at the I-E CRISPR/*cas* system of *E. coli* and the role of *casD* therein, figure 2 shows the crystal structure of CRISPR/*cas* and another look at the CRISPR operon of *E. coli* to give an in-depth look at the structure of the *cas* proteins with the mature crRNA, forming a Cascade complex. Cascade is the CRISPR-associated complex for antiviral defense which binds to and degrades the invading DNA after a mature crRNA is added to the CRISPR/*cas* system (Jiang, 2015). *CasD* is integral in the CRISPR-cas system for the formation of an R-loop, consisting of six B-hairpins from casD to cas7 interrupting the crRNA-DNA from base pairing with the target phage DNA (The Jiang and Doudna, 2015; UniProt Consortium, 2018). This prevents the CRISPR-cas system from going through conformational changes which would twist and separate the components of the CRISPR-cas system, not allowing for it to be immediately reused in the cell for further immune response. As the immune response of bacterial cells is important, not only for future work in pathology and immunology, but in the furthering of scientific knowledge for the Tree of Life, *Meiothermus ruber* was used to better understand gene organization that may not be present in model systems as well as novel genes.

In this study, through the use of various bioinformatics tools, we will determine whether or not mrub_3015 and *E. coli* b2757, both coding for *casD* protein, are orthologous to each other.

Methods

Data was collected on the model organism, *E. coli* K12 MG1655 by perusing the Ecocyc (Keseler *et al.*, 2012) site specific to the *CasD*, *b2757*, gene to obtain background information on the known structure of the CRISPR-*Cas* system. From Ecocyc the FASTA-formatted amino acid sequence, location of the protein, and biosynthesis pathway were collected (Keseler *et al.*, 2012). As the gene is part of an operon we also obtained the structure of the operon from Ecocyc, as shown by Figure 1 in the introduction, obtained from https://biocyc.org/gene?orgid=ECOLI&id=G7427#tab=TU.

Next, we used KEGG (Kanehisa *et al.*, 2019), to predict if the genome of strain *M. ruber* DSM1279 encoded an ortholog of *casD*. From this analysis, KEGG identified mrub_3013 as a likely ortholog within the Type I-E CRISPR-*Cas* system. Before using the amino acid of mrub_3013 as a query for further bioinformatics tools, we confirmed that the start codon had been correctly called using two programs. First, the DOE Joint Genome Institute's platform called IMG/M (Markowitz) was used to analyze the 5' upstream region of Mrub_3013 for potential Shine-Dalgarno sequence(s), and identify nearby in-frame alternative start codons. Second, the NCBI protein BLAST (pBLAST) was used to create a multi-sequence alignment of the amino termini of related proteins to determine if they all aligned at the same "M" methionine amino acid (Madden, 2013).

Once the start codon was confirmed, another pBLAST was performed to produce a pairwise alignment between *E. coli* and the *M. ruber casD* sequences to determine if there existed a strong alignment between the two, which would support the hypothesis that the two shared a common ancestor and were orthologous.

A collection of bioinformatics tools assembled for this project by Dr. Scott were used to compare cellular localization and functional features of mrub_3013 and *E. coli* CasD. Three programs were used to determine the cellular location of the two proteins. TMHMM predicts the existence of transmembrane alpha helices that might be responsible for embedding a protein into the membrane. PRED predicts the existence of beta-barrels embedded in the outer membranes (Bagos**)**. PSORT-B predicts a protein's cellular location within Gram-negative bacteria, which have five primary localization sites—the cytoplasm, the inner membrane, the periplasm, the outer membrane and the extracellular space. determines the location probability of the protein based on how the sequence properties would interact with various local in the cell (via hydrophobic and hydrophilic regions) (Yu N.Y. *et al.*, 2010).

Next, structure-based evidence was gathered for *E. coli* CasD and mrub_3013 using three databases organized into consensus sequences for specified functional units. The Conserved Domain Database CDD (Markowitz VM *et al.,* 2012) possesses a collection of domain consensus sequences called Clusters of Orthologous Genes (COGs) to which a query can be aligned. TIGRfam (Haft DH *et al.*, 2001) contains a collection of consensus sequences derived from protein families. Pfam is a collection of domain, family and clan consensus sequences (Finn, R.D.). Since amino acid sequence determine the structure and function of the protein, high variability would lead to varying protein structures between species, resulting in different functions. The bioinformatic tool PDB, the Protein DataBase, is a collection of protein which have been crystalized and 3D models created. From this database, we pulled proteins were the same function regardless of species (Berman). The structure of the protein sequences is important, as similar structure is a strong indicator of similar function within the respective organism.

IMG/M Gene Neighborhood program was used to analyze the flanking regions of mrub_3013 and *E. coli casD* to determine if these genes are components of a CRISPR-Cas operon and their order on the chromosome. This was done to note if the genes flanking *casD* in either system were the same, as variations in operon organization can occur over time and between species, leading to a variety of operon organizations which are divided into types I through VI today (Wright *et al.*, 2016). Finally, a pBLAST was performed by using mrub_3013 as a query against the *M. ruber* genome to determine if paralogs, or repeated similar genes, of *CasD* existed within the same genome.

Results

Table 1 summarizes the KEGG output for the genes involved in the CRISPR-Cas system for *E. coli* and *M. ruber*. The first column is the generic KEGG assigned locus tags of the *Cas* in both organisms as gene product names vary between studies. The second column defines the common *cas* protein names. The third column is the locus tags of *E. coli* K12 for the CRISPR-Cas genes found in its genome. The fourth column are the locus tags of *M. ruber*, which contains multiple paralogs of cas1 and cas2 denoted by multiple locus tags for each gene. This provides an organized reference to the various locus tags and names that may be used in this study for both *E. coli* and *M. ruber*.

Table 1. Comparison of *E. coli* and *M. ruber* locus tags for a Type I-E CRISPR-Cas system as predicted by KEGG databases

Locus Tag(s)	Common *cas* name	*E. coli K12* MG1655	*M. ruber DSM1279*
K15342	cas1	b2755	Mrub_0224, Mrub_1477, Mrub_3013
K09951	cas2	b2754	Mrub_1476, Mrub_0225, Mrub_3012
K07464	cas3	b2761	Mrub_3020
K19123	casA	b2760	Mrub_3019
K19046	casB	b2759	Mrub_3018
K19124	casC	b2758	Mrub_3016
K19125	casD	b2757	Mrub_3015

K19126	casE	b2756	Mrub_3014

Table 2 summarizes the results of a variety of bioinformatics tools that were used to compare the *E. coli* CasD to Mrub_3015. The first row of data shows the results of the initial pBLAST analysis comparing E. coli CasD to Mrub_3013. Due to the varying length of the proteins, the percent identity and bit score is not as meaningful as the E-value. The E-value of the pBLAST result is 5e-13, which shows the alignment of the two sequences was not due to chance and shows they share a significant amount of amino acids. This provides evidence the two genes are evolutionarily related by sharing a common ancestor.

In the second row of data, Ecocyc and TIGRfam site searches came up with slightly different gene product names (type I-E CRISPR system Cascade subunit CasD and CRISPR-associated protein Cas5 family, respectively). However, upon additional digging between various sites, CasD and Cas5 are synonyms for one another. The Ecocyc, TIGRfam, and synonymous *casD* and *cas5* join together to show *E. coli* b2757 and mrub_3015 have similar structures and functionality as they produce similar gene products.

The bioinformatics tools used to analyze cellular location (TMHMM and PRED) suggested both proteins are located in the cytoplasm of the cell. Neither protein possesses a transmembrane region that would allow the protein to pass through or embed in a membrane. pSORT-b produced an “indeterminate” output for the E. coli CasD, but it provided localization scores for mrub_3013 that coincide with the previous findings of TMHMM and PRED, the localization of the protein was predicted to be in the cytoplasm. From the TMHMM and PRED data, it can be concluded both proteins have the same cellular location in the cytoplasm, adding to the evidence suggesting the two genes are orthologs.

When the COG numbers were collected from the CDD site there were no COG hits for either protein sequence. However, *E. coli* b2757 did have a TIGRfam hit (TIGR01868) with an E-value of 5.63e-99 indicating it was significant but was not of much use in comparison to the *M. ruber* sequence. The only hit collected for Mrub_3015 from the CDD was the superfamily cl21479 with an E-value of 6.27e-57. The significant piece of data gathered from the CDD was the similar superfamily for both protein sequences as the TIGRfam number was situated under the same superfamily even though no E-value could be obtained for *E. coli*. Then moving to the TIGRfam database: the TIGRfam numbers collected from the protein sequences were the same for both organisms (TIGR01868 and TIGR02593) which were called casD_Cas5e: CRISPR system CASCADE complex p and CRISPR_cas5: CRISPR-associated protein Cas5 respectively. Building off of the TIGRfam data, Pfam confirmed both proteins contained the same three domains: Cas_Cas5d (PF09704), HD_6 domain (PF18019), and DEAD box helicase domain (PF00270). Lastly, the protein database generated a protein name of CRISPR-associated protein (Cas_Cas5) for both protein sequences, although the protein number for *E. coli* was 4QYZ with

an extremely small E-value of 5.64501e-123 while the protein number for mrub_3015 was 5U07 with an E-value of 4.27097e-21. The three-dimensional structure of both proteins, taken from PDB, was seemingly identical, further confirming the structural similarity of *E. coli* b2757 and Mrub_3015.

Table 2. Comparison of *E. coli* b2757 and Mrub_3015 using various bioinformatics tools

Bioinformatics Tools Used	*E.coli (CasD gene) b2757*	*M.ruber (CasD gene) Mrub_3015*
pBLAST b2757 against Mrub_3015	Identities: 31% E-value: 5e-13 Bit Score: 52.4	
GenBank and Ecocyc product name	type I-E CRISPR system Cascade subunit CasD	CRISPR-associated protein Cas5 family
Cellular Localization	Cytoplasm of the cell	
CDD Data (COG hit)	TIGR01868	cl21479
	E-value: 5.63e-99 Bit Score: 286.31	E-value: 6.27e-52 Bit Score: 167.28
TIGRfam - protein family	TIGR01868	
	casD_Cas5e: CRISPR system CASCADE complex p	
	E-value: 9.4e-129 Bit Score: 437.3	E-value: 1e-58 Bit Score: 204.6
	TIGR02593	
	CRISPR_cas5: CRISPR-associated protein Cas5	
	E-value: 9.4e-16 Bit Score: 61.9	E-value: 2.5e-14 Bit Score: 57.1
Pfam - protein family	PF09704 (Cas_Cas5d)	
	PF18019 (HD_6 domain)	
	PF00270 (DEAD box helicase)	
Protein Database (PDB)	4QYZ (CRISPR-associated protein [Cas_Cas5])	5U07 (CRISPR-associated protein [Cas_Cas5])
	E-value: 5.64501e-123 Bit Score: 437.958	E-value: 4.27097e-21 Bit Score: 99.7525

Figure 6 shows the 3-dimensional structure of b2757 and mrub_3015, respectively, taken from the Protein Database (PDB) based off of the best E-value and most relevant entry. The structure for *E. coli* has the

database sequence of 4QYZ, named CRISPR-associated protein [Cas_Cas5] while the structure for the *M. ruber casD* protein has the database sequence of 5U07, named CRISPR-associated protein [Cas_Cas5].

E. coli 3D structure

M. ruber 3D structure

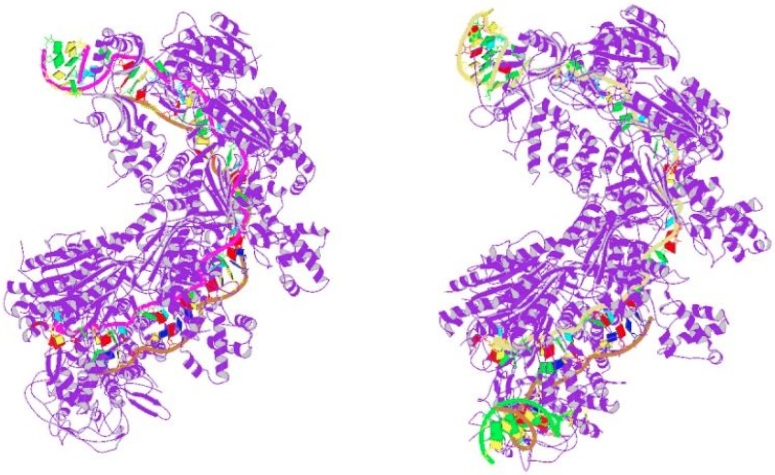


Figure 6. 3-dimensional structures from the Protein Database showing similar structures between *E. coli* b2757 and mrub_3015 proteins. Data was collected from https://www.rcsb.org via sequence searches, with the best E-value and relevant structures chosen.

Figure 7 shows the results of a protein BLAST of *E. coli* b2757 against Mrub_3015. According to the BLAST, 31% of the amino acids were identical between both sequences. Of greater importance is the E-value from the BLAST which is given as 5e-13. The E-value is well-below the cut-off of 0.001, which shows that the two sequences did not align by chance. As the sequence of amino acids determines the folding and structure of the protein, a similar sequence alignment is evidence in support of the two genes being orthologs and sharing the same function within the cell.

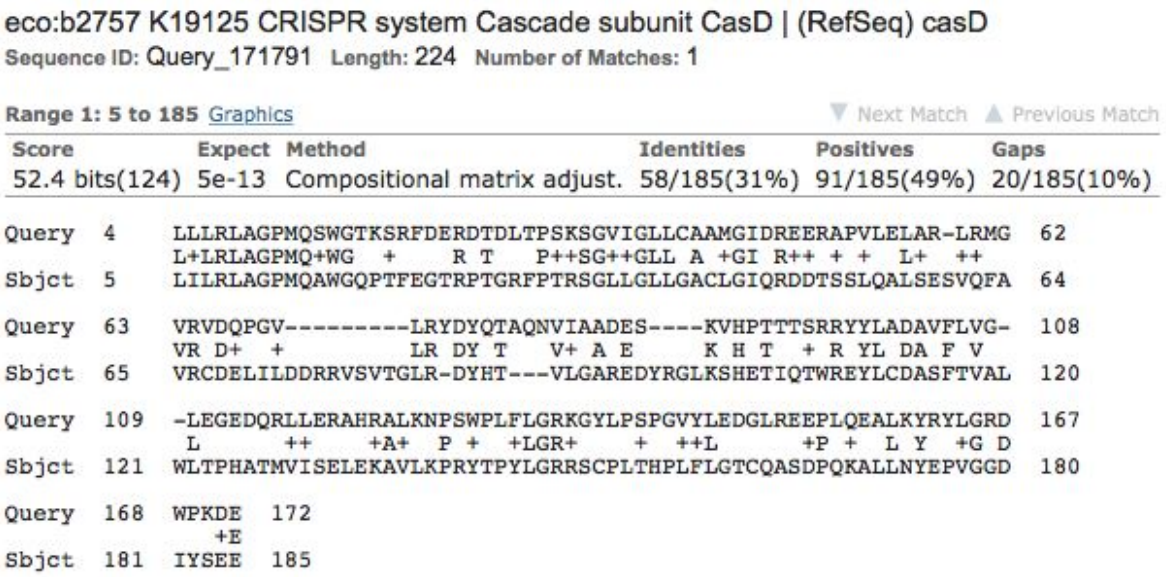


Figure 7. Protein BLAST showing a similar protein sequence with *E. coli* b2757 as the query sequence and Mrub_3015 as the subject sequence. Analysis was performed using the NCBI BLAST bioinformatics tool at https://blast.ncbi.nlm.nih.gov/Blast.cgi

Figure 8a shows the sequence viewer for alternate ORFs of the *M. ruber* (mrub_3015) protein sequence with possible start codons highlighted in yellow and possible Shine-Dalgarno sequences highlighted in light blue. In order for the correct start codon in the sequence to be categorized, there must be a viable Shine-Dalgarno sequence around 8bp upstream of plausible start codon. As there are no Shine-Dalgarno sequences in which ribosomes can bind and facilitate protein synthesis except the given start codon, it can be assumed there is only one possible start codon. Figure 8b then shows a multiple sequence alignment of 12 different *Meiothermus* species with *M. ruber* being the first sequence. By comparing multiple sequences it can be determined if the correct start codon has been recognized. Based on the repetitive alignment of the first Met it can be assumed the correct start codon have been recognized.

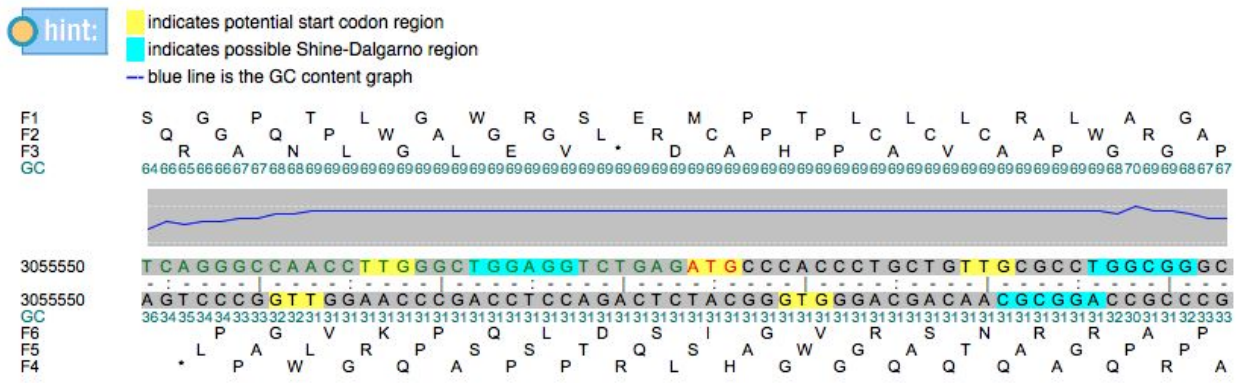
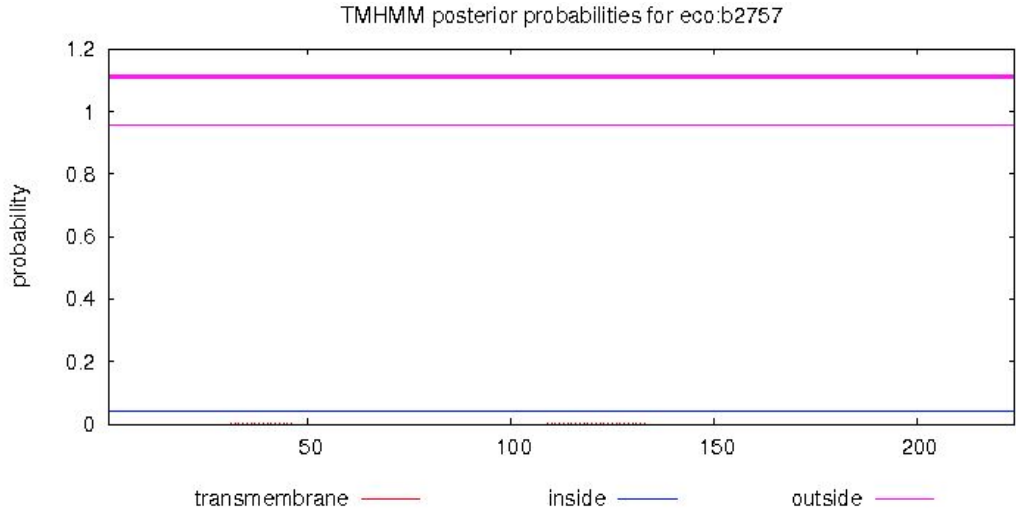


Figure 8a. IMG sequence viewer for alternate ORF of *M. ruber* protein sequence showing alternative start codons and Shine-Dalgarno sequences. The claimed start codon is colored red with a yellow highlight. Analysis was performed using the IMG gene details page at https://img.jgi.doe.gov/cgi-bin/m/main.cgi

Figure 8b. Protein BLAST multiple sequence alignment of *Meiothermus* species starting from the top: *ruber, taiwanesis, silvanus, luteus, timidus, cerbereus, roseus, rufus, chliarophilus, terrae, hypogaeus, and sp. QL-1*. Analysis was performed using NCBI BLAST bioinformatics tool at https://blast.ncbi.nlm.nih.gov/Blast.cgi

Images in Figure 9 show the TMHMM graphs for *E. coli* b2757 and Mrub_3015 detailing the hypothesized location of the protein in the cell, specifically the presence of transmembrane alpha helices. Peaks signifying the presence of alpha helices are not found anywhere, demonstrating both casD genes in *E. coli* and *M. ruber* are localized in the cytoplasm rather than in the membrane or secreted by the cell.

Panel A



Panel B

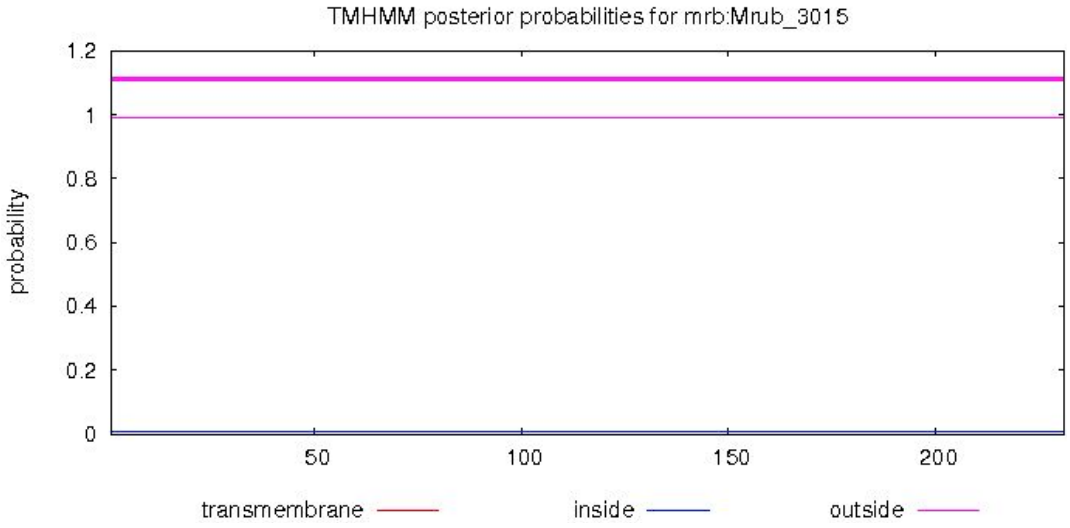
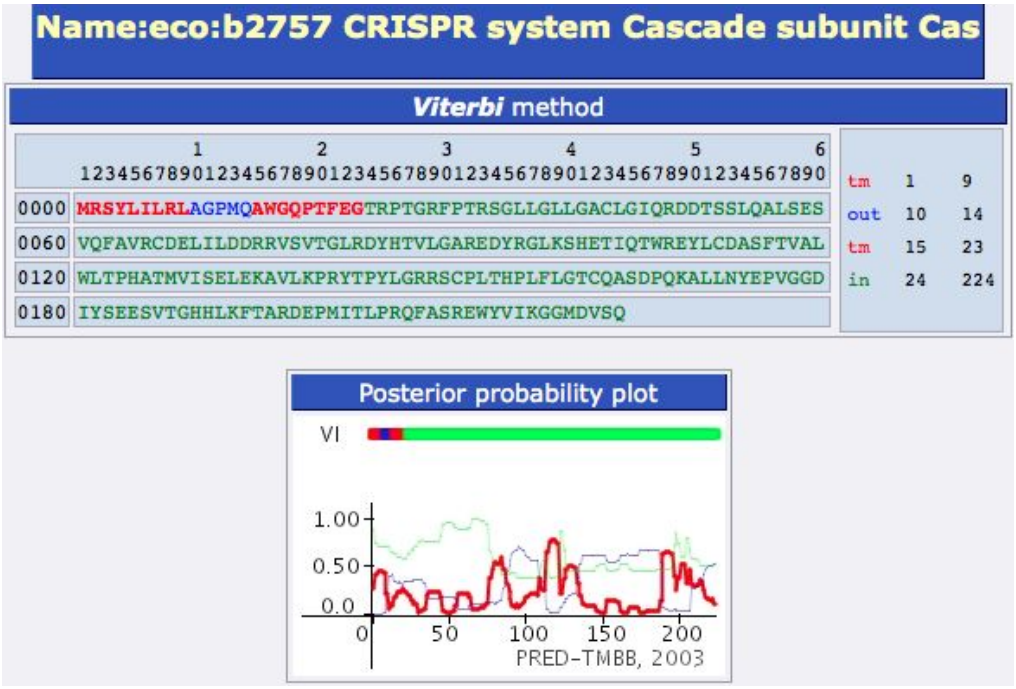


Figure 9. TMHMM graphs lacking peaks signifying alpha helices in both *E. coli* b2757 and Mrub_3015. A cytoplasmic protein location is predicted for both protein sequences. Panel A shows the TMHMM data for *E. coli* b2757 while panel B shows the TMHMM data for Mrub_3015. http://www.cbs.dtu.dk/services/TMHMM/ was used to create these two charts.

The graphs and pictures in Figure 10 displays the PRED predictions on the presence of transmembrane beta-barrels in *E. coli* b2757 and mrub_3015. Peaks signifying the presence of beta-barrels are not seen in either protein sequence, further signifying the locality of the proteins in the cytoplasm of the cell.

Panel A



Panel B

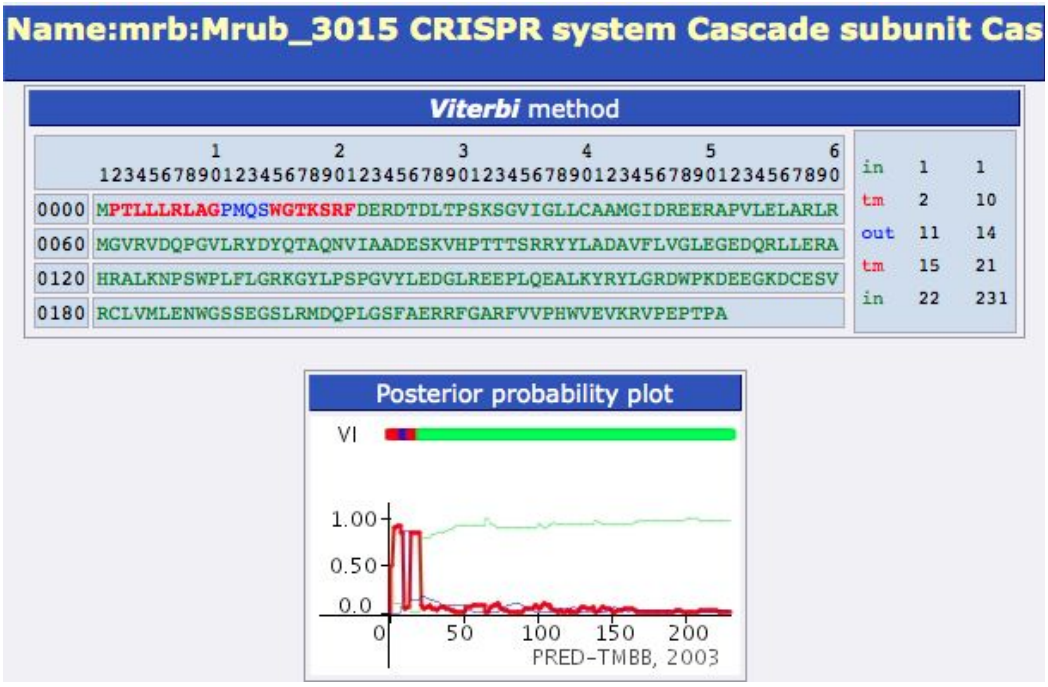
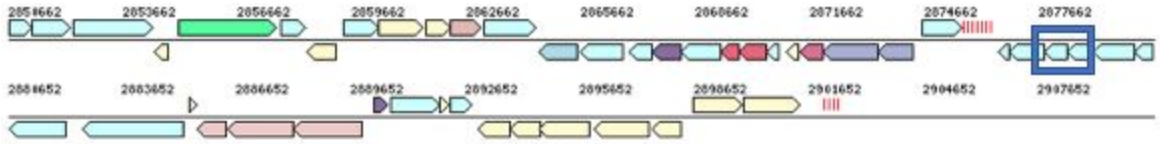


Figure 10. PRED data showing the lack of peaks denoting the presence of beta-barrels in the protein sequences of both proteins, hypothesizing a cytoplasmic location for *E. coli* b2757 and

Mrub_3015. Panel A shows the PRED data for *E. coli* b2757 while panel B shows the PRED data for Mrub_3015. http://bioinformatics.biol.uoa.gr/PRED-TMBB/input.jsp was used to create these two charts.

Figure 11 illustrates the IMG/M Gene Neighborhood Chromosome map, colored according to KEGG demonstrating the genes involved in the same pathway. As shown by the similarly colored genes surrounding them, both *b2757* and *mrub_3015* are part of an operon. Both are the same shade of blue, which is indicative of a gene involved in defense mechanisms, and are noted by the box.The organization of any set of genes into an operon decreases the genetic load on the organism and the chance of mutations that may lead to dysfunctionality (Nuñez *et al.,* 2012). Organizing multiple genes close together also decreases the probability of point mutations and rearrangements of the genome, decreasing genome instability; however, even between evolutionarily related species there can be variation to the organization of the CRISPR-cas system (Darmon and Leach, 216). Therefore the Chromosome Viewer map is additional evidence *E. coli* b2757 and Mrub_3015 are orthologs.

Panel A



Panel B

Figure 11. IMG/M Gene Neighborhood Chromosome map showing *E. coli* b2757 and Mrub_3015 are part of an operon. Chromosome Viewer maps were colored according to KEGG, with light blue denoting functionality as part of defense mechanisms. Panel A shows the Chromosome Viewer of *E. coli* b2757 while panel B shows the Chromosome Viewer of Mrub_3015. Images were taken from https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=GeneDetail&page=geneDetail&gene_oid=646314719 and https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=GeneDetail&page=geneDetail&gene_oid=2555895482.

Discussion

By gaining a deeper understanding about the various genes in *M. ruber* we can gain insights into the diversification of microbes and discover genes not present in well studied organisms. The results obtained from this study revealed *E. coli* b2757 and Mrub_3015 are orthologous genes. This means the *E. coli* and *M. ruber* may be related through a common ancestry and share similar protein sequences for the CasD gene in the CRISPR-Cas system as seen by the BLAST analysis. The similarity between the sequences, as well as the structural similarities seen via the bioinformatics tools TMHMM, PRED, and PDB point to a cellular location of the cytoplasm within the organism. In addition to the structural characteristics and placement of the protein, COG, Pfam and TIGRfam expanded on the similarity between *b2757* and *mrub_3015* by matching the protein sequences of both to the same superfamily (cl21479), to similar proteins and domains (Cas_Cas5d, HD_6, and DEAD). There were several other bioinformatics tools and sites used for this project which pointed to the orthologous relationship between Mrub_3015 and *E. coli* b2757. Although some information could not be obtained for *E. coli* (cellular localization from pSORT-B) the evidence against the orthologous relationship between the two proteins can be due to the species differences of *E. coli* and *M. ruber* and the lack of information and study being done on lesser known and less global/human health related bacteria. Based on the gathered data from the various bioinformatics tools, we can conclude Mrub_3015 is orthologous to *E. coli* b2757 in the CRISPR-Cas system.

Works 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.

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/.

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

Ekman J, Kosonen M, Jokela S, Kolari M, Korhonen P, Salkinoja-Salonen M. 2007. Detection and quantitation of colored deposit-forming *Meiothermus* spp. in paper industry processes and end products. Society for Industrial Microbiology (34):203-11.

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.

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

Jiang and Doudna. 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).

Keseler I.M., Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, Fulcher C, Huerta A,M., Kothari A, Krummenacker M, et al. 2012. EcoCyc: Fusing model organism databases with systems biology. Nucleic Acids Research 41(D1):D612.

Loginova, L. G., Egorova, L. A., Golovacheva, R. S., Seregina, L. M. (1984). Thermus ruber sp. nov., nom. rev. Int J Syst Bacteriol (34):498– 499.

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

Nuñez PA, Romero H, Farber MD, Rocha EPC. 2012. Natural selection for operons depends on genome size. Genome Biology and Evolution 5(11):2242-54.

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

Scott L. 2018. Functional studies comparing *Escherichia coli* proC to a putative ortholog mrub_1345.

The UniProt Consortium. 2018. UniProt: The universal protein knowledgebase. Nucleic Acids Research 46(5):2699.

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

Wright AV, Nuñez JK, Doudna JA. 2016. Biology and application of CRISPR systems: Harnessing nature's toolbox for genome engineering. Cell (164):29-44.