2019

Mrub_3020, a paralog of mrub_1489, is orthologous to *E. coli* casC (locus tag b2761)

Alfred Dei-Ampeh
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_3020, a paralog of mrub_1489, is orthologous to E. coli *casC* (locus tag b2761)?

Alfred Kwabena Dei – Ampeh

Dr. Lori Scott.

Introduction

Strain 21^T (= DSM 1279, ATCC 35948 = VKM B - 1258) is a type strain of the species *Meiothermus ruber* (*M.ruber*). Strain 21^T was initially placed in the *Thermus* genus, but it was later moved to the genus, *Meiothermus* (Nombre *et al*., 1996; Loginova LG and Egorova LA, 1975). The species name "ruber" is a latin epithet, which translates into english as "red". Thus, "ruber" describes the red pigmentation of *M.ruber* (Loginova LG and Egorova LA, 1984; Euzéby JP, 1997). The genus name *Meiothermus* is coined from two greek works: "meion" means "lesser" and "thermus" means "hotter" - that describe the less hot habitat of the bacteria. (Nombre et al., 1996; Euzéby JP, 1997). Eight species in the *Meiothermus* genus were isolated from artificial thermal environment and hot springs found in six different countries (Euzéby JP, 1997; Tindall, B. J. *et al.*, 2010; Nombre *et al.*, 1996).

M. ruber DSM 1279 genome's (GenBank name ASM2442v1) was sequenced, finished, and annotated as part of GEBA (Genomic Encyclopedia of Bacteria and Archaea) project, a collaboration between U.S. Department of Energy Joint Genome Institute and Leibniz - Institut DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH). The goal of the project is to fill the gaps by sequencing along the archaeal branches and bacterial branches of the tree of life. This bacteria strain is one of the many non-pathogenic bacteria species whose genomes were sequenced because they are part of the poorly studied diverse bacterial phyla (Tindall B.J, *et al.*, 2010; Scott *et al.*, n.d; Wu *et al.*,2009). Sequencing *M.ruber*'s genome, together with that of other species of bacteria, would lead to novel gene discoveries, novel biochemical processes, and increase understanding in the underlying processes of evolutionary diversification microbes (Scott *et al.*, n.d.).

Meiothermus ruber is a gram negative, non-motile, rod shaped bacteria with rounded ends (Figure 1). This bacterium is obligate aerobic. *M.rub* grows in normal media that is supplemented with 0.15% (w/w) peptones as a source of nitrogen, 0.05% (w/v) yeast extract, and 0.25% (w/v) carbon sources like D-glucose (Loginova LG and Egorova LA, 1984).

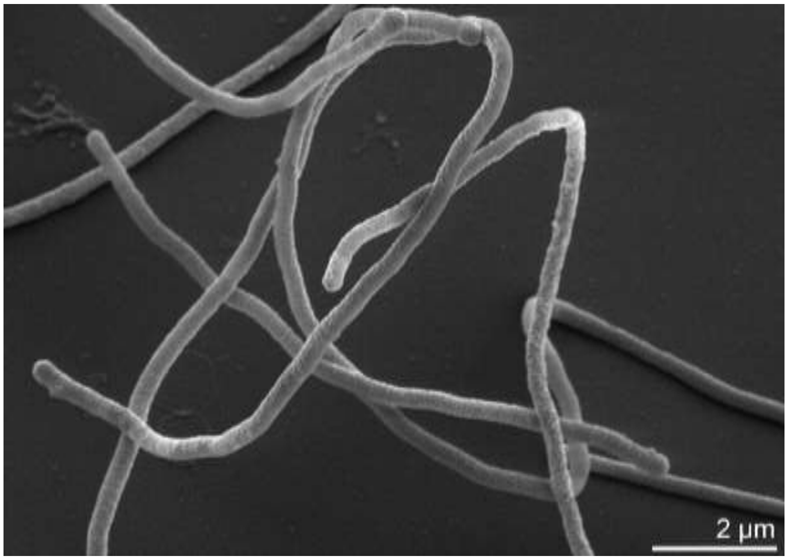


Figure 1. Scanning electron micrograph of rod-shaped M. ruber 21^T.
http://standardsingenomics.org

M.ruber's sequenced genome has some unique properties such as being a 3,097,457 b.p. long chromosome with 3052/3105 genes that encode proteins. 53 of the genes encode RNA, and 38 of the genes are pseudogenes. 71.8% of the protein-encoding genes have been assigned putative functions (Tindall B.J, *et al.*, 2010).

Living organisms have many systems that help them survive. Microbes have several defense mechanisms that enable them to recognize and distinguish many "foreign" DNA from "self" DNA, in addition to getting protection from invasive elements. The Clustered Regularly Interspaced Short-Palindromic Repeats (CRISPR)-Cas system is an adaptive immune system of bacteria and archaea. The acronym CRISPR was coined in 2002, but its structure was discovered in *Escherichia coli* in 1987. CRISPR is a family of DNA repeats that make up 90% of the archaeal genome and 40% of the bacterial genome. Its size varies from 23 - 47 base pairs and 21-72 base pairs (Jansen *et al.*, 2002; Van der oost *et al.*, 2009; Sorek R, Kunin V, Hugenholtz P. 2008, Horvath P and Barrangou R. 2010).

CRISPR provides an acquired immunity for the microbes against viruses and plasmids (Horvath P and Barrangou R. 2010). CRISPR loci is made up of non-adjoining direct repeat separated by spacer (stretches of variable sequences), usually next to cas genes - protein encoding genes that serve as nucleases, helicases, polymerases, and polynucleotide-binding proteins - (Haft DH *et al.*, 2005). The repeat sequences are partially palindromic, and they can form stable secondary structures (Kunin V, Sorek R, Hugenholtz P. 2007). Cas proteins and CRISPR come together to form the CRISPR-Cas system, which uses a range of small CRISPR RNAs (crRNAs) that are transcribed from the CRISPR loci and the cas proteins to detect specific, non-self-DNA sequences and silence them. This defense system targets non-self-DNAs through base pairing between the DNAs and the crRNA guide sequence that result in cas-protein mediated DNA cleavage (Barrangou et al. 2007; Browns, SJ. *et al.*, 2008; Garneau, JE *et al.*, 2010; Marraffini, LA *et al.*, 2008). This functional relationship between cas and CRISPR was obtained by inferring from congruence obtained between sequence of patterns (Figure 2) (Barrangou R *et al.*, 2007).

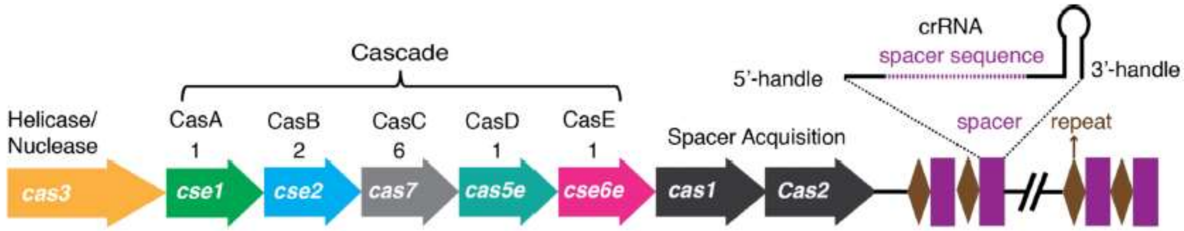


Figure 2. The *E. coli* CRISPR/Cas operon system. A type I subtype E operon system (a sequence of genes whose transcription is regulated in a sequential manner starting with the promoter) shows the genes that encode the proteins that make the cascade (Complex of proteins that assemble with crRNA to form the Cascade complex; targets the invading DNA/RNA sequence). Numbers under Cas A-E identify the number of copies of each protein in a cascade complex that recruits Cas3, which is a helicase & nuclease to catalyze target DNA degradation (Jiang & Doudna, 2015).

The CRISPR-Cas adaptive immune system consists of three stages by which the organism is provided with immunity against invaders. The adaptation or acquisition stage involves identification of the PAM (protospacer adjacent motif) sequence, which lies 2-4 bases upstream of a protospacer region. The protospacer, now called the spacer, is then incorporated into a CRISPR array in the host's genome. This enables the host to keep genetic records of the prior invasion; thus, facilitating future immune response against the same invader. The next stage is the CRISPR-cas expression during which the CRISPR array is transcribed into pre-crRNA, and further processed into matured crRNAs. The last stage is DNA interference. In this step, complexes formed between the matured crRNAs and the cas proteins are used to target viral DNA for degradation; therefore, preventing propagation (Figure 3) (Van der oost *et al.*, 2014; Marraffini LA and Sontheimer EJ, 2010; Wiedenheft B, Sternberg SH, Doudna JA, 2012; Brouns SJ, et al. 2008; Heler R, Marraffini LA, Bikard D., 2014; Mojica FJM *et al.*, 2009).

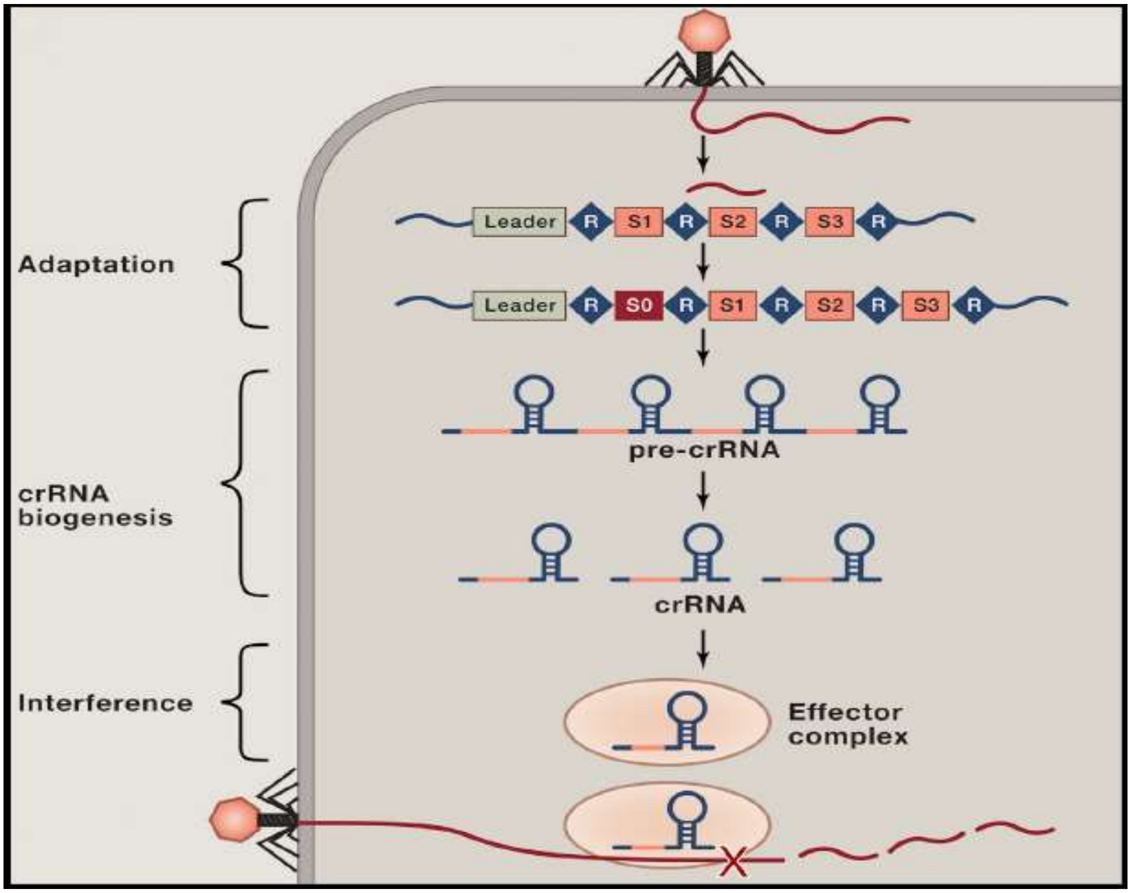


Figure 3. Schematic representation of the CRISPR-Cas immune response. The three major stages begin with adaptation, then commences crRNA biogenesis, and finally destroys the invading DNA during the interference stage. The effector complex acts in the interference stage to spot future foreign nucleic acids and degrade them. Key: "R" (repeats region), "S" (spacer region of the host's sequence), "SO" (selected protospacers from the invading nucleic acids that are inserted in front of the "leader" end of the CRISPR locus); Blue wavy lines (host's nucleic acid sequence); and red wavy lines (invader's nucleic acid sequence) (Wright *et al.*, 2016).

There are three main types of the CRISPR-Cas systems: types I, II, and III. These main types are further divided into 11 sub-types: I-A to I-F, II-A to II-C, and III-A to III-B (Figure 4) (Markarova KS, et al. 2011; Chylinski K, et al. 2014). All these CRISPR-Cas systems work to provide adaptive immunity, but there exist some mechanistic diversities. These diverse mechanisms are, especially, prominent during the CRISPR-Cas expression stage, and the DNA interference stage of an immune response (van der Oost J, et al. 2014). For example, type I CRISPR-Cas system uses Cas 3 proteins to aid in unwinding and cleaving the non-self-DNA when a cascade is formed that binds the complementary DNA target sequence (Westra ER *et al.*, 2012; Sinkunas T *et al.*, 2013). Cas 3 proteins are nuclease helicases, Cas 1 proteins (proposed to be double-stranded DNA endonuclease) and Cas 2 proteins (proposed to be sequence-specific endonuclease that cleave uracil-rich single stranded RNAs) are universal markers of the CRISPR-Cas system (Westra ER *et al.*, 2012; Sinkunas T *et al.*, 2013; Sorek R, Kunin V, Hugenholtz P. 2008; Haft DH *et al.*, 2005; Beloglazova N *et al.*, 2008).

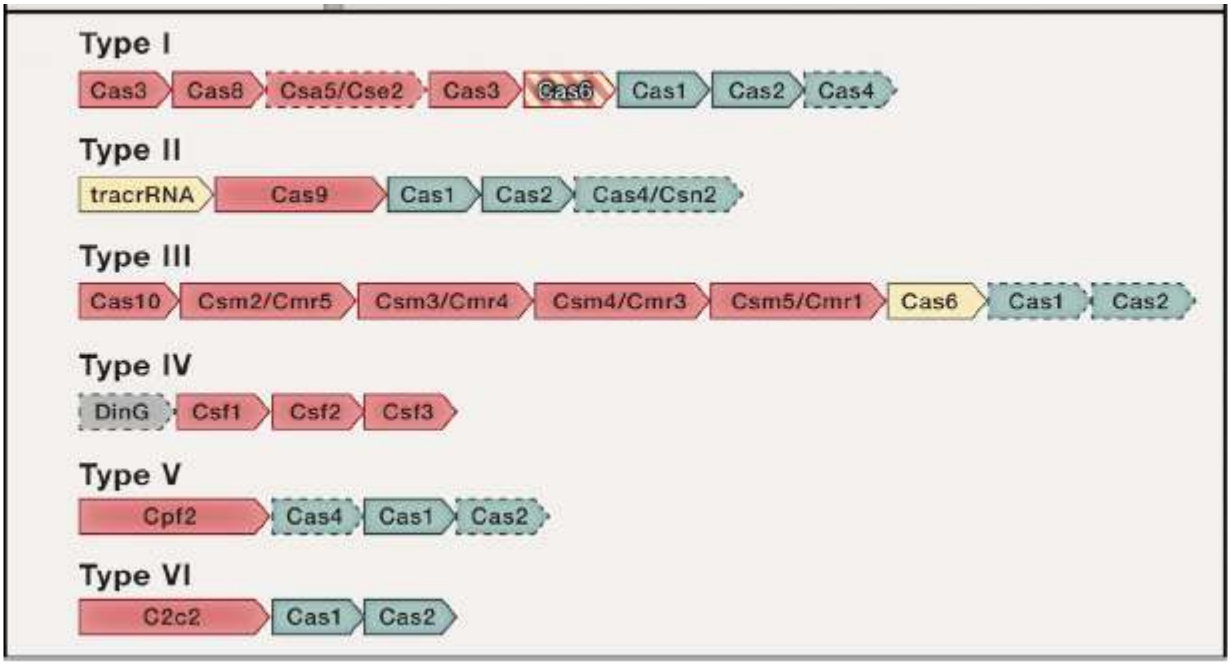


Figure 4. Representation of the components of the different types of the CRISPR-Cas systems. Class 1 (Types IA-E, III, IV): multi-subunit Cascade complex, and Class 2 (Types II, V, VI): single-subunit effector. The names in the colored boxes represent the different gene encoding proteins. Key: Cmr (Chlorophenicol resistance protein); Csm (Cutoff Scanning Matrix, which is a recent protein function prediction and structural classification method); Cas (CRISPR-associated system) (Wright *et al.*, 2016).

The CRISPR-Cas system is an important system to study because of the current scaled up implementation of an engineered version of the TypeII Cas 9 targeting complex for site-specific genome engineering in animals, plants, fungi, and bacteria (Jinek M et al. 2012; Jiang F and Doudna JA. 2015). Although several research works have informed the scientific world on the CRISPR-Cas system of organisms, such as *Escherichia coli*, there are still some unknowns. For instance, the structural mechanism by which PAM recognition triggers Cas3 mediated DNA cleavage is yet to be elucidated (Anders C et al.; 2014). Also, it will be beneficial to probe the reason behind a comparatively smaller percentage of bacteria CRISPR genome, that is 50 % lower than the percentage found in archaea (Horvath P and Barrangou R. 2010).

This journal style research paper seeks to answer this question: Is mrub_3020, a possible paralog of mrub_1489, orthologous to *E. coli* cas3 (locus tag b2761)? *E. coli* K12 MG1655 strain is used as the model organism because extensive published research is available on its CRISPR-Cas system (EcoCyc; Jiang and Doudna, 2015). *E.coli* cas3 gene encodes the CRISPR-associated endonuclease/helicase Cas3, a signature protein of the Type I CRISPR systems It has been shown to support the Cascade complex to provide resistance to the host cell against certain phages. Some genomic regions of these phages are complementary to some elements of the CRISPR repeat (Brouns *et al.*, 2008). CasA recruits Cas3 to the Cascade complex and positions it near the "protospacer adjacent motif" (PAM) (Hochstrasser *et al.*, 2014). Cas3 proteins support the complex by cutting the target viral DNA, unwinding it, and then degrading it through a joint ATP-dependent helicase activity and Mg^{2+}-dependent HD-nuclease activity (Bailey and Mulepati, 2013; Westra *et al.*, 2012).

M. ruber DSM1279 is the test organism whose potential CRISPR-Cas gene (mrub_3020) would be primarily studied. Part of this research would seek to determine if mrub_1489 is a possible paralog (homologous structures or organisms that have their evolution reflects gene duplication) of mrub_3020 (Fitch WS, 1970). This because there could be the possibility that such paralogs would exist for the test organism because previous studies using the genome of the model organism showed that 66% of the putative proteins encoded by the model organism's genome were paralog proteins (Labedan B and Riley M, 1995; 1995; 1997). Additionally, probing the possibility of mrub_3020 having the paralog would give way to appreciating evolutionary processes that may have generated the biochemical and other biological differences between the proteins encoded by the paralog genes (Ewen-Campen B, *et al.*, 2017). Studying paralogs is a great way to fathom the nature of the paralog: Is it a "phenotype gap" (the existence of large number of genes that result from gene duplication and do not have detectable phenotypic effect when the genes are altered)?, did it acquire new functions?, did it retain varying degrees of overlapping functions, or did it acquire new functions and still retain some of its initial roles? (Rogers, R.L. *et al.*, 2009; Ewen-Campen B, *et al.*, 2017).

Materials and Methods.

Ecocyc, a sister site of MetaCyc and Biocyc, (Keseler *et al.*, 2013) was the first bioinformatics tool used to learn more about the CRISPR-Cas system in the model organism of this study, *Escherichia coli* K-12 MG1655. This bioinformatics tool is a scientific database dedicated to the prokaryote *Escherichia coli* K-12 MG1655. It includes an extensive literature-based curation of the many processes known to occur in this specific bacterial strain.

Next, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000; Kanehisa *et al.*,2017, Kanehisa *et al.*,2019) was used to determine the similarity or the difference between the CRISPR/Cas system of *M. ruber* DSM1279 and of *E. coli* K12 strain. The Integrated Genome and Metagenome comparative data analysis system (IMG/M) database (Markowitz *et al.*, 2012), together with National Center for Biotechnology Information (NCBI) Basic Local Assignment Search Tool (BLAST) (Madden, T., 2002) were used to confirm the start codon of mrub_3020 and mrub_1489 amino acid sequences. The NCBI BLAST was performed to further investigates the similarity between the model *E. coli* Cas3 protein (locus tag b2761) and the query mrub_3020 protein using their respective amino acid sequences. GenBank® is another essential tool that provided access to the complete genome of the study organism: *Meiothermus ruber* DSM 1279, which was used to retrieve the specific genetic information for mrub_3020 and mrub_1489 (GenBank Overview). E.coli b2761, mrub_3020 and mrub_1489 were the genes used for this study.

Three different bioinformatics tools were used to determine the cellular location of the *E. coli* Cas 3 protein and the two *M. ruber* putative proteins. TMHMM was used to predict whether the proteins could have membrane-embedded transmembrane alpha helices (Krogh A and Rapacki K, 2016). PRED (Bagos *et al.,* 2004, 2004), on the other hand, was used to determine if the two proteins are composed of membrane-embedded transmembrane beta-barrel, which would be localized to the outer membrane. PSORT-B (Yu *et al.,* 2010) uses refined localization subcategories and predictive probabilities to predict the protein's subcellular localization. The potential cellular locations in Gram negative bacteria include the cytoplasm, the inner membrane, the periplasm, the outer membrane and the extracellular space.

Conserved Domain Database (CDD), TIGRfam, Pfam, and PDB are the different bioinformatics tools used to gather structure-based evidences to determine if the three proteins proteins/polypeptide strands share the same physical properties such as protein domains and families. CDD (Marchler *et al.,* 2016) is a database of annotated multiple sequence alignment models for full-length proteins and ancient domains. TIGRfam (Haft *et al.,* 2001) is a resource of protein families that facilitates functional identification of proteins. Pfam (Finn *et al.,* 2016) is a database of protein families. It was used to analyze the amino acid sequences of the model gene and mrub_3020. PDB (Berman *et al.,* 2000) gives access to three-dimensional information for large biological molecules like proteins, DNA, and RNA. It was used to gather additional information on *E. coli* Cas3 (locus tag b2761) and the two putative *M. ruber* proteins.

Another set of bioinformatics tools were used to find evidence to determine whether mrub_3020 and mrub_1489 are paralogs. Protein Blastp was performed using mrub_3020 as the query against the *M. ruber* DSM1279 genome. The nucleotide sequence of the likely paralog, together with that of mrub_3020 obtained using NCBI BLAST, was used on T-Coffee (Notredame *et al.,* 2000) to make multiple-sequence alignment. Phylogeny.fr was used to create a phylogenetic tree. In addition, we returned to IMG/M to study the chromosomal organization of the genes flanking mrub_3020 and mrub_1489 to determine if one or both genes are part of a CRISPR-Cas operon.

Results.

Is *E. coli cas3* orthologous to *Mrub* _3020?

Taken from EcoCyc, Figure 5 shows the order of genes near the *cas3* gene on map position 2,884,553 . . . 2,887,219 of *E. coli* K-12 MG1655. The symbols "σ32" and "σ70" denote the sigma factor recognition sites (*aka* promoter region) for this set of genes. The *E. coli cas3* gene is shown to be outside but adjacent to the CRISPR-Cas operon, which begins with *casA* and ends with *cas2* on the far left. The same bioinformatics tool indicated that the translated *cas3* gene of *E. coli* yields a sequence of 888 amino acids (Figure 6). This amino acid sequence is folded into the Cas 3 protein and, EcoCyc predicts that the protein is in the cytosol of the cell.

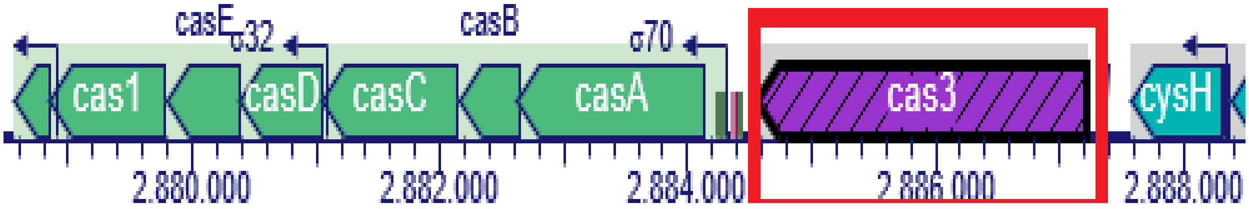


Figure 5. A section of the chromosomal map of *E. coli* K-12 MG1655 organism that contains the components of the Type I-3 CRISPR-Cas system. The *E. coli cas3* gene is highlighted in red and it is located between map position: 2,884,553 <- 2,887,219. The neighboring genes are colored green and they each encode a unique Cascade protein. The direction of the arrows show the transcription occurs from the right to left. The symbols "σ32" and "σ70" denote the sigma factor recognition sites (*aka* promoter region) for this set of genes.

>*E. coli*:b2761 K07012 CRISPR-associated endonuclease/helicase Cas3

```
MEPFKYICHYWGKSSKSLTKGNDIHLLIYHCLDVAAVADCWWDQSVVLQNTFCRNEMLSKQRVKA
WLLFFIALHDIGKFDIRFQYKSAESWLKLNPATPSLNGPSTQMCRKFNHGAAGLYWFNQDSLSEQSLG
DFFSFFDAAPHPYESWFPWVEAVTGHHGFILHSQDQDKSRWEMPASLASYAAQDKQAREEWISVLEA
LFLTPAGLSINDIPPDCSSLLAGFCSLADWLGSWTTTNTFLFNEDAPSDINALRTYFQDRQQDASRVLE
LSGLVSNKRCYEGVHALLDNGYQPRQLQVLVDALPVAPGLTVIEAPTGSGKTETALAYAWKLIDQQI
ADSVIFALPTQATANAMLTRMEASASHLFSSPNLILAHGNSRFNHLFQSIKSRAITEQGQEEAWVQCC
QWLSQSNKKVFLGQIGVCTIDQVLISVLPVKHRFIRGLGIGRSVLIVDEVHAYDTYMNGLLEAVLKAQ
ADVGGSVILLSATLPMKQKQKLLDTYGLHTDPVENNSAYPLINWRGVNGAQRFDLLAHPEQLPPRFSI
QPEPICLADMLPDLTMLERMIAAANAGAQVCLICNLVDVAQVCYQRLKELNNTQVDIDLFHARFTLN
DRREKENRVISNFGKNGKRNVGRILVATQVVEQSLDVDFDWLITQHCPADLLFQRLGRLHRHHRKYR
PAGFEIPVATILLPDGEGYGRHEHIYSNVRVMWRTQQHIEELNGASLFFPDAYRQWLDSIYDDAEMDE
PEWVGNGMDKFESAECEKRFKARKVLQWAEEYSLQDNDETILAVTRDGEMSLPLLPYVQTSSGKQL
```

Figure 6. The FASTA formatted amino acid sequence of *E. coli* b2761 is shown above. The single-letter abbreviations for each amino acid are used. The first amino acid "M" (methionine) at the top left and the last amino acid L (lysine) denote the N-terminal and C-terminal, respectively, of the Cas3 protein.

The IMG/M chromosome maps generated for *E. coli* b2761, mrub_3020, and mrub_1489 show that these three genes are adjacent to a CRISPR array and other *cas* genes, which suggests they are likely components of a CRISPR-Cas system (Figure 7). The *cas* gene order for the region containing *E. coli cas3* and mrub_3020 are nearly identical (See Table 1), which strongly suggests that both have a Type I-E system. The set of *cas* genes positioned near mrub_1489 suggests a Type I-C and/or Type III-C CRISPR-Cas system (Table 1).

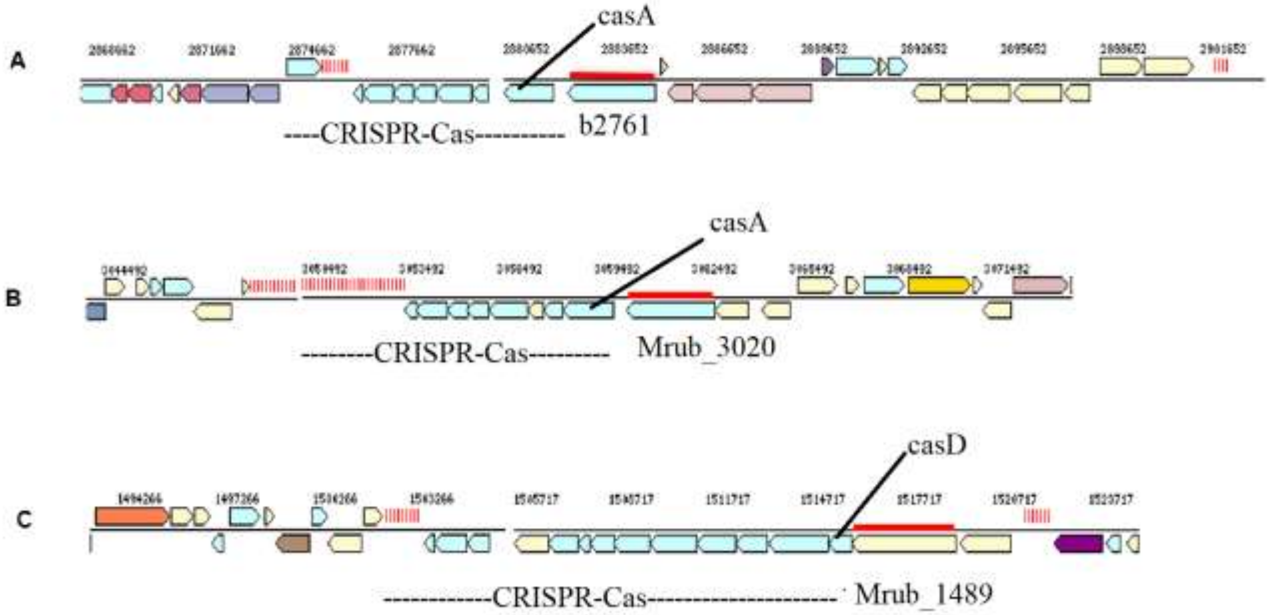


Figure 7. IMG/M chromosomal map suggests that *E. coli* b2761/*cas3* (Panel A) and the two putative *cas3* genes of *M. ruber* (mrub_3020 in Panel B, and mrub_1489 in Panel C) are components of a CRISPR-Cas system. The multiple red vertical lines represent the CRISPR array.

KEGG was used to compare the CRISPR/Cas systems in *M. ruber* and *E. coli*. Both CRISPR/Cas systems could have specific genes that encode the same protein (e.g. Cas A proteins). But, there could be some proteins that are only encoded by certain genes in the CRISPR/Cas system of either the model organism or the test organism. (Cmr: Chlorophenicol resistance protein; CSM: Cutoff Scanning Matrix, which is a recent protein function prediction and structural classification method; cas: CRISPR-associated system) (Table 1).

Table 1A. Comparison between the genes involved in the CRISPR/Cas system of *E. coli* and *M. ruber* using KEGG.

Locus Tag(s)	Gene product name	*E. coli K12*	*M. ruber*
Cas 1	multifunctional nuclease Cas1	b2755	Mrub_0224 Mrub_1477 Mrub_3013
Cas 2	CRISPR-associated endoribonuclease Cas2	b2754	Mrub_1476 Mrub_0225 Mrub_3012
Type I CRISPR-Cas system.			
Type I signature cas proteins			
Cas 3	CRISPR-associated helicase	b2761(has endonuclease role)	Mrub_3020
Subtype I-A factors			
Cas 4	CRISPR-associated protein	N/A	Mrub_1478
Cas 6	CRISPR-associated protein		Mrub_0222
I-B factors			
Cas 4	CRISPR-associated protein	N/A	Mrub_1478
I-C factors			
Csd1	CRISPR-associated protein	N/A	Mrub_1487
Csd2	CRISPR-associated protein		Mrub_1486
Cas4	CRISPR-associated protein		Mrub_1478
Cas5 family	CRISPR-associated protein		Mrub_1488
I-D factors			
Cas4	CRISPR-associated protein	N/A	Mrub_1478
Cas6	CRISPR-associated protein		Mrub_0222
I-E factors			
CasA	CRISPR system Cascade subunit CasA	b2760	Mrub_3019
CasB	CRISPR system Cascade subunit CasB	b2759	Mrub_3018
CasC	CRISPR system Cascade subunit CasC	b2758	Mrub_3016
CasD	CRISPR system Cascade subunit CasD	b2757	Mrub_3015
CasE	pre-CRISPR RNA endonuclease	b2756	Mrub_3014
I-F factors	N/A	N/A	N/A
I-U factors	N/A	N/A	N/A

Table 1B. Comparison between the genes involved in the CRISPR/Cas system of *E. coli* and *M. ruber* using KEGG (Continued)

Locus Tag(s)	Gene product name	*E. coli K12*	*M. ruber*
Type II CRISPR-Cas System			
Type II signature *cas* proteins			
Subtype			
II-A factors	N/A	N/A	
II-B factors			
Cas 4	CRISPR-associated protein	N/A	Mrub_1478
Type III signature *cas* proteins			
Csm 1	CRISPR-associated protein	N/A	Mrub_0215
Subtype			
III-A factors		N/A	
Csm 2	CRISPR-associated protein		Mrub_0216
Csm 3	CRISPR-assoc. RAMP protein		Mrub_0217
Csm 4	CRISPR-assoc. RAMP protein		Mrub_0218
Csm 5	CRISPR-assoc. RAMP protein		Mrub_0219
III-B factors			
Cmr 1	CRISPR-assoc. RAMP protein	N/A	Mrub_1485
Cmr 2	CRISPR-associated protein		Mrub_1484
Cmr 3	CRISPR-associated protein		Mrub_1483
Cmr 4	CRISPR-assoc. RAMP protein		Mrub_1482
Cmr 5	CRISPR-associated protein		Mrub_1481
Cmr 6	CRISPR-assoc. RAMP protein		Mrub_1480
III-C factors			

Mrub_3020 could be orthologous to *E. coli* b2761 because both genes are predicted to encode the Cas 3 protein of the CRISPR/Cas system. But, the amino acid sequence of the translated mrub_3020 gene (Figure 8 A) is different from that of *E. coli* b2761 (Figure 6). Mrub_1489 (predicted as a metal dependent phosphohydrolase) might be a paralog to mrub_3020, which may have diverged in sequence and function over evolutionary time since it would no longer be under the same functional constraints as mrub_3020: (Figure 8 B).

A

```
>646674485 YP_003508784 CRISPR-associated helicase, Cas3 family [Meiothermus ruber DSM 1279
chromosome: NC_013946]
MSLSETARALWAKSDRGREQGAWHPLIAHLLDVAACAEAILEREPPKTLELYAHDLSLEPQQAKAWVCAL
AGLHDIGKASPAFQQKWPEGKERLWATGLTWSSDPTPPPHDLSHSIISEVVLPELLEARGWKYRAAQNVA
AAVGEHHGFRATRGDLDKATTREKGNANWDEVRRELFEAVLEVLGVGEAPKVKLYGGAAFERLAGLTSF
ADWIGSSLDFHPLGDDLAGYYREAKARAAQKLDGIGWFQRKTLMPEPQSLEEVFAYLGSPEAPFRARPLQ
AAIERLLEGVDCPALLLVEAPMGEGKTEAAFYAHLRLQAANGHRGMYVALPTQATGNLMFERAKAFLDR
WGQSRKLDLQLLHGASELVEAYQEIRVRPNSPEEREEGVEAQVWFSHRKRGLLSEYAVGTVDQALLGILPT
KHQFVRLWGLGNRVVVLDEVHAYDTYTSGLIEMLVRWLRALDSSVVLMSATLPRAKRENLLRAFGAEKI
TEDKPYPRITRVVKDNPMPVVETFEACKQLTLQLRALPLDLEAIAEQALEQARRGGCVACIVNTVQRAQEL
YRALAGNSDGVEVYLFHARYPLEERLNREQLVLAKFGKQGQRPKRAILVATQVVEQSLDLDFDVMFTDLA
PVDLVLQRAGRLHRHARSAEERHSHTEPVLWVAGLECEGVPDFGTAERIYERYVLLRSWLALRNRTRIGLP
GDIDRLVQEVYSDMPQGPSEAWKRALEEAQARMEKRDARDQDEAFYAPFGDPDETGWLEPRDFTRLPDD
EPNPDDDPSLLKTRKGPPSATVVLLHRVGGQLCLDAGGKEGVSLASQLELAQARRIFARSVKLSRYELVQN
NLEALEAHRKAHDLPTKPWSETPLLAHAHPVVLEGGCAVLGELVLELHPELGVVYGSAL
```

B

```
>646672954 YP_003507271 metal dependent phosphohydrolase [Meiothermus ruber DSM 1279 chromosome:
NC_013946]
MEATYHQNKAHRLLRLLELLEQKAWRPHELRRELGLGERAIFDYLLEAQALAERLGLEFRHDRLRGL
YWVEVRERLSLTETVVAHAALRMLAHHAPGSNKAYQESLRKLARSLPEPLRSIALRSTEALNQRPPSL
SGANLETLTQGWLNQQVVAFEYRLPQARVIRVELETYFIEVSRANMAVYVIGKDRLYGRGLHYLENL
KTYKLERIQRPRLLDETYTIPDDFDPSQYLSSAWGIVRSEPPMRVRLRFNPEASERIREGGYPNLQILEQ
LEGGSTLVQITVGTDTEGFPLELLPWIQSWGPRVEVLEPESLRQAWLAEARAVLEQYGQPGLAFRTYW
AHTHPNPARWQPLREHLHQVAERAAAKARPFGEEENARLAGRLHDLGKYGDLFQRRLEGREKGLDH
WSAGAHLALFEYRQPAVALAVQGHHIGLQSGARESLMEMKLREDGKGVPPELRLSETDLEVLKARL
QKDGLELPPPSQTRISPPAGAAAMLDTRMLFSALVDADFLDTEAHIKGPEARPAPPELRAREALERLEA
HLAQLSQAGHIPQKTLELRRVVAEAAASAAEQAASVFTLTAPTGLGKTLAMLRFALRRAARDPRIRRI
VVVLPYLSILDQTAKVYRELFADFGPHYILEDHSLAYRPLSRELSDEQDLQERERRLLSENWEAPIVLTT
HVQLLESLHANRPGACRKLHNLAGSVLLFDEVQTLPTHLAVPTLKTLARLASQKYGAVVVFATATQPA
FDTLHEQIQRGEPQGWQPVEMVPEPERLFAQSRRVELEWWLKNPIPWPHLATLLEAEPQVLAVLNLKR
QAYALFQESQARNLEGLYHLSTALCPAHRRRVLEEVQRRLEQGQPCRLVATQVVEAGVELDFPAGYRA
LGPLEAIAQTAGRINRHGLRPQGRLVVFLPQEEAYPDRAYGRAAALTRALQAEGPLTLEPTTFRRYYQSL
YALQQVSDPAIEALIQTQNYTELARRYRIIESVAVNVVVPYNDEALALMQEARDHGISAAWIHRARPYT
VPYFLPKDGPPPFLETVFLRYGRGEAPDWFLSADPALYDARLGFTPQDGASVGLVV
```

Figure 8. FASTA formatted amino acid sequence of mrub_3020 (A) and mrub_1489 (B). The amino acid sequences include the appropriate locus tags for correct identification. The first amino acid and the last amino acid of the sequences denote the N-terminal and C-terminal, respectively.

GenBank® was used to collect some basic information for mrub_3020 and mrub_1489, whiles EcoCyc was used to collect the corresponding information for *E. coli* b2761. IMG/M was used to generate the chromosomal maps for the two systems that focuses on the chromosomal locations of the three genes. Mrub_3020 and *E. coli* b2761 proteins have helicase activity in the CRISPR/Cas system. But, the Cas 3 protein of the *E. coli* CRISPR/Cas system has an additional role: endonuclease activity. Mrub_1489 encodes a metal dependent phosphodydrolase whose name does not suggest a role in the CRISPR/Cas system. All the genes are found at different location of their respective chromosome. Neither genes have the same nucleotide sequences because their amino acid sequences appear different (figures 6 and 8). All three genes appear to be part of an operon system, but they are generally flanked by different genes. The only exception is that both *E. coli* b2761 and mrub_3020 genes are flanked by a Cas A protein encoding gene (Table 2).

Table 2. Comparing basic information for the three genes: *E. coli* b2761 (using EcoCyc), mrub_3020 and mrub_1489 (using GenBank®, and IMG/M).

Feature	*E. coli* b2761	Mrub_3020	mrub_1489
Gene product name	CRISPR-associated helicase / endonuclease	CRISPR-associated helicase	Metal dependent Phosphohydrolase
DNA coordinates	2884553 . . . 2887219	3060491. . . 3063190	1516312..1519548
Nucleotide sequences	Different		
Positioned adjacent to a *cas* operon	Yes		
Upstream Gene from gene of interest (G.O.I)	casA: CRISPR-associated protein, Cse1 family	mrub_3019 (cas A: CRISPR-associated protein, Cse1 family)	mrub_1488 (CRISPR-associated protein Cas5d, family)
Downstream gene from G.O.I	cysH: phospho adenylylsulfate reductate (thioredoxin)	mrub_3021 (Transcriptional regulator protein)	mrub_1490 (tetratricopeptide domain containing protein)

IMG/M was used to analyze mrub_3020 and mrub_1489 to confirm their start position for translation. The image of the upstream region shows the predicted start codon for mrub_3020 (Figure 9 A) and mrub_1489 (Figure 9 B). A putative Shine-Delgarno sequence is noted for mrub_3020, but not mrub_1489. The start position is marked by the Shine-Delgarno (SD) sequence, but there appeared to be two possible options for mrub_3020. Because a typical SD sequence is 8-13 bases from a start codon, the sequence closest to the ATG is the most likely one of the two. Regarding mrub_1489, it is not unexpected that a suitable SD region might not be identified for an *M. ruber* gene, as a consensus sequence for the *M. ruber* Shine-Delgarno sequence has not been determined (Scott, personal communication). It is more surprising when an SD is predicted. Regardless, there is no evidence to suggest that a wrong start codon has been called for either *M. ruber* gene.

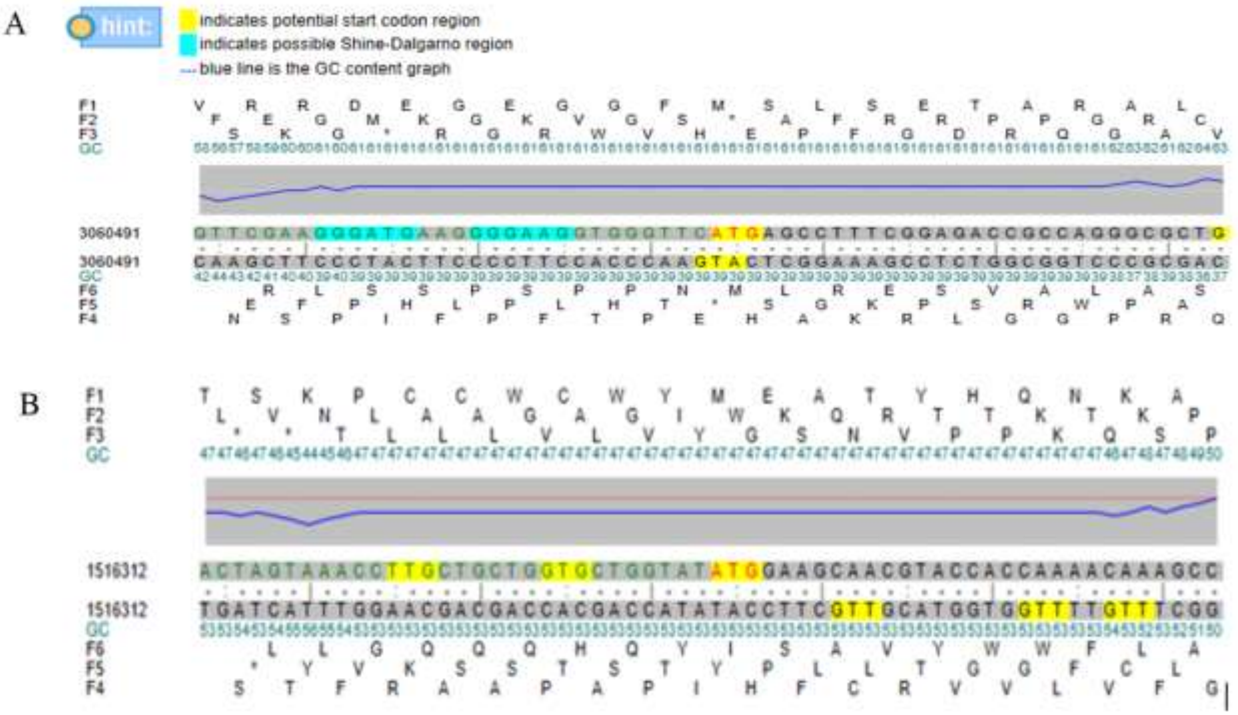
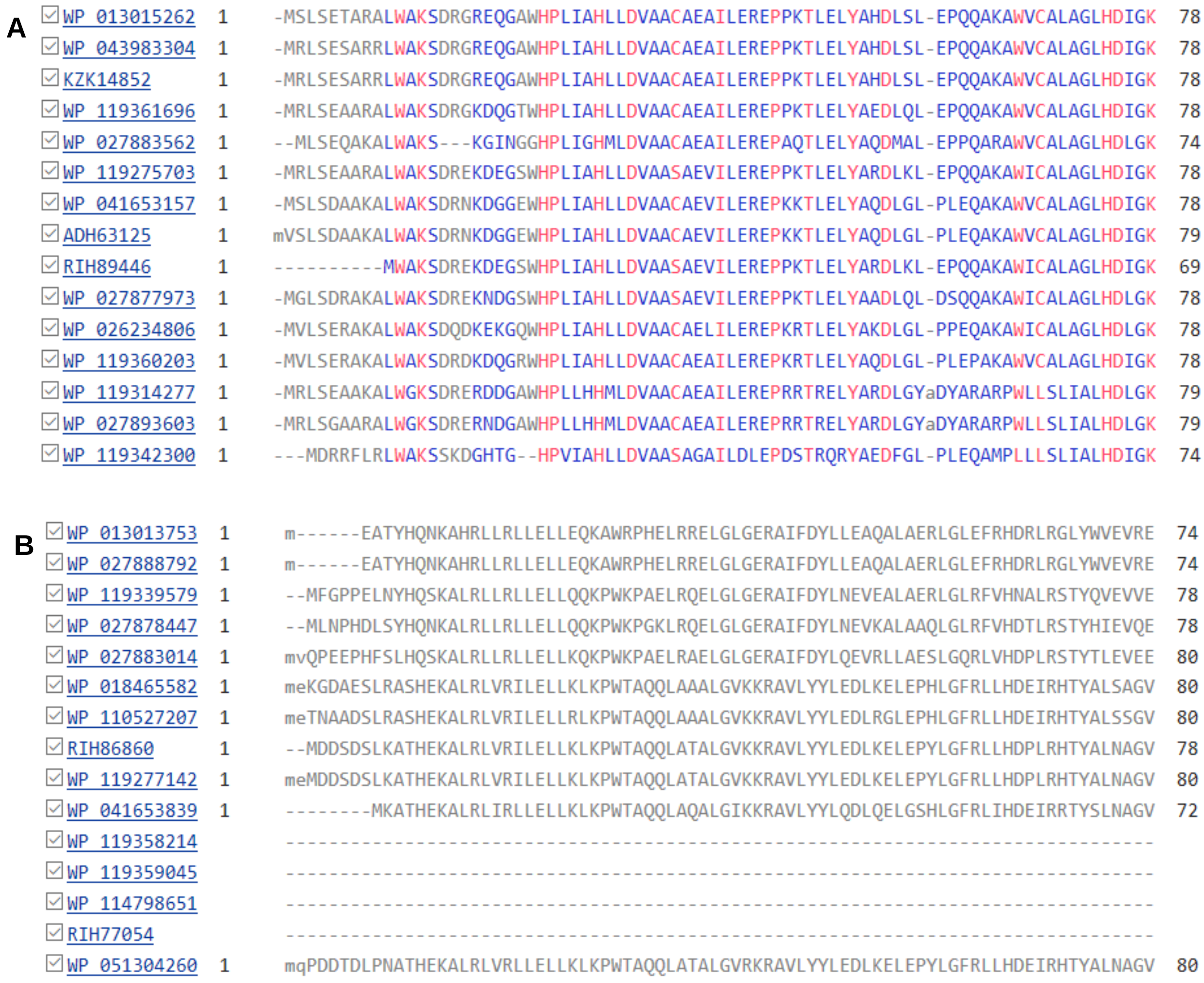

A
hint:
indicates potential start codon region
indicates possible Shine-Dalgarno region
blue line is the GC content graph
F1 V R R D E G E G G F M S L S E T A R A L
F2 F E G M K G K V G S * A F R R P P G R C
F3 S K G * R G R W V H E P F G D R Q G A V
GC
3060491 GTTCGAAGGGATGAAGGGGAAGGTGGGTTCATGAGCCTTTCGGAGACCGCCAGGGCGCTG
3060491 CAAGCTTCCCTACTTCCCCTTCCACCCAAGTACTCGGAAAGCCTCTGGCGGTCCCGCGAC
GC
F6 R L S S P S P P N M L R E S V A L A S
F5 E F P H L P L H T * S G K P S R W P A
F4 N S P I F P F T P E H A K R L G G P R Q
B
F1 T S K P C C W C W Y M E A T Y H Q N K A
F2 L V N L A A G A G I W K Q R T T K T K P
F3 * * T L L L V L V Y G S N V P P K Q S P
GC
1516312 ACTAGTAAACCTTGCTGCTGGTGCTGGTATATGGAAGCAACGTACCACCAAAACAAAGCC
1516312 TGATCATTTGGAACGACGACCACGACCATATACCTTCGTTGCATGGTGGTTTTGTTTCGG
GC
F6 L L G Q Q Q H Q Y I S A V Y W W F L A
F5 * Y V K S S T S T Y P L L T G G F C L
F4 S T F R A A P A P I H F C R V V L V F G

Figure 9. The 5' region of both mrub_3020 (A) and mrub_1489 (B) show only one likely start codon. The nucleotide sequence provided start at map position 3060491 for mrub_3020 and position 1516312 for mrub_1489. All the six reading frames (*e.g.* F1, F2, *etc.)* are translated into the single-letter amino acid abbreviations. Reading frame F1 starts translating 30 bases upstream of the start codon. The starting position of this protein is the methionine (M) above the start codon (marked by ATG in red) in F1. Other ATG codons are highlighted in yellow. Two potential Shine-Delgarno regions (*aka* the ribosome binding site) are identified (GGGATG and GGGAAG highlighted in blue) for mrub_3020, but none is identified for mrub_1489. The actual Shine-Delgarno (SD) region is 8-15 bases upstream of the start codon and there should be only one ATG codon in the same reading frame as the start codon.

The NCBI BLAST was used to create a multiple sequence alignment using the amino acid sequence for both *M. ruber* genes. The amino terminus of 15 likely orthologs is shown in Figure 10. Most of the amino acid sequences have methionine as their first amino acid for mrub_3020 (Fig. 10A). Along with the information in Figure 9, this observation supports the hypothesis that the correct start codon has been identified. The interpretation is more difficult for mrub_1489 (10B), however. The amino terminus of the chosen orthologs show greater amino acid variability and a clear alignment was difficult to make. On further analysis of the downstream region from the initial M/methionine, however, there are no nearby M amino acids but the overall alignment improves. We conclude that the best start codon has likely been identified for mrub_3020.

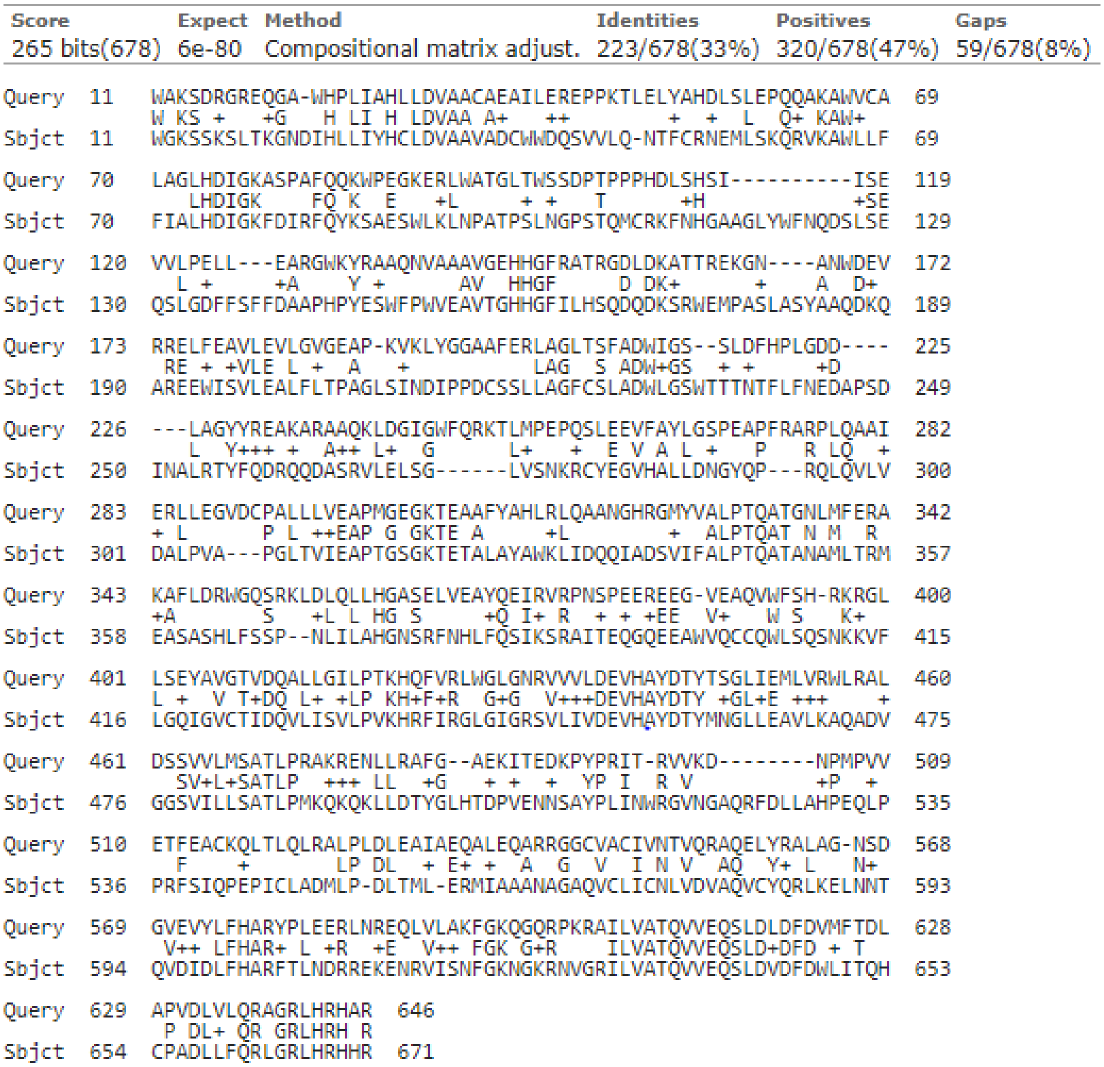


```
A
☑ WP_013015262  1    -MSLSETARALWAKSDRGREQGAWHPLIAHLLDVAACAEAILEREPPKTLELYAHDLSL-EPQQAKAWVCALAGLHDIGK  78
☑ WP_043983304  1    -MRLSESARRLWAKSDRGREQGAWHPLIAHLLDVAACAEAILEREPPKTLELYAHDLSL-EPQQAKAWVCALAGLHDIGK  78
☑ KZK14852      1    -MRLSESARRLWAKSDRGREQGAWHPLIAHLLDVAACAEAILEREPPKTLELYAHDLSL-EPQQAKAWVCALAGLHDIGK  78
☑ WP_119361696  1    -MRLSEAARALWAKSDRGKDQGTWHPLIAHLLDVAACAEAILEREPPKTLELYAEDLQL-EPQQAKAWVCALAGLHDIGK  78
☑ WP_027883562  1    --MLSEQAKALWAKS---KGINGGHPLIGHMLDVAACAEAILEREPAQTLELYAQDMAL-EPPQARAWVCALAGLHDLGK  74
☑ WP_119275703  1    -MRLSEAARALWAKSDREKDEGSWHPLIAHLLDVAASAEVILEREPPKTLELYARDLKL-EPQQAKAWICALAGLHDIGK  78
☑ WP_041653157  1    -MSLSDAAKALWAKSDRNKDGGEWHPLIAHLLDVAACAEVILEREPKKTLELYAQDLGL-PLEQAKAWVCALAGLHDIGK  78
☑ ADH63125      1    mVSLSDAAKALWAKSDRNKDGGEWHPLIAHLLDVAACAEVILEREPKKTLELYAQDLGL-PLEQAKAWVCALAGLHDIGK  79
☑ RIH89446      1    ----------MWAKSDREKDEGSWHPLIAHLLDVAASAEVILEREPPKTLELYARDLKL-EPQQAKAWICALAGLHDIGK  69
☑ WP_027877973  1    -MGLSDRAKALWAKSDREKNDGSWHPLIAHLLDVAASAEVILEREPPKTLELYAADLQL-DSQQAKAWICALAGLHDLGK  78
☑ WP_026234806  1    -MVLSERAKALWAKSDQDKEKGQWHPLIAHLLDVAACAELILEREPKRTLELYAKDLGL-PPEQAKAWICALAGLHDLGK  78
☑ WP_119360203  1    -MVLSERAKALWAKSDRDKDQGRWHPLIAHLLDVAACAEAILEREPKRTLELYAQDLGL-PLEPAKAWVCALAGLHDIGK  78
☑ WP_119314277  1    -MRLSEAAKALWGKSDRERDDGAWHPLLHHMLDVAACAEAILEREPRRTRELYARDLGYaDYARARPWLLSLIALHDLGK  79
☑ WP_027893603  1    -MRLSGAARALWGKSDRERNDGAWHPLLHHMLDVAACAEAILEREPRRTRELYARDLGYaDYARARPWLLSLIALHDLGK  79
☑ WP_119342300  1    ---MDRRFLRLWAKSSKDGHTG--HPVIAHLLDVAASAGAILDLEPDSTRQRYAEDFGL-PLEQAMPLLLSLIALHDIGK  74

B
☑ WP_013013753  1    m------EATYHQNKAHRLLRLLELLEQKAWRPHELRRELGLGERAIFDYLLEAQALAERLGLEFRHDRLRGLYWVEVRE  74
☑ WP_027888792  1    m------EATYHQNKAHRLLRLLELLEQKAWRPHELRRELGLGERAIFDYLLEAQALAERLGLEFRHDRLRGLYWVEVRE  74
☑ WP_119339579  1    --MFGPPELNYHQSKALRLLRLLELLQQKPWKPAELRQELGLGERAIFDYLNEVEALAERLGLRFVHNALRSTYQVEVVE  78
☑ WP_027878447  1    --MLNPHDLSYHQNKALRLLRLLELLQQKPWKPGKLRQELGLGERAIFDYLNEVKALAAQLGLRFVHDTLRSTYHIEVQE  78
☑ WP_027883014  1    mvQPEEPHFSLHQSKALRLLRLLELLKQKPWKPAELRAELGLGERAIFDYLQEVRLLAESLGQRLVHDPLRSTYTLEVEE  80
☑ WP_018465582  1    meKGDAESLRASHEKALRLVRILELLKLKPWTAQQLAAALGVKKRAVLYYLEDLKELEPHLGFRLLHDEIRHTYALSAGV  80
☑ WP_110527207  1    meTNAADSLRASHEKALRLVRILELLRLKPWTAQQLAAALGVKKRAVLYYLEDLRGLEPHLGFRLLHDEIRHTYALSSGV  80
☑ RIH86860      1    --MDDSDSLKATHEKALRLVRILELLKLKPWTAQQLATALGVKKRAVLYYLEDLKELEPYLGFRLLHDPLRHTYALNAGV  78
☑ WP_119277142  1    meMDDSDSLKATHEKALRLVRILELLKLKPWTAQQLATALGVKKRAVLYYLEDLKELEPYLGFRLLHDPLRHTYALNAGV  80
☑ WP_041653839  1    -------MKATHEKALRLIRLLELLKLKPWTAQQLAQALGIKKRAVLYYLQDLQELGSHLGFRLIHDEIRRTYSLNAGV  72
☑ WP_119358214       --------------------------------------------------------------------------------
☑ WP_119359045       --------------------------------------------------------------------------------
☑ WP_114798651       --------------------------------------------------------------------------------
☑ RIH77054           --------------------------------------------------------------------------------
☑ WP_051304260  1    mqPDDTDLPNATHEKALRLVRLLELLKLKPWTAQQLATALGVRKRAVLYYLEDLKELEPYLGFRLLHDEIRHTYALNAGV  80
```

Figure 10. Multiple amino acid sequence alignment analysis confirms the correct start codon has been identified for only mrub_3020 (A), but not for mrub_1489 (B). NCBI blast was performed using fifteen different amino acid sequences from species of the same *Meiothermus* genus. There is a good alignment for methionine (the first amino acid) and other amino acids when the same amino acids line up for all the fifteen different sequences. There is a good alignment for mrub_3020 (accession number: WP_013015262), but there is a poor alignment for mrub_1489 (accession number: WP_013013753) because only about four methionine line up and four of the supposed amino acid sequences are made up of only dashed lines. Dashed lines could mean that no amino acids were encoded for by that region of the nucleotide sequence.

Protein BLAST of the *E. coli* gene against the putative mrub_3020 was performed. The two amino acid sequences have some identical amino acids thus, a non-zero percent identities (33%) was obtained for the pairwise alignment. The recorded expect value (E-value) of the pairwise alignment ($6 * 10^{-80}$) is lower than the cut-off value of 0.001. There are 320 chemically similar amino acids between the *E. coli* and *M. ruber* amino acid sequence. The alignment specifically started at amino acid 11 (Figure 11).



Score	Expect	Method	Identities	Positives	Gaps
265 bits(678)	6e-80	Compositional matrix adjust.	223/678(33%)	320/678(47%)	59/678(8%)

```
Query  11   WAKSDRGREQGA-WHPLIAHLLDVAACAEAILEREPPKTLELYAHDLSLEPQQAKAWVCA  69
            W KS +   +G   H LI H LDVAA A+   ++        +  +  L  Q+ KAW+
Sbjct  11   WGKSSKSLTKGNDIHLLIYHCLDVAAVADCWWDQSVVLQ-NTFCRNEMLSKQRVKAWLLF  69

Query  70   LAGLHDIGKASPAFQQKWPEGKERLWATGLTWSSDPTPPPHDLSHSI----------ISE  119
               LHDIGK    FQ K  E   +L     + +   T      +H            +SE
Sbjct  70   FIALHDIGKFDIRFQYKSAESWLKLNPATPSLNGPSTQMCRKFNHGAAGLYWFNQDSLSE  129

Query  120  VVLPELL---EARGWKYRAAQNVAAAVGEHHGFRATRGDLDKATTREKGN----ANWDEV  172
              L +     +A    Y +      AV  HHGF     D DK+      +    A  D+
Sbjct  130  QSLGDFFSFFDAAPHPYESWFPWVEAVTGHHGFILHSQDQDKSRWEMPASLASYAAQDKQ  189

Query  173  RRELFEAVLEVLGVGEAP-KVKLYGGAAFERLAGLTSFADWIGS--SLDFHPLGDD----  225
             RE + +VLE L +  A   +          LAG  S ADW+GS  + +     +D
Sbjct  190  AREEWISVLEALFLTPAGLSINDIPPDCSSLLAGFCSLADWLGSWTTTNTFLFNEDAPSD  249

Query  226  ---LAGYYREAKARAAQKLDGIGWFQRKTLMPEPQSLEEVFAYLGSPEAPFRARPLQAAI  282
               L  Y+++ +  A++ L+  G      L+   +  E V A L +   P   R LQ  +
Sbjct  250  INALRTYFQDRQQDASRVLELSG------LVSNKRCYEGVHALLDNGYQP---RQLQVLV  300

Query  283  ERLLEGVDCPALLLVEAPMGEGKTEAAFYAHLRLQAANGHRGMYVALPTQATGNLMFERA  342
            + L      P L ++EAP G GKTE A     +L         +  ALPTQAT N M  R
Sbjct  301  DALPVA---PGLTVIEAPTGSGKTETALAYAWKLIDQQIADSVIFALPTQATANAMLTRM  357

Query  343  KAFLDRWGQSRKLDLQLLHGASELVEAYQEIRVRPNSPEEREEG-VEAQVWFSH-RKRGL  400
            +A       S   +L L HG S     +Q I+ R  + + +EE  V+   W S   K+
Sbjct  358  EASASHLFSSP--NLILAHGNSRFNHLFQSIKSRAITEQGQEEAWVQCCQWLSQSNKKVF  415

Query  401  LSEYAVGTVDQALLGILPTKHQFVRLWGLGNRVVVLDEVHAYDTYTSGLIEMLVRWLRAL  460
            L +  V T+DQ L+ +LP KH+F+R  G+G  V+++DEVHAYDTY +GL+E +++    +
Sbjct  416  LGQIGVCTIDQVLISVLPVKHRFIRGLGIGRSVLIVDEVHAYDTYMNGLLEAVLKAQADV  475

Query  461  DSSVVLMSATLPRAKRENLLRAFG--AEKITEDKPYPRIT-RVVKD--------NPMPVV  509
              SV+L+SATLP  +++ LL  +G   + +  +  YP I  R V          +P  +
Sbjct  476  GGSVILLSATLPMKQKQKLLDTYGLHTDPVENNSAYPLINWRGVNGAQRFDLLAHPEQLP  535

Query  510  ETFEACKQLTLQLRALPLDLEAIAEQALEQARRGGCVACIVNTVQRAQELYRALAG-NSD  568
              F    +       LP DL  + E+ +  A  G  V  I N V  AQ  Y+ L   N+
Sbjct  536  PRFSIQPEPICLADMLP-DLTML-ERMIAAANAGAQVCLICNLVDVAQVCYQRLKELNNT  593

Query  569  GVEVYLFHARYPLEERLNREQLVLAKFGKQGQRPKRAILVATQVVEQSLDLDFDVMFTDL  628
             V++ LFHAR+ L +R  +E  V++ FGK G+R    ILVATQVVEQSLD+DFD + T
Sbjct  594  QVDIDLFHARFTLNDRREKENRVISNFGKNGKRNVGRILVATQVVEQSLDVDFDWLITQH  653

Query  629  APVDLVLQRAGRLHRHAR  646
             P DL+ QR GRLHRH R
Sbjct  654  CPADLLFQRLGRLHRHHR  671
```

Figure 11. Protein BLAST pairwise alignment between *E. coli* gene and putative mrub_3020 ortholog show that there is a significant similarity between the two sequences. 33% of the amino acids of both sequences are identical. 320 amino acids are either identical or come from chemically similar groups. These 320 amino acids are represented by the + signs (positives). The alignment begins at amino acid sequence 11.

Three different bioinformatics tools were used to determine the cellular localization of each protein: inside or outside the cell. Both *E. coli* and *M.ruber* are Gram negative Eubacteria. The TMHMM bioinformatics tool was used to predict the number of transmembrane alpha helices. All the three genes had no predicted transmembrane alpha helices (Figure 12). None of the proteins are predicted to be transmembrane alpha helices.

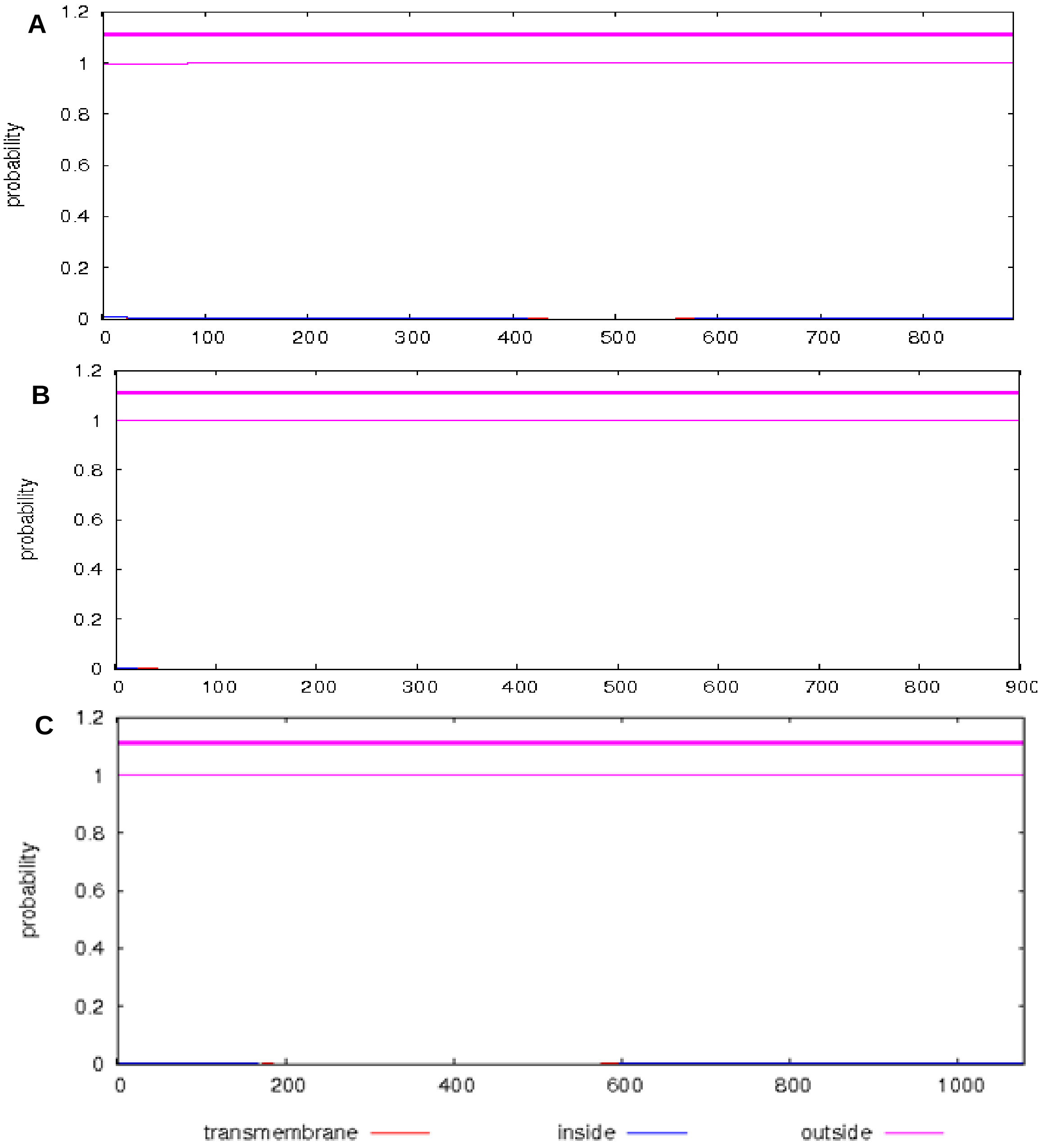


Figure 12. TMHMM posterior probabilities plots for *E. coli* b2761 (A), mrub_3020 (B), and mrub_1489 (C) predict no transmembrane alpha helices for the three proteins encoded by the respective gene. There is little to no probability of finding either proteins as an internal or transmembrane alpha helix.

PRED analysis is performed using the amino acid sequence of *E. coli* and mrub_3020 Cas 3 proteins, in addition to the sequence of the mrub_1489 encoded protein, to predict if the proteins could be beta-pleated sheets. The posterior probability plots (Figure 13) show that the two proteins are not hydrophobic transmembrane beta-pleated sheets that span the entire cell membrane. The corresponding amino acid sequences are polar and not non-polar.

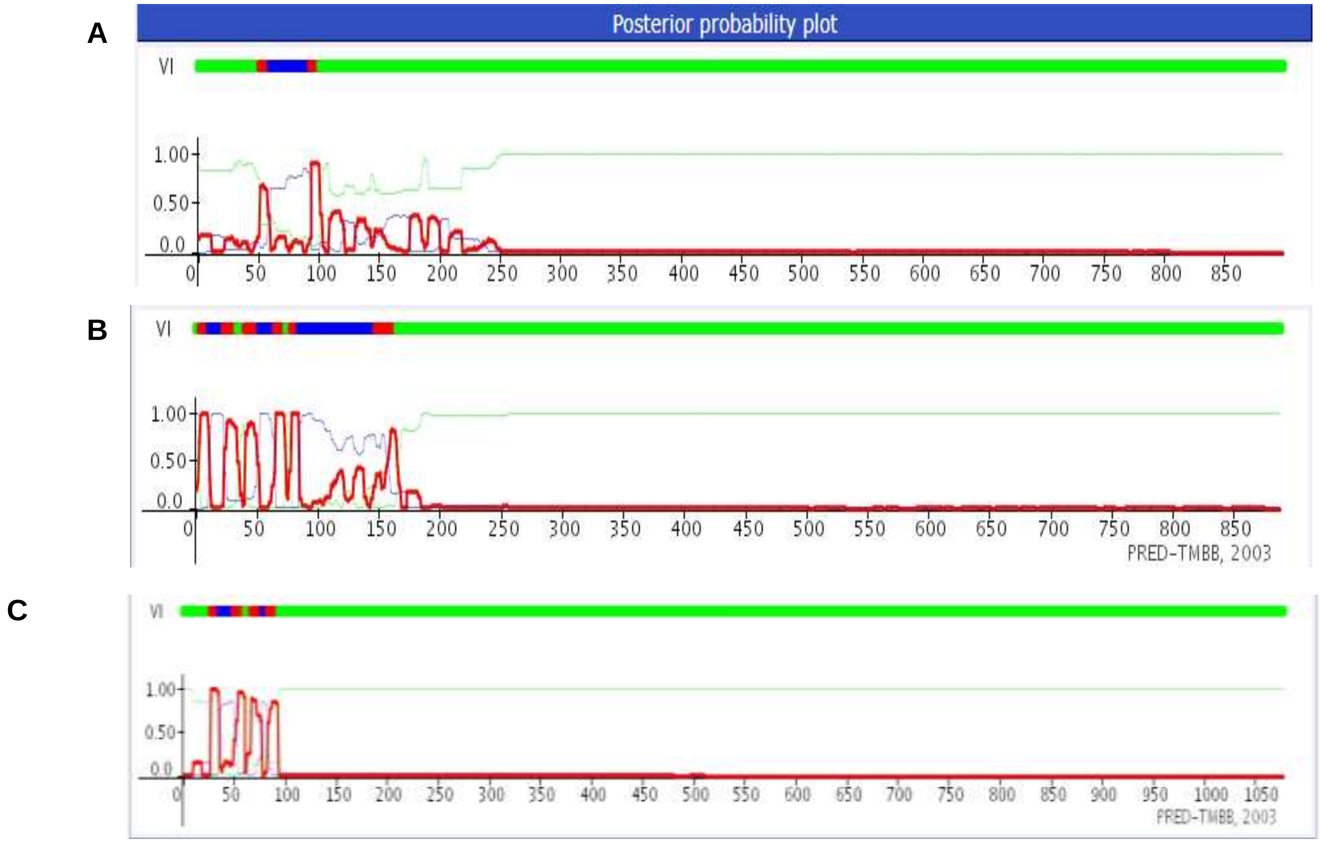


Figure 13. Posterior probability plots indicate that neither *M. ruber* putative Cas 3 protein (A), nor *E. coli* Cas 3 protein (B), nor *M. ruber* putative Metal dependent phosphohydrolyase (C) are hydrophobic, beta-pleated sheet that span the entire cell membranes. Only a small portion of the cell membrane is spanned by hydrophobic, beta pleated sheets.

PSORT-B is the last bioinformatics tool that helps determine the cellular location of the proteins. The *E. coli* protein has equal chance of been in six different cellular location (all cellular locations had a score of 2.00). *M.ruber* putative Cas 3 protein had scores for all the different cellular locations. But, the highest of the six scores was the cytoplasmic score of 8.96. Mrub_1489 putative metal-dependent phosphohydrolase had different scores for all the different cellular locations.

Table 3. PSORT-B predict that *M.ruber* Cas 3 is located in the cytoplasm, whiles it predict an unknown location for *E. coli* Cas 3 protein and mrub_1489 encoded protein.

Cellular location	Scores		
	Mrub_3020	*E. coli* b2761	Mrub_1489
Cytoplasmic score	8.96	2.00	5.48
Cytoplasmic Membrane score	0.51	2.00	0.10
Cell wall score	N/A (Gram negative)	N/A (Gram negative)	N/A (Gram negative)
Periplasmic score	0.26	2.00	0.48
Outer Membrane score	0.01	2.00	1.93
Extracellular score	0.26	2.00	2.01

Four different bioinformatics tools were used to probe the functions of the Cas 3 protein and the metal-dependent phophohydrolase. CDD found one matching protein of *E. coli* – Helicase Cas 3 (PRK09694) and a different matching protein for *M.ruber* - CRISPR/Cas system-associated endonuclease/helicase Cas3 [Defense mechanisms] (COG1203). But, mrub_1489 had multiple domains matches – a CRISPR/Cas system-associated helicase (COG1203); WYL domain (Pfam13280); and a CRISPR/Cas system-associated protein Cas3’’(cd09641). TIGRfam analysis of the different amino acid sequences identified the same two different proteins for *E. coli* b2761 and mrub_3020 – cas3_core: CRISPR-associated helicase Cas3 (TIGR01587), and cas3_HD: CRISPR-associated endonuclease Cas (TIGR01596). The same hits were found for mrub_1489, together with six other proteins and domains – DEAD: DEAD/DEAH box helicase (PF00270); HDIG: uncharacterized domain HDIG (TIGR00277); HD: HD domain (PF01966); ResIII: type III restriction enzyme, restriction subunits (PF04851); CRISPR-associated protein, TIGR0 (TIGR03985); and Zot: zonula occludens toxin (Zot), (PF05707).

The same PDB hit was pulled by the *E. coli* and *M.ruber* amino acid sequences that were analyzed – 4Q2C: entity contains Chain A. Crystal Structure of CRISPR-associated helicase Cas 3 (Figure 14 A). A different protein hit (6C66: entity contains Chain G CRISPR RNA-guided surveillance complex) was obtained following the analysis of the mrub_1489 amino acid sequence (Figure 14 B). Pfam analysis of the amino acid sequences for *E. coli* b2761 and mrub_3020 yielded the same protein domain – Cas3 C-terminal domain (PF01966). IMG/M gene search was performed using mrub_1489 as the GOI. Two Pfam numbers were retrieved (PF00270: DEAD domain, PF13280: WYL domain). The DEAD domain serves as a helicase that unwinds nucleic acids (Aubourg *et al.,* 1999; de la Cruz *et al.*,1999). WYL domain is a negative regulator of the I-D CRISPR-Cas system in *Synechocystis sp* (Hein *et al.,* 2013). All the reported findings are the top hits because they had high score numbers and E-values below the cut-off (0.001).

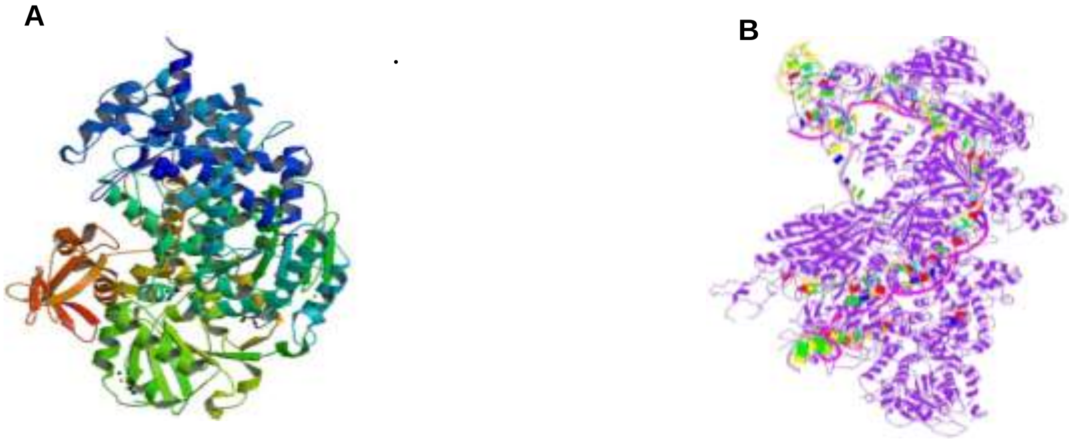
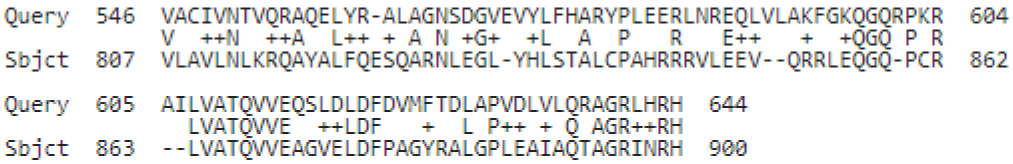


Figure 14. The crystal structure of the CRISPR-associated helicase Cas 3 (A) and the CRISPR RNA-guided surveillance complex (B). The different colored regions represent all the unique domains that interact to form the quaternary structure of the Cas 3 protein.

Protein BLASTp analysis of mrub_3020 as a query against the *Meiothermus ruber* DSM 1279 genome identified mrub_1489 (Metal dependent phosphohydrolase) as a potential paralog. This probable paralog is the top hit with an E-value of $7*10^{-5}$. Pairwise alignment of mrub_3020 amino acid sequence with that of mrub_1489 showed that 36% of the amino acids are identical (Figure 15).



```
Query  546  VACIVNTVQRAQELYR-ALAGNSDGVEVYLFHARYPLEERLNREQLVLAKFGKQGQRPKR  604
            V  ++N  ++A  L++ + A N +G+  +L  A  P   R   E++   +  +QGQ P R
Sbjct  807  VLAVLNLKRQAYALFQESQARNLEGL-YHLSTALCPAHRRRVLEEV--QRRLEQGQ-PCR  862

Query  605  AILVATQVVEQSLDLDFDVMFTDLAPVDLVLQRAGRLHRH  644
              LVATQVVE  ++LDF   +  L P++ + Q AGR++RH
Sbjct  863  --LVATQVVEAGVELDFPAGYRALGPLEAIAQTAGRINRH  900
```

Figure 15. Pairwise alignment data of mrub_3020 as the query and mrub_1489 as the subject. The amino acids between the query and the subject (sbjct) are the conserved amino acids. The (+) between the two sequences represent similar amino acids. 36% of the amino acids are identical.

A phylogenetic tree was generated to show the evolutionary relationship between mrub_3020 and mrub_1489. Only mrub_3020 shares a branch with other species like *M. rufus* and *M. taiwanensis*. Mrub_1489 appears distantly related to the other species because it shares no branch with other species.

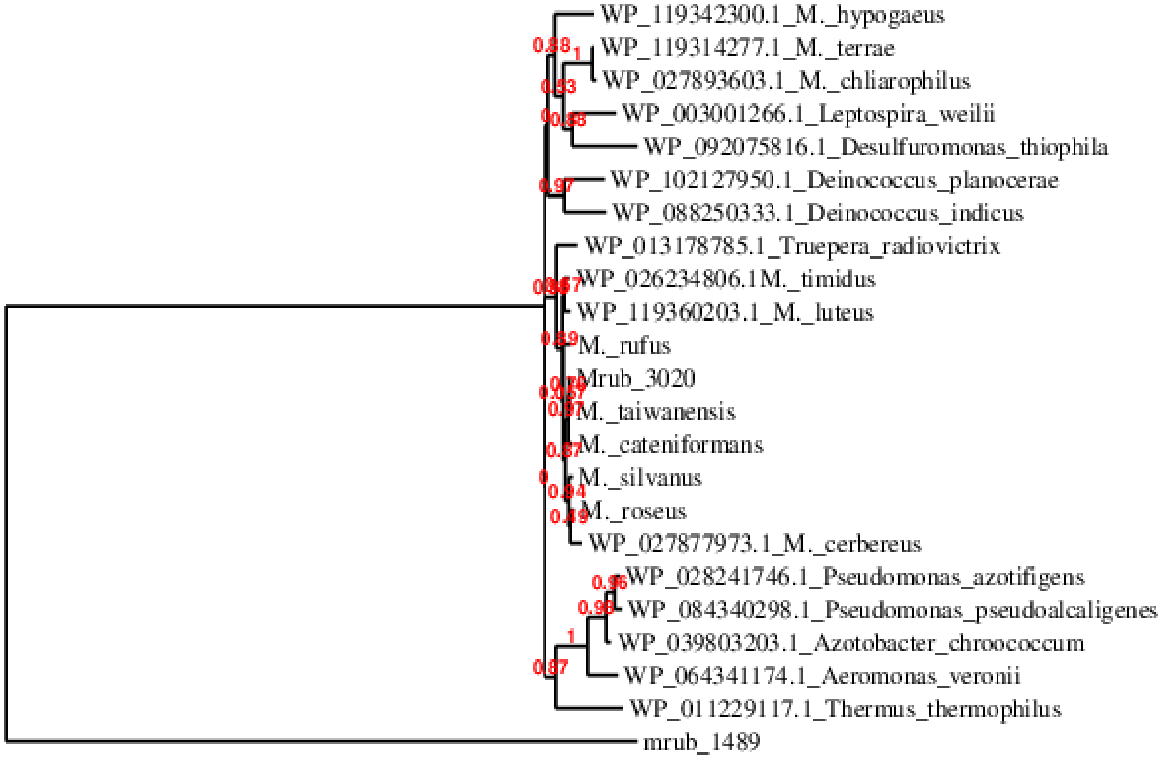

WP_119342300.1_M._hypogaeus
0.88
1
WP_119314277.1_M._terrae
0.53
WP_027893603.1_M._chliarophilus
0
0.88
WP_003001266.1_Leptospira_weilii
WP_092075816.1_Desulfuromonas_thiophila
0.97
WP_102127950.1_Deinococcus_planocerae
WP_088250333.1_Deinococcus_indicus
WP_013178785.1_Truepera_radiovictrix
WP_026234806.1M._timidus
WP_119360203.1_M._luteus
0.89
M._rufus
Mrub_3020
M._taiwanensis
0.97
M._cateniformans
0.87
M._silvanus
0
0.94
M._roseus
0.49
WP_027877973.1_M._cerbereus
0.96
WP_028241746.1_Pseudomonas_azotifigens
0.98
WP_084340298.1_Pseudomonas_pseudoalcaligenes
1
WP_039803203.1_Azotobacter_chroococcum
0.87
WP_064341174.1_Aeromonas_veronii
WP_011229117.1_Thermus_thermophilus
mrub_1489

Figure 16. A phylogenetic tree showing the evolutionary relationship between mrub_3020 and mrub_1489. Multiple sequence alignment between the different nucleotide sequences were done using T-coffee. The tree was generated using the maximum method on Phylogeny.fr and the numbers above are the maximized probability of the genetic data used to generate the tree. mrub_3020 shares a branch with other species like *M._rufus*, but mrub_1489 shares no branch with other species. *Meiothermus*, *Deinococcus*, *Thermus*, and *Truepera* are of the same phylum: *Deinococcus-Thermus* (Carbone).

Discussion.

The CRISPR/Cas system that includes mrub_3020 is similar to the model system. In a nutshell, the collection of bioinformatics tools used in this project all suggest that the *E. coli cas3* and the mrub_3020 are *cas3* orthologs and they are involved in the CRISPR/Cas systems as helicases. This conclusion was made by comparing *Meiothermus ruber* and *Escherichia coli* using the Prokaryotic Defense system in KEGG (Table 1), as well as the outputs of a collection of other bioinformatics tools. Both systems have genes, identified by their locus tags, that encode orthologs of the Type I-E CRISPR-Cas system, which includes the Cascade complex, the universal *cas1* and *cas2*, and the *cas3* signature gene. The CRISPR-Cas system in *M. ruber* is more complex, however. It appears to encode genes for additional CRISPR-Cas types.

There are compelling reasons to believe that mrub_3020 and mrub_1489 are paralogs. In support of this hypothesis is the observation that both genes are positioned adjacent to a likely Type I CRISPR-Cas operon, which is defined by its *cas3* signature gene. Without mrub_1489, there would be no *cas3* for its adjacent CRISPR-Cas system. A BLAST alignment between the two amino acid sequences produced an E-value well below the cut-off of 0.001. CDD, Pfam TIGRfam and PDB outputs all suggest a function within the CRISPR immune response for both proteins, usually as a helicase, which is the known function of Cas3. Interestingly, although the two paralogs were matched to different protein 3-D structures (PDB database), they are both predicted to serve as helicases and to be involved in the CRISPR/Cas system. The overwhelming evidence suggests that both proteins are localized to the cytoplasm.

The differences between these two proteins could be attributed to a duplication event that occurred in the distant past, followed by the acquisition of mutations within the mrub_1489 gene as its evolutionary constraints were reduced. For example, when compared to similar sequences drawn from GenBank, the amino terminus appears to be particularly variable (Figure 10B). We propose that the system containing mrub_3020 is the original CRISPR-Cas system because of its similarity to the *E. coli* CRISPR-Cas system. The phylogenetic tree generated to show the evolutionary relationship between mrub_3020 and mrub_1489 (Figure 16) is consistent with another tree generated using 6srRNA data of Deinococcus-Thermus phylum (Tindal *et al.*, 2010). In both trees, *Meiothermus ruber* or mrub_3020 is close to the same two species: *M. rufus* and *M. taiwanensis*. The key difference is that only the phylogenetic tree generated for this study includes the genus *Deinococcus*, which is in the same Deinococcus-Thermus phylum as *Meiothermus.* It is unlikely for a recent gene duplication to have resulted in the formation of the two paralogs. This is because a significant (based on the low E value), but smaller identity number of 36% was obtained following the NCBI BLAST alignment between mrub_3020 and mrub_1489 amino acid sequences (Figure 15). This lower identity number could imply that the gene duplication occurred well in the past.

Works Cited

1. Anders C, Niewoehner O, Duerst A, Jinek M. 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease [print]. Nature. 2014; 513:569–573. [PubMed: 25079318]
2. Aubourg S, Kreis M, Lecharny A. 1999. The DEAD box RNA helicase family in Arabidopsis thaliana [web]. Nucleic Acids Res. 27 (2): 628–36. doi:10.1093/nar/27.2.628
3. Bagos PG., Liakopoulos TD., Spyropoulos IC., and Hamodrakas SJ. 2014. PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins [web]. Nucleic Acids Res, 2004 Jul 1;32 https://www.ncbi.nlm.nih.gov/pubmed/15215419
4. Bagos PG, Liakopoulos TD, Spyropoulos IC and Hamodrakas SJ. 2004. A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins [web]. BMC Bioinformatics. https://www.ncbi.nlm.nih.gov/pubmed/15070403
5. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science [print]. 315:1709–1712. [PubMed: 17379808]
6. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes [print]. Science 315(5819):1709.
7. Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, et al. 2008. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats [print]. J. Biol. Chem. 283(29):20361-71.
8. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. 2000. The Protein Data Bank [web]. http://www.rcsb.org/.
9. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes [print]. Science. 321:960–964. [PubMed: 18703739]
10. Chylinski K, Makarova KS, Charpentier E, Koonin EV. 2014. Classification and evolution of type II CRISPR-Cas systems [print]. Nucleic Acids Res. 42:6091–6105. [PubMed: 24728998]
11. Carbone, Alessandra. Maximum Likelihood [web] [cited February 2019]. http://www.ihes.fr/~carbone/MaximumLikelihood2.pdf .
12. De la Cruz J, Kressler D, Linder P. 1999. Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families [web]. Trends Biochem. Sci. 24 (5): 192–8. doi:10.1016/S0968-0004(99)01376-6
13. Euzéby JP. 1997. List of bacterial names with standing in nomenclature: A folder available on the Internet [print]. Int J Syst Bacteriol. 47:590-592. PubMed doi:10.1099/00207713-47-2-590.
14. Ewen-Campen B, Mohr SE, Hu Y, Perrimon N. 2017. Accessing the phenotype gap: Enabling systematic investigation of paralog functional complexity with CRISPR [print]. Dev. Cell 43(1):6-9.
15. Fitch WS. 1970. Distinguishing homologous from analogous proteins [print]. Syst Zool.19:99-113.

16. 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 [web]. Nucleic Acids Res., 44:D279-D285. http://pfam.xfam.org/
17. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, et al. 2008. The pfam protein families database [web]. Nucleic Acids Res 36:D288. https://www.ncbi.nlm.nih.gov/pubmed/18039703.
18. Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA [print]. Nature. 468:67–71. [PubMed: 21048762].
19. García-Martínez J, Maldonado RD, Guzmán N,M., Mojica FJM. 2018. The CRISPR conundrum: Evolve and maybe die, or survive and risk stagnation [print]. Microbial Cell (Graz, Austria) 5(6):262-8.
20. GenBank Overview. [web] [cited 2019 February 5]. Available from: https://www.ncbi.nlm.nih.gov/genbank/.
21. Horvath P and Barrangou R. 2010. CRISPR/cas, the immune system of bacteria and archaea [print]. Science 327(5962):167.
22. Haft DH, Selengut J, Mongodin EF, Nelson KE. 2005. A guild of 45 CRISPR-associated (cas) protein families and multiple CRISPR/cas subtypes exist in prokaryotic genomes [print]. PLoS Computational Biology 1(6):e60.
23. Heler R, Marraffini LA, Bikard D. 2014. Adapting to new threats: the generation of memory by CRISPR-Cas immune systems [print]. Mol Microbiol. 93:1–9. [PubMed: 24806524]
24. 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 [print]. Nucleic Acids Res 29(1):41-3.
25. Hein S, Scholz I, Voss B, Hess WR 2013. Adaptation and modification of three CRISPR loci in two closely related cyanobacteria [web]. RNA Biol. 10:852-864. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3737342/
26. Hochstrasser ML, Taylor DW, Bhat P, Guegler CK, Sternberg SH, Nogales E, Doudna JA (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference [print]. Proc Natl Acad Sci U S A 111(18);6618-23. PMID: 24748111
27. Jansen, R., Embden, J. D., Gaastra, W. and Schouls, L. M. (2002), Identification of genes that are associated with DNA repeats in prokaryotes [print]. Molecular Microbiology, 43: 1565-1575. doi:10.1046/j.1365-2958.2002.02839.x
28. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity [print]. Science. 337:816–821. [PubMed: 22745249]
29. Jiang F and Doudna JA. 2015. The structural biology of CRISPR-cas systems [print]. Curr. Opin. Struct. Biol. 30:100-11.
30. Kunin V, Sorek R, Hugenholtz P. 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats [print]. Genome Biol 8(4):R61.

31. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. 2019. New approach for understanding genome variations in KEGG [web]. Nucleic Acids Res. 47, D590-D595. https://www.ncbi.nlm.nih.gov/pubmed/30321428
32. Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs [web]. Nucleic Acids Res. 45, D353-D361. https://www.ncbi.nlm.nih.gov/pubmed/27899662
33. Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes [web]. Nucleic Acids Res. 28, 27-30. Available from: https://www.kegg.jp/kegg/
34. Krogh A, Rapacki K. TMHMM Server, v. 2.0. Cbs.dtu.dk. 2016 [accessed 2016 Dec 6]. http://www.cbs.dtu.dk/services/TMHMM/.
35. 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 [web] Nucleic Acids Research 41: D605-612. https://www.ncbi.nlm.nih.gov/pubmed/23143106
36. Loginova LG, Egorova LA. 1975. Obligate thermophilic-bacterium Thermus ruber in hot springs of Kam-chatka. Mikrobiologiya [print]. 44:661-665.
37. Loginova LG, Egorova LA, Golovacheva RS, Sere-gina LM. 1984. Thermus ruber sp. nov., nom. rev. Int J Syst Bacteriol [print]. 34:498-499. doi:10.1099/00207713-34-4-498
38. Labedan B, Riley M. 1995. Widespread protein sequence similarities: origins of Escherichia coli genes [print]. J Bacteriol. 177:1585-1588.
39. Labedan B, Riley M. 1995. Gene products of Escherichia coli: sequence comparisons and common ancestries [print]. Mol Biol Evol 12:980-987.
40. Labedan B, Riley M. 1997. Protein evolution viewed through Escherichia coli protein sequence introducing the notion of a structural segment of homology, the module [print]. J Mol Biol. 268:857-868.
41. Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA [print]. Science. 322:1843–1845. [PubMed: 19095942]
42. Marraffini LA, Sontheimer EJ. 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea [print]. Nat Rev Genet. 11:181–190. [PubMed: 20125085]
43. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system [print]. Microbiology. 155:733–740. [PubMed: 19246744].
44. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV. 2011. Evolution and classification of the CRISPR-Cas systems [print]. Nat Rev Microbiol. 9:467–477. [PubMed: 21552286]
45. 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. 2016.CDD: NCBI's conserved domain database [web]. Nucleic Acids Res.28(43): D222-2: https://www.ncbi.nlm.nih.gov/pubmed/25414356?dopt=AbstractPlus

46. Mulepati S, Bailey S (2013). In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target [print]. J Biol Chem 288(31);22184-92. PMID: 23760266
47. Nobre MF, Trüper HG, Da Costa MS. 1996. Transfer of Thermus ruber (Loginova et al. 1984), Thermus silvanus (Tenreiro et al. 1995), and Thermus chlia-rophilus (Tenreiro et al. 1995) to Meiothermus gen. nov. as Meiothermus ruber comb. nov., Mei-othermus silvanus comb. nov., and Meiothermus chliarophilus comb. nov., respectively, and emendation of the genus Thermus.Int J Syst Bac-teriol [print]. 46:604-606. doi:10.1099/00207713-46-2-604
48. Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology [print]. 302 (1):205-17. http://www.ebi.ac.uk/Tools/msa/tcoffee/
49. 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 [print]. Bioinformatics 26(13):1608-1615
50. Rogers, R.L., Bedford, T., and Hartl, D.L. (2009). Formation and longevity of chimeric and duplicate genes in Drosophila melanogaster [print]. Genetics 181, 313–322.
51. Scott et al. Functional studies comparing Eschericia coli proC to putative ortholog Mrub_1345 [web]. https://digitalcommons.augustana.edu/biolmruber/43/
52. Sorek R, Kunin V, Hugenholtz P. 2008. CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea [print]. Nature Reviews Microbiology 6:181.
53. Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, Siksnys V. 2013. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus [print]. Embo J. 32:385–394. [PubMed: 23334296]
54. Tindall, B. J., Sikorski, J., Lucas, S., Goltsman, E., Copeland, A., Glavina Del Rio, T., … Lapidus, A. (2010). Complete genome sequence of Meiothermus ruber type strain (21T). Standards in Genomic Sciences [print]. 3(1), 26–36. http://doi.org/10.4056/sigs.1032748.
55. Van der Oost J, et al. 2009. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends in Biochemical Sciences [print]. 34(8):401-7.
56. Van der Oost J, Westra ER, Jackson RN, Wiedenheft B. 2014. Unravelling the structural and mechanistic basis of CRISPR-Cas systems [print]. Nat Rev Microbiol. 12:479–492. [PubMed: 24909109]
57. Wiedenheft B, Sternberg SH, Doudna JA. 2012. RNA-guided genetic silencing systems in bacteria and archaea [print]. Nature. 482:331–338. [PubMed: 22337052]
58. Westra ER, van Erp PB, Kunne T, Wong SP, Staals RH, Seegers CL, Bollen S, Jore MM, Semenova E, Severinov K, de Vos WM, Dame RT, de Vries R, Brouns SJ, van der Oost J (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3 [print]. Mol Cell 46(5);595-605. PMID: 22521689
59. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea [print]. Nature. 462:1056-1060. doi:10.1038/nature08656.
by Cascade and Cas3 [print]. Mol. Cell. 46:595–605. [PubMed: 22521689]