2018

Predicted ortholog pairs between *E. coli* and *M. ruber* are b3456 and mrub_2379, b3457 and mrub_2378, b3456 and mrub_2374, b3455 and mrub_2376, and b3454 and mrub2377, which each code for components of a prokaryotic-type ABC transporter for branched-chain amino acids

Elizabeth Paris
Augustana College, Rock Island Illinois

Tony Steinle
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Predicted ortholog pairs between *E. coli* and *M. ruber* are b3458 and mrub_2379, b3457 and mrub_2378, b3456 and mrub_2374, b3455 and mrub_2376, and b3454 and mrub2377, which each code for components of a prokaryotic-type ABC transporter for branched-chain amino acids

Elizabeth Paris and Tony Steinle
Dr. Lori R. Scott Laboratory
Biology Department, Augustana College
639 38th St., Rock Island, IL 61201

INTRODUCTION

Meiothermus Ruber

Meiothermus ruber is a species in the *Meiothermus* genus (Tindall et al., 2010). The species name derives from the Latin word "ruber" which means red. The organism has the title "ruber" due to its red cell pigmentation. *M. ruber* preferentially grows in high-temperature environments ranging from 35-70°C. The species was first isolated from a hot spring in Kamchatka Russian (Loginova *et al*.,1975). *M. ruber* is also a gram-negative bacterium. Compared to other strains of bacteria such as *E. coli* and *Salmonella*, both of which have over 30,000 publications, *M. ruber* only has 28 publications (Scott). Due to the lack of research on the organism, there is a lot of information missing about *M. ruber*. By studying lesser-known bacteria such as *M. ruber*, it may give scientists information about genes or different cellular processes in other organisms (Phylogenetic Diversity 2018).

Importance

Genome sequencing has transformed scientists' understanding of different microorganisms and the role they play in important processes (JGI). These processes include pathogenesis, energy production, bioremediation, global nutrient cycles, and many others. However, there is an imbalance in the phylogenetic distribution of known genome sequences. In other words, certain portions of the phylogenetic tree are studied more and have more genomes sequenced than others. This has created large gaps in information about microbial complexity and understanding of the evolution, physiology, and metabolic capacity of microbes. By researching a more diverse range of organisms, such as *M. ruber*, it could improve the identification and classification of protein families and ortholog groups across species, therefore strengthening the annotation of other microbial genes. It could also give the scientific community a better understanding of the

processes underlying evolutionary diversification among organisms and help with gene identification.

E. coli **as a Control**

To help fill in information gaps with understudied organisms, a model organism can be used (Cooper, 2000). In this specific project, *E. coli* was used as the model organism. *E. coli* is a great fit for this role because it is relatively easy to grow in the laboratory, and has frequently been studied so its entire genome has been sequenced. It is also a gram-negative protein like *M. ruber*. A BLAST search was performed where selected *E. coli* genes were BLAST'ed against *Meiothermus ruber* DSM 1279. There were 5 *E. coli* genes BLAST'ed, and each had a similar amino acid sequence to genes in the *Meiothermus ruber* DSM 1279 genome. This suggest the selected *E. coli* genes have orthologs in the *M. ruber* genome. Orthologs are genes in different species that evolved from a common ancestral gene by speciation (Koonin, 2005). Usually, orthologs retain the same function during the course of evolution, and therefore the same structure. If the structures of the two proteins are similar, their amino acid sequences should be too. Overall, *E. coli* is used as a control not only because it is easy to grow and has frequently been studied, but also because it contains genes that may be orthologous to genes in the *M. ruber* genome. *E. coli* is a Gram-negative bacillus native to the intestinal flora of many animals, including humans (Moussatova, 2008). The K-12 strain of *E. coli* is a non-virulent strain which does not have O and K antigens. It is also the most commonly used strain in laboratories. For this reason, it is referred to as the standard *E. coli* culture. The complete genome of the K-12 serotype was sequenced, and the largest single family of proteins in the *E. coli* K-12 genome is the ABC transporter family. This family accounts for 5% of the entire genome.

ABC Transportation

The 5 *E. coli* genes BLAST'ed against the *M. ruber* DSM 1279 genome all code for proteins that play a role in ABC transport, specifically branched- chain amino acid transport. ATP-binding cassette (ABC) transporters are integral membrane proteins that transport molecules across the lipid membrane of a cell, against the molecule's concentration gradient (Wilkens 2015). ABC transporters can do this using energy obtained from the hydrolysis of ATP into ADP (Moussatova, 2008). This class of transporters is present in nearly all living organism, including *E. coli* and *M. ruber*. These proteins belong to a very ancient family of transporters believed to have existed for over 3 billion years. There is phylogenetic evidence supporting the idea that the ABC transporter family diversified before archaea, eukaryotes and bacteria diverged on separate evolutionary paths.

All ABC transporters have a common basic structure regardless of what they transport across a membrane (Moussatova, 2008). The ABC transporters have two transmembrane domains (TMDs) which are integral membrane proteins. Along with this, they have two nucleotide binding domains, both of which are located on the same side of the membrane, but are

not integrated into the membrane (figure 1).

The TMD's form the transport channel through the membrane and consist of several membrane-spanning alpha-helices (Moussatova, 2008). The number of helices varies between 8-20 for importers, and 12 for exporters. The NBDs are highly conserved compared to the TMDs and are also the engines of the ABC transporter because they bind and hydrolyze ATP. The hydrolysis of ATP powers transport. When ATP binds to the NBDs, it induces a conformational change and forces the NBDs into closer contact, forming the characteristic nucleotide sandwich dimer. The changes experienced by the NBDs are transmitted to the TMDs, causing a conformational change that opens a section of the transmembrane channel to the inside or outside of the cell. After ATP hydrolysis, the structure returns to its original state.

There are two major classes of ABC transporters found in bacteria (Moussatova, 2008). The first is prokaryotic-type (PK-type), which are importers requiring additional extracellular proteins called substrate binding proteins (SBPs). Specifically, for Gram-negative bacteria, these proteins can be called periplasmic binding proteins (PBPs), because they are found in the periplasmic space between the inner cell membrane and outer cell membrane. The presence of SBPs determines the direction of transport. The other class of ABC transporter proteins is the eukaryotic-type (EK-type). This class of transporters are exporters, and move substances either from the cytoplasm out of the cell or from the cytoplasm into organelles.

Usually each component of PK-type ABC transporters is coded as a separate protein, which arises from an individual gene (Moussatova, 2008). Typically, the genes coding for a complete ABC transporter are found in a cluster of genes. For example, if there are five genes responsible for coding an ABC transporter, these five genes would be found in a gene cluster. The two TMD and two NBD domains may or may not be identical in a transporter. This means if a protein has two TMD domains, two NBD domains, and one substrate binding protein, there would be five genes necessary to code for an ABC transporter and therefore there would be a five-gene cluster. However, with a PD-type ABC transporter, there could be more than one substrate binding protein.

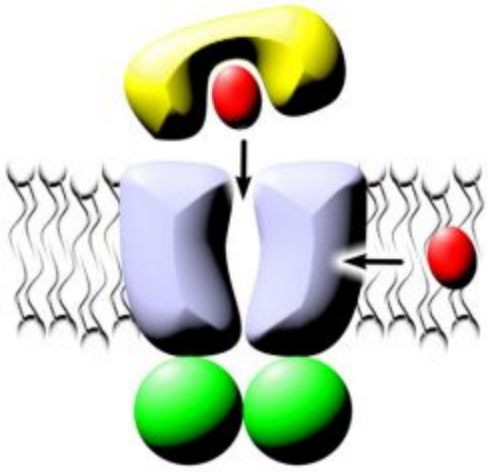


Figure 1. The figure shows the typical structure of a transmembrane protein. Represented in grey are the TMDs of the protein. The green spheres are the NBDs of the protein, and in yellow is the substrate binding protein (Moussatova, 2008).

Branched-chain Amino Acid Transport

For this study, a PK-type ABC transporter is described in *E. coli.* With this specific type of transporter, there are two TMDs and two NBDs as well as one substrate binding protein (Keseler IM et al., 2013). This means there are five genes that code for the PK-type ABC transporter. The transporter is used to move branched-chain amino acids across the membrane. The three branched-chain amino acids are leucine, valine and isoleucine. The *E. coli* genes involved in coding for the ABC transporter have the locus tags b3458, b3457, b3456, b3455, b3454 which code for the proteins livK, livH, livM, livG and livF respectively (figure 2). These five genes were observed to be in an operon with one another. This conclusion was made based on information gathered from the EcoCyc website. In figure 2 below, there are five genes shown and one promoter region. All genes are transcribed the same direction, and code for proteins involved in ABC transportation of branched-chain amino acids. This evidence supports the claim that the five genes are in an operon with one another. It would make sense for genes coding for proteins involved in ABC transportation to be in an operon so they could be regulated together. This ensures that for every one livF protein produced, there is one livG, one livM, one livH, and one livK protein produced.

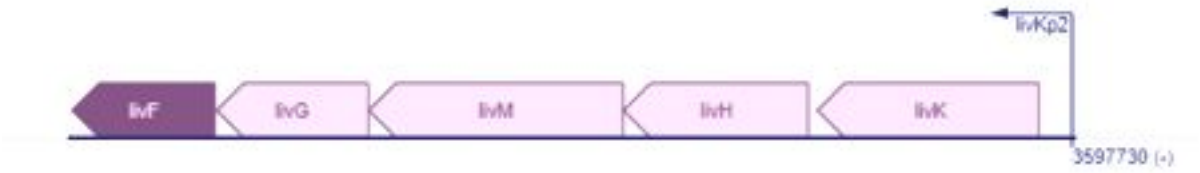


Figure 2. The image was obtained from EcoCyc and shows an operon with one promoter region, livKp2, and 5 genes, livK, livH, livM, livG and livF. The genes are all pointing the same direction, meaning they are transcribed the same way. They also possess the same color, which shows they code for proteins involved in the same process. Along with this, there is only one promoter region. These pieces of evidence strongly support the claim that these proteins are in an operon with one another (Keseler et al., 2013).

LivKHMGF is an ATP-dependent high-affinity branched-chain amino acid transport system, also referred to as the Liv-I system (Keseler et al., 2013). It is a member of the ABC superfamily of transporters. Liv-I is a common transporter of L-leucine, L-isoleucine and L-valine. Along with this, it is able to transport phenylalanine. LivF and LivG are the ATP-binding components of the ABC transporter complex, while LivH and LivM are the integral membrane proteins. LivK is the periplasmic binding protein. If a strain of *E. coli* is lacking LivK and unable to express LivHMGF, then it is unable to carry out high-affinity transport of leucine. Expression of LivKHMGF from a plasmid can restore high affinity leucine transport. According to EcoCyc, the liv genes are all a part of the same transcription unit.

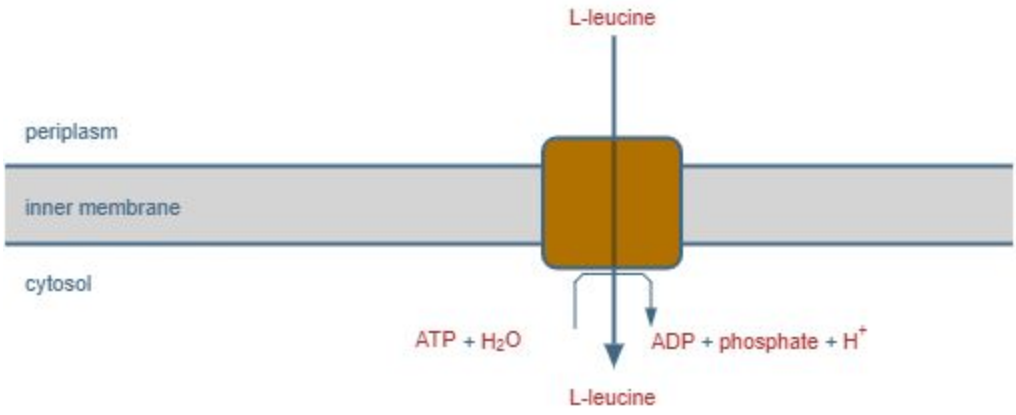


Figure 3. A branched-chain amino acid transporter moves L-leucine across the inner membrane (Keseler *et al.*, 2013). L-leucine, a branched chain amino acid, is shown to be in the periplasm. Once ATP binds to the NBD domains, the domains undergo a conformational change. This conformational change triggers

the TMDs to also undergo a conformational change, creating an opening on the periplasm side of the inner membrane. After this, ATP is hydrolyzed and the energy from this reaction is used to pump L-leucine into the cytosol of the cell.

Bioinformatics

Understanding how to use bioinformatics tools and knowing how to interpret their results is important because all careers in the biological sciences utilize bioinformatics tools to some extent (Persidis, 1999). The tools are available for free and can be efficient for those who know how to use them (Persidis, 1999).

Purpose/ Hypothesis

During this project, we use a variety of bioinformatics tools to determine if there are orthologs between the *E. coli* genome and the *M. ruber* genome. The hypothesis of this experiment is that b3458, b3457, b3456, b3455, b3454 from *E. coli* K-12 are orthologous to Mrub_2379, Mrub_2378, Mrub_2374, Mrub_2376, Mrub_2377 respectively. With the bioinformatics programs, we can determine similarities and differences between the *E. coli* genes coding for branched-chain amino transporter proteins and the *M. ruber* genes obtained through a BLAST search. To understand these programs, an understanding of E-values is crucial and how this determines the significance of results obtained through the programs. The lower the E-value, the stronger the evidence is.

METHODS

***M. ruber* genes have *E. coli* orthologs**

To confirm that each of the genes in the assigned *M. ruber* gene set are orthologous to the assigned genes found in *E. coli*, we performed a BLASTp (Altschul *et al.*, 1990; Madden, 2002) of each *E. coli* strain against the entire *M. ruber* genome and identifying the degree of similarity the strain had to its respective ortholog .

Correctly calling the start codons for the *M. ruber* genes

The start codons of each of the *E. coli* sequences is known because of the previous research available. To determine if the start codons of each *M. ruber* gene were called correctly, the same series of programs was used. First, the locus tag was entered into IMG/M and the alternate open reading frame viewer was examined (Markowitz *et al.*, 2012). Next, TCoffee (Notredame *et al.*, 2000) was used to create a multiple sequence alignment with strains obtained from a BLAST (Altschul *et al.*, 1990; Madden, 2002) search of the *M. ruber* amino acid sequence of interest. The resulting multiple sequence alignment was then put into the Weblogo program to create a colored Weblogo demonstrating the degree of conservation of amino acid residues throughout the sequence (Crooks *et al.*, 2004).

M. ruber **genes have comparable features to their** ***E. coli*** **orthologs**

Comparing the features of a given *M. ruber* gene to its *E. coli* ortholog required a series of programs to assess similarity of the cellular localization and family and domain names. The amino acid sequence of the gene of interest was entered into TMHMM (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998), SignalP (Petersen *et al.*, 2011), LipoP (Juncker *et al.*, 2003), PSORT-B (Yu *et al.*, 2010), and Phobius (Kall *et al.*, 2004; Kall *et al.*, 2007) programs for information of the location of the gene in the cell. Information about the families and domains of the genes was obtained by entering the same sequences into Pfam (Finn *et al.*, 2014; Finn *et al.*, 2016), TIGRfam (Haft *et al.*, 2001), BLAST/CDD (Marchler-Bauer *et al.*, 2015), and PDB (Berman *et al.*, 2000; Berman *et al.*, 2000) databases.

M. ruber **and** ***E. coli*** **genes are part of functional units**

To answer whether the *M. ruber* gene set and its orthologous *E. coli* gene set are each part of functional units or operons, the IMG/M chromosome map was utilized (Markowitz *et al.*, 2012). Chromosome maps were viewed as colored by Kegg and by top COG hit neighborhoods. Additional confirmation of the presence of an operon was obtained from EcoCyc for *E. coli* genes only (Keseler *et al.*, 2013).

RESULTS

E. coli **gene** ***b3454*** **and** ***M. ruber*** **gene** ***Mrub_2377***

The first step in this research project was BLASTing the first gene in the set, b3454, against the *M. ruber* genome (Altschul *et al.*, 1990; Madden, 2002). This was done to see if *M. ruber* had any potential orthologs to the b3454 gene. The results from the BLAST test are recorded in figure 4, including the pairwise alignment between the two sequences. It is important to note that during the BLAST result, Mrub_2377 was not the first hit, however, the hits with lower E-values than Mrub_2377 were found to be paralogs. The E-value found from the BLAST result is 2e-52, which is a very low E-value, showing the two genes have a high degree of similarity between them. Due to the fact the E-value is so low, it is promising these two genes could be orthologs, but further test should be done to confirm this.

Score	Expect	Method	Identities	Positives	Gaps
167 bits(424)	2e-52	Compositional matrix adjust.	93/230(40%)	136/230(59%)	6/230(2%)

```
Query  14   HYGKIQALHEVSLHINQGEIVTLIGANGAGKTTLLGTLCG-----DPRATSGRIVFDDKD  68
            + G I AL  VS+    GE V L+G NGAGK+TL+  + G     D R   GRI    +D
Sbjct  13   YRGVILALQGVSMRAGAGEAVALLGPNGAGKSTLVRAISGLLPQYDGRVLDGRITLGGED  72

Query  69   ITDWQTAKIMREAVAIVPEGRRVFSRMTVEENLAMGGFFAERDQFQERIKWVYELFPRLH  128
            I+      K+    +  + EGR +F  +TV ENL   G       + +E    ++  FPRL+
Sbjct  73   ISHLPALKVAGLGLTAILEGRPIFRYLTVIENLRAAGHKLTPQRQKELTDEIFTRFPRLY  132

Query  129  ERRIQRAGTMSGGEQQMLAIGRALMSNPRLLLLDEPSLGLAPIIIQQIFDTIEQL-REQG  187
            ERR ++ G +SGGEQQML +G AL++ PR+L++DEPSLGL+P + +++    +++L R++G
Sbjct  133  ERRFEQGGYLSGGEQQMLLLGMALLTEPRILVVDEPSLGLSPKLTEEVMRVLDELRRDKG  192

Query  188  MTIFLVEQNANQALKLADRGYVLENGHVVLSDTGDALLANEAVRSAYLGG  237
            +T+ LVEQNA   A  + +R YV+E G VV    T     A+  V   YLGG
Sbjct  193  LTLVLVEQNARAAFSIVERVYVMEQGRVVFEGTAQEAQADADVMEFYLGG  242
```

Figure 4. Mrub_2377 is the "Sbjct" sequence and b3454 is the "Query" sequence. Analysis was performed using the NCBI BLAST bioinformatics tool at http://www.ncbi.nlm.nih.gov (Altschul *et al.*, 1990; Madden, 2002).

The next couple of tests run were used to determine if the proper start codon was called. This is necessary because an incorrectly called start codon may lead to an inaccurate alignment between the *E. coli* and *M. ruber* sequences. By using the programs IMG/M (Markowitz *et al.*, 2012), T-Coffee (Notredame *et al.*, 2000) and WebLogo(Crooks et al., 2004), we analyzed if the proper start codon was called. The IMG Sequence tool used to analyze alternate start codons listed the predicted start codon as a methionine amino acid in the first reading frame, approximately 10 amino acids downstream from a potential Shine-Dalgarno sequence (Markowitz *et al.*, 2012). While it is not necessary to rely on the Shine-Dalgarno sequence for *M. ruber* genes of interest, the fact the highlighted methionine starts in RF1 is a good indication of a correctly called start codon. For the T-Coffee analysis, the only sequences beginning with M are those that belong to the *Meiothermus* genus, even though the species vary (Notredame *et al.*, 2000). Sequences belonging to a different genus are still similar to *M. ruber* throughout the sequence, but they have different start codons. This result is then exhibited in the Weblogo created from the TCoffee multiple sequence alignment (Crooks et al., 2004); the starting methionine is not highly conserved in the image because half of the sequences chosen do not belong to the genus *Meiothermus* and do not share the same start codon. However, it is still logical to conclude that the start codon was called correctly for mrub_2377 since the start codon is consistent among the species.

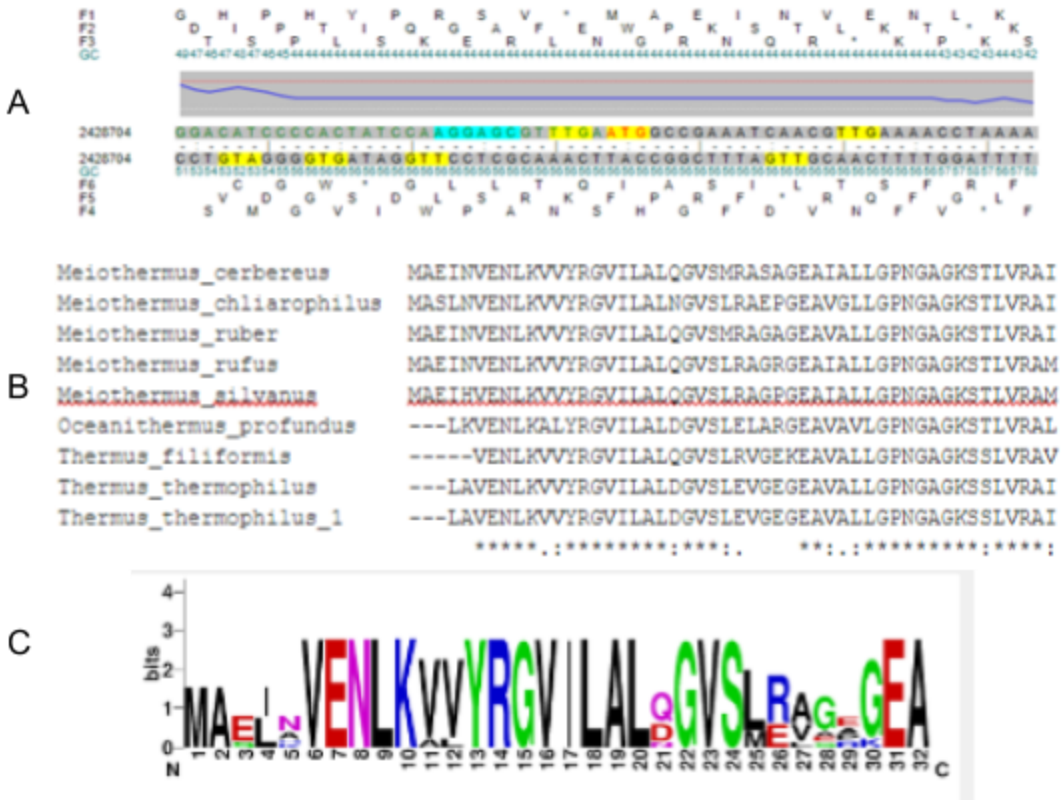


Figure 5. The start codon of Mrub_2377 is correctly called. Panel A: IMG/M alternate ORF viewer with the suggested start codon highlighted in red (Markowitz *et al.*, 2012); Panel B: first line of TCoffee MSA (Notredame *et al.*, 2000); Panel C: first line of Weblogo created from MSA (Crooks et al., 2004). The programs used are described in the Methods section.

To properly analyze the the structure and function of b3454 and Mrub_2377, a variety of bioinformatics tools were used. The table below summarizes the results from the programs used by providing any pertinent family names and numbers as well as E-values. It is important to analyze the E-values of each output to confirm that the gene are related and that the alignment is not due to random chance.

Bioinformatics Tool Used	***E. coli b3454***	***M. ruber Mrub_ 2377***
BLAST *E. coli* against *M. ruber*	Score: 167 bits (424) E-value:2e-52	
CDD Data (COG category)	COG Number: COG0410 LivF	

	E-value: 5.11e-136	E-value: 5.84e-91
Cellular Localization	Cytoplasm	
TIGRfam (protein family)	TIGR03410 (urea trans UrtE) TIGR03411 (urea trans UrtD)	
	E-value: 1.6e-59 E-value: 3.0e-17	E-value: 1.5e-29 E-value: 2.8e-17
Pfam (protein family)	PF00005 (ABC Transporter)	
	E-value: 3.3e-33	E-value: 5.1e-22
Protein Database (PDB)	1JI0 Crystal Structure Analysis of the ABC transporter from *Thermotoga maritima*	
	E-value: 2.66694e-66	E-value: 7.23404E-43
KEGG Pathway Map	Prokaryotic-type ABC Transporters (02010)	

Table 1 summarizes the results from an assortment of bioinformatics tools that were used to compare *E. coli* b3454 to Mrub_ 2377. The first row of data shows the results of the initial BLAST analysis performed by BLASTing the amino acid sequence of *E. coli* b3454 to the *M. ruber* genome (Altschul *et al.*, 1990; Madden, 2002). The results yielded a low E-value of 2e-52, which means the sequences of the proteins are fairly similar to one another and the two sequences do not align because of chance. The BLAST test provides evidence the two genes might be orthologous to one another. The CDD search gave the same COG number (COG0410) and name (Liv F) from the database (Marchler-Bauer *et al.*, 2015). For both genes, the E-values were extremely small, indicating significance. This is a strong indication the genes code for the same protein involved in branched-chain amino acid transport, which is a NMD protein. All the bioinformatics tools used to analyze cellular location (TMH (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998), SignalP (Petersen *et al.*, 2011), LipoP (Juncker *et al.*, 2003), Phobius (Kall *et al.*, 2004; Kall *et al.*, 2007) and PSORT-B (Yu *et al.*, 2010)) suggested both proteins are located in the cytoplasm. These tools also showed an absence of a cleavage site and signal peptide sequence, which makes sense because the protein would not have to cross any membranes if it is on the cytoplasmic side of the cytoplasmic membrane. The two genes have the same cellular location, which further supports the idea they are orthologs. The TIGRfam number obtained for both protein sequences was the same, TIGR03410, which is named urea trans UrtE (Haft *et al.*, 2001). The E-values for both genes were very low indicating a strong match. When the sequences were run on Pfam (Finn *et al.*, 2014; Finn *et al.*, 2016), the first and only hit was the ABC Transporter family (PF00005). The low E-values associated with this hit and the fact it is the ABC transporter family strongly suggests the genes

are orthologs coding for proteins involved in ABC transportation. For PDB, both Mrub_ 2377 and b3454 yielded the result of 1JI0 Crystal Structure Analysis of the ABC transporter from *Thermotoga maritima* (Berman *et al.*, 2000; Berman *et al.*, 2000). The fact that both genes showed the same top PDB hit with fairly low E-values is strong evidence for their orthologous relationship. Finally, both groups were predicted to be Prokaryotic-type ABC Transporters, further suggesting the two genes are orthologs (Kanehisa *et al.*, 2016).

After confirming the correct start codon was called for Mrub_2377, a series of test were run to determine the cellular location of the proteins coded from b3454 and Mrub_2377. Figure 6 shows the results of each program for the *E. coli* sequence b3454, and figure 7 represents the result of the same programs used but for the Mrub_2377 sequence. From the TMHMM graphs, for both the b3454 and Mrub_2377, there is an absence of transmembrane helices (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998). This makes sense because b3454 is one of the NBDs. NBDs are located on in the cytoplasm and are not integrated into the inner membrane. If Mrub_2377 is an ortholog to b3454, it would be an NBD domain and therefore would be expected to be found in the cytoplasm. The Phobius graphs for both the genes did not have any transmembrane helices present (Kall *et al.*, 2004; Kall *et al.*, 2007). This makes sense because the NBDs are not transmembrane proteins, therefore they should not have any transmembrane helices. For both genes, the SignalP graphs predicted zero signal peptides (Petersen *et al.*, 2011). This makes sense because a signal peptide sequence is needed for proteins to cross a membrane, and because NBDs are located in the cytoplasm, they do not cross any membranes and therefore do not need a signal peptide sequence. LipoP also predicted the absence of a signal peptide sequence for both genes (Juncker *et al.*, 2003). The PSORT-B test however did have conflicting results (Yu *et al.*, 2010). It predicted the protein made from b3454 would be found in the cytoplasm, giving it a score of 9.12. However, for the Mrub_2377 gene, its protein was predicted to be found in the cytoplasmic membrane. This seems contradictory to the other results that predicted Mrub_2377 to have 0 transmembrane helices, which would seem illogical for a cytoplasmic membrane protein to have. Due to the evidence supporting the location of the Mrub_2377 protein being in the cytoplasm despite the result from PSORT-B, it is confident to say the genes from *M. ruber* and *E. coli* are localized to the same location in the cell, which further supports the hypothesis that the two genes are orthologs.

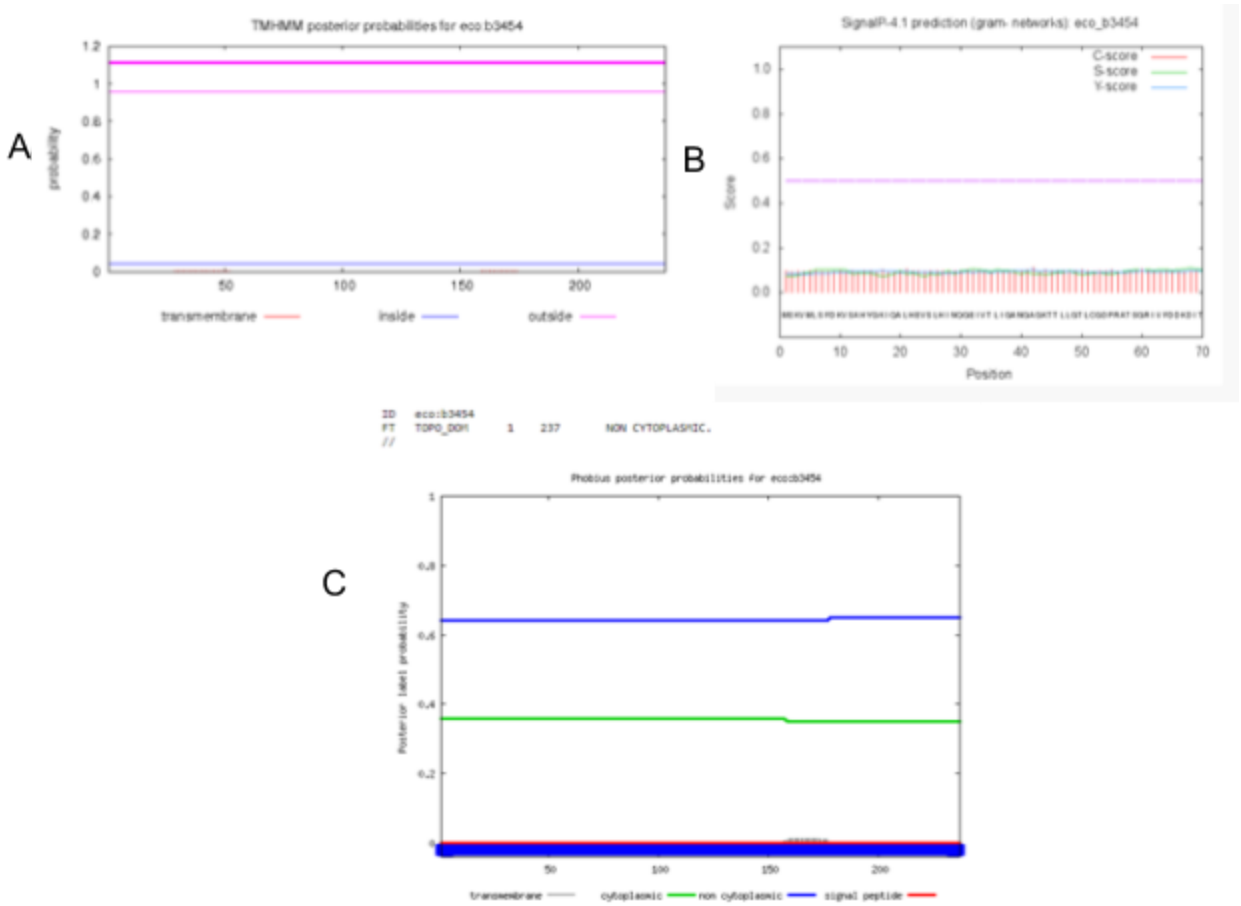


Figure 6. Cellular location determination of b3454. Panel A: TMHMM shows zero transmembrane helices because there are not any red peaks present (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); Panel B: SignalP shows the absence of a signal peptide because the there are no peaks past the central cut-off line (Petersen *et al.*, 2011); Panel C: Phobius shows transmembrane helices in gray peaks, where the gene is cytoplasmic in green, and where the gene is non-cytoplasmic in blue (Kall *et al.*, 2004; Kall *et al.*, 2007). The absence of grey peaks means there are not any transmembrane helices. The bioinformatics tools used are described in Methods.

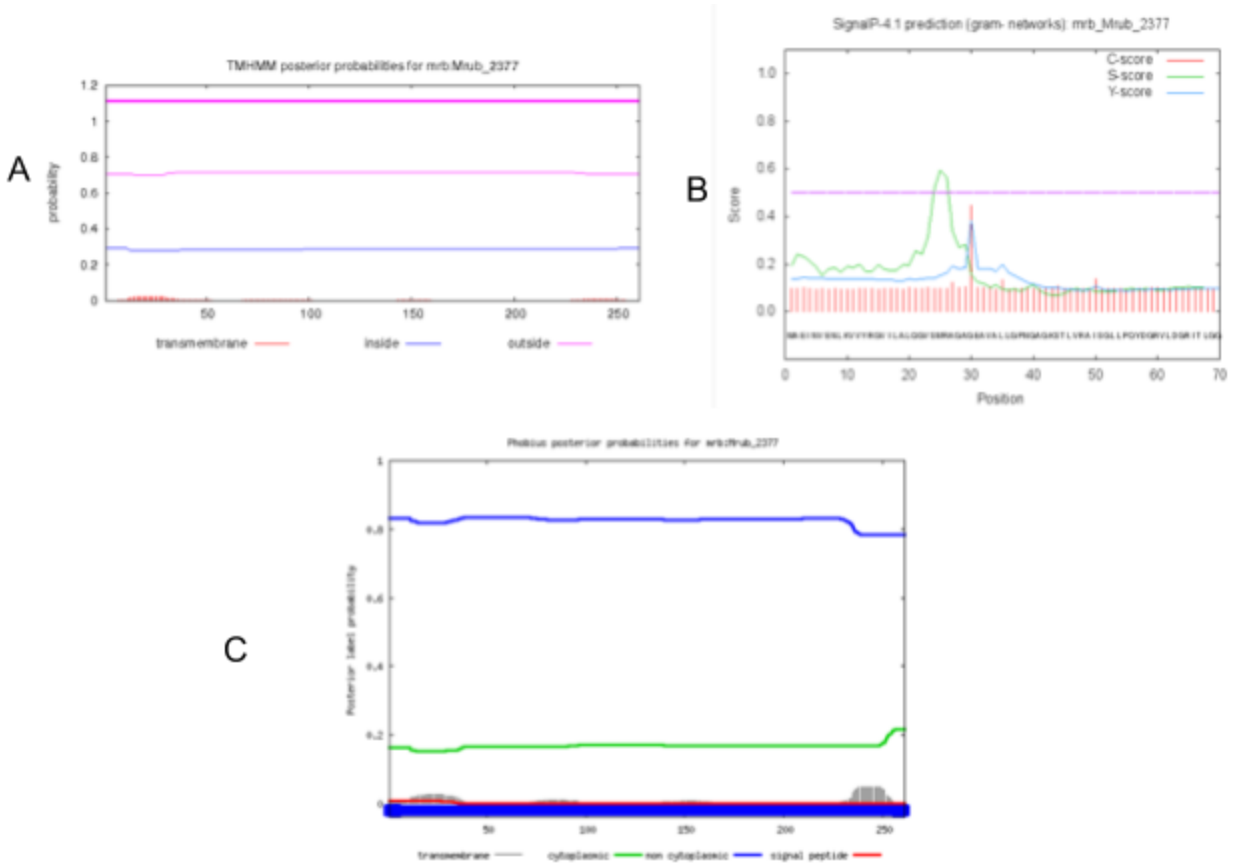


Figure 7. Cellular location determination of Mrub_2377. Panel A: TMHMM shows the lack of transmembrane helices because there are not any red peaks present (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); Panel B: SignalP shows the lack of a signal peptide because the there are no peaks past the central cut-off line (Petersen *et al.*, 2011); Panel C: Phobius shows transmembrane helices in gray peaks, where the gene is cytoplasmic in green, and where the gene is non-cytoplasmic in blue (Kall *et al.*, 2004; Kall *et al.*, 2007). The absence of grey peaks means there are not any transmembrane helices. The bioinformatics tools used are described in Methods.

The Pfam test (Finn *et al.*, 2014; Finn *et al.*, 2016) showed both b3454 and Mrub_2377 belong to the PF00005 (ABC Transporter) group. Alignments were also obtained from the test and are shown in figure 8. Unlike the BLAST alignment, the Pfam alignment is a pairwise alignment that compares the sequences of both b3454 and Mrub_2377 to a consensus sequence obtained from hundreds of other proteins. Both b3454 and Mrub_2377 were matched to the same consensus sequence. Both b3454 and Mrub_2377 have several glycine residues conserved with the consensus sequence along with a lysine, leucine, glutamic acid and aspartic acid residue. The commonality of shared amino acid residues with the consensus sequence between b3454 and Mrub_2377 is another piece of evidence supporting the hypothesis that the two gene are orthologous to one another.

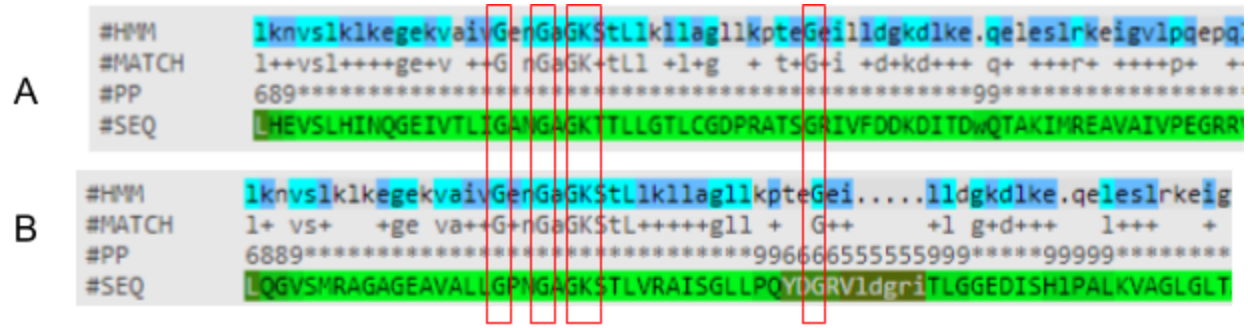


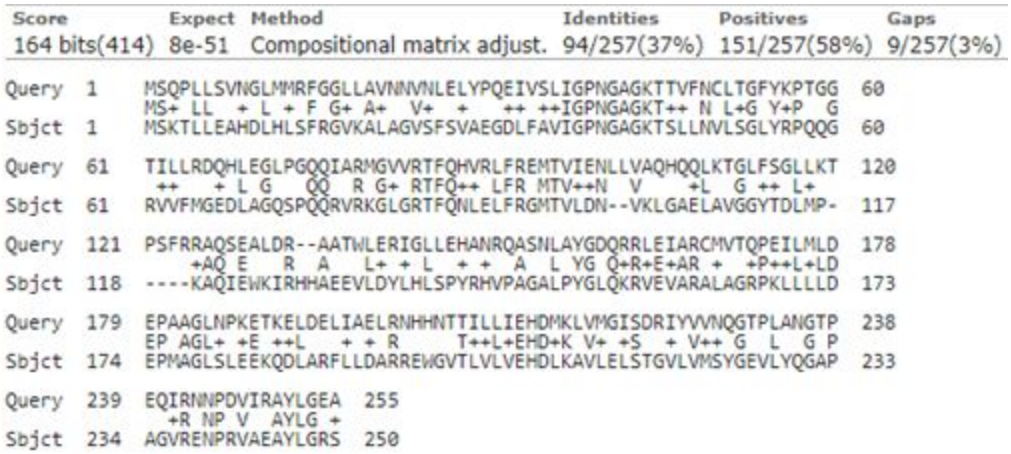
Figure 8. Panel A shows the alignment between b3454 and the consensus sequence (Finn *et al.*, 2014; Finn *et al.*, 2016); Panel B shows the alignment between Mrub_2377 and the consensus sequence. Conserved amino acids are written in capital letters in the "#MATCH" line. Both b3454 and Mrub_2377 have multiple glycine residues conserved with the consensus sequence as well as a lysine residue. The red boxes surround the conserved amino acids. The #HMM line is the consensus sequence and the #SEQ line is the gene being analyzed (either b3454 or Mrub_2377). The pairwise alignment was produced by the Pfam website http://pfam.sanger.ac.uk/search.

In addition to the consistency between cellular localization data, support of the orthologous relationship between b3454 and Mrub_2377 can be observed through their familial similarity exhibited by the names and E-values presented in Table 1. Further support is available by confirming that both genes are parts of operons and are involved in the same molecular pathway. Both b3454 and Mrub_2377 belong to the branched-chain amino acid category and are each part of a 5-gene operon. The IMG/M Color by Kegg feature presented clear images of each gene within its operon and in relation to the flanking regions upstream and downstream (Markowitz *et al.*, 2012). The presence of an operon is indicated by the same color identifier and direction of transcription. The fact that b3454 and Mrub_2377 are part of operons and within the same biochemical pathway are strong indications that the genes are orthologous.

Figure 9. Both b3454 and Mrub_2377 exist as units of distinct operons (Markowitz *et al.*, 2012). Panel A: The Color by Kegg Chromosome Map viewer of the area surrounding b3454 with the GOI indicated by the red dash; Panel B: The output of the same program for the area surrounding Mrub_2377.

E. coli* gene *b3455* and *M. ruber* gene *Mrub_2376

The first step in this research project was BLASTing b3455 against the *M. ruber* genome (Altschul *et al.*, 1990; Madden, 2002). This step was done to see if *M. ruber* had any potential orthologs to the b3455 gene. The results from the BLAST test are recorded in figure 10, including the pairwise alignment between the two sequences. It is important to note that Mrub_2376 was not the first hit obtained from the BLAST search, however, the hits with lower E-values than Mrub_2376 were found to be paralogs. The E-value found from the BLAST result is 8e-51, which is a very small E-value, showing the two genes have a high degree of similarity between them. The low E-value shows these two could be orthologs, but further tests should be done to confirm this.



Score	Expect	Method	Identities	Positives	Gaps
164 bits(414)	8e-51	Compositional matrix adjust.	94/257(37%)	151/257(58%)	9/257(3%)

```
Query  1    MSQPLLSVNGLMMRFGGLLAVNNVNLELYPQEIVSLIGPNGAGKTTVFNCLTGFYKPTGG  60
            MS+ LL   + L + F G+ A+   V+  +    ++ ++IGPNGAGKT++ N L+G Y+P  G
Sbjct  1    MSKTLLEAHDLHLSFRGVKALAGVSFSVAEGDLFAVIGPNGAGKTSLLNVLSGLYRPQQG  60

Query  61   TILLRDQHLEGLPGQQIARMGVVRTFQHVRLFREMTVIENLLVAQHQQLKTGLFSGLLKT  120
             ++   + L G   QQ  R G+ RTFQ++ LFR MTV++N  V    +L  G ++ L+
Sbjct  61   RVVFMGEDLAGQSPQQRVRKGLGRTFQNLELFRGMTVLDN--VKLGAELAVGGYTDLMP-  117

Query  121  PSFRRAQSEALDR--AATWLERIGLLEHANRQASNLAYGDQRRLEIARCMVTQPEILMLD  178
               +AQ E   R  A   L+ + L  + +  A  L YG Q+R+E+AR +  +P++L+LD
Sbjct  118  ----KAQIEWKIRHHAEEVLDYLHLSPYRHVPAGALPYGLQKRVEVARALAGRPKLLLLD  173

Query  179  EPAAGLNPKETKELDELIAELRNHHNTTILLIEHDMKLVMGISDRIYVVNQGTPLANGTP  238
            EP AGL+ +E ++L   + + R     T++L+EHD+K V+ +S  + V++ G  L  G P
Sbjct  174  EPMAGLSLEEKQDLARFLLDARREWGVTLVLVEHDLKAVLELSTGVLVMSYGEVLYQGAP  233

Query  239  EQIRNNPDVIRAYLGEA  255
              +R NP V  AYLG +
Sbjct  234  AGVRENPRVAEAYLGRS  250
```

Figure 10. Mrub_2376 is the "Sbjct" sequence and b3455 is the "Query" sequence. Analysis was performed using the NCBI BLAST bioinformatics tool at http://www.ncbi.nlm.nih.gov (Altschul *et al.*, 1990; Madden, 2002).

The next couple of tests run were used to determine if the proper start codon was called. This is necessary because an incorrectly called start codon may lead to an inaccurate alignment between the *E. coli* and *M. ruber* sequences. By using the programs IMG/M (Markowitz *et al.*, 2012), T-Coffee (Notredame *et al.*, 2000) and WebLogo (Crooks et al., 2004), we analyzed whether or not the proper start codon was called. The alternate open reading frame program through IMG/M provided a sequence having a start codon of methionine in the first reading frame, approximately 10 positions downstream from the potential Shine-Dalgarno sequence (Markowitz *et al.*, 2012). There were no other options for a start codon in the template strand. These results support the idea the correct start codon was called, because start codons are usually 10 positions downstream of the Shine-Dalgarno sequence and start with methionine. To go along with this, there is only one potential start codon, so the one called is the only option and therefore should be the correct option. The multiple sequence alignment (MSA) was created from selected amino acid

sequences similar to Mrub_2376 and the subsequent Weblogo supports this claim (Crooks et al., 2004). At first glance, the MSA and Weblogo are not supportive of the claim the correct start codon was called because the starting M residue is not highly conserved among the organisms (figure 3B). However, this is because only species belonging to the Meiothermus genome conserved the start codon and the species in the Thermus genome began with a different start codon. After the start codon, the remainder of the alignment and Weblogo show a majority of the amino acids are highly conserved among the species. Looking at all the data, we can conclude the start codon was called correctly.

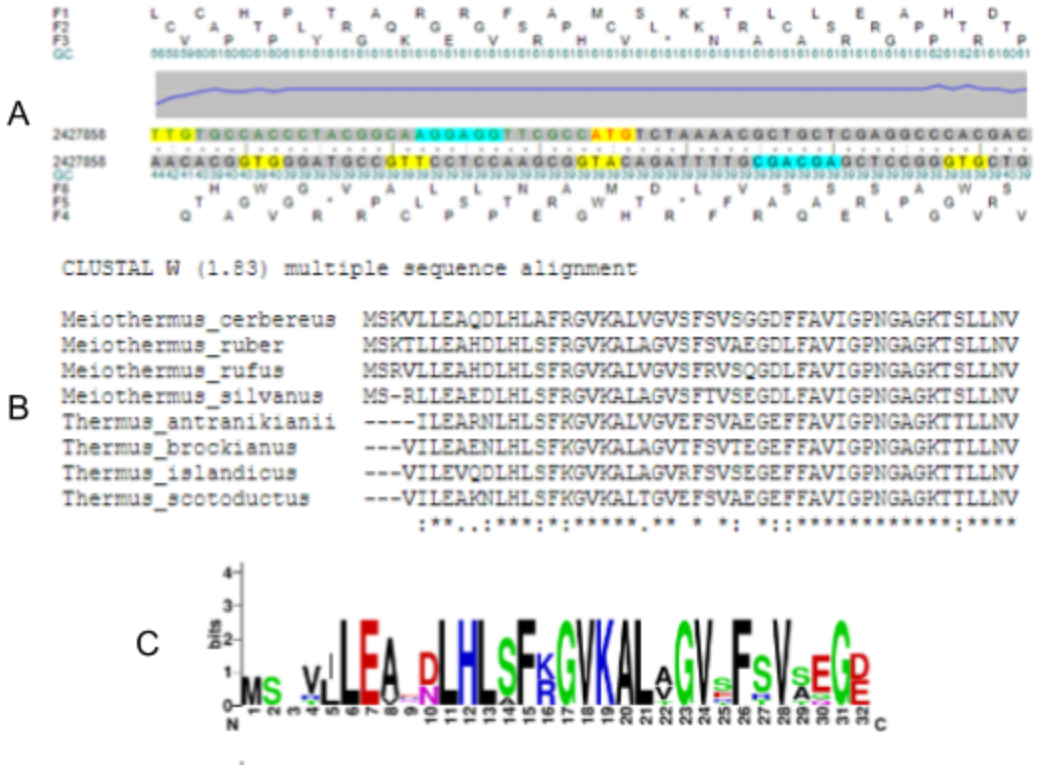


Figure 11. The start codon of Mrub_2376 is correctly called. Panel A: IMG/M alternate ORF viewer with the suggested start codon highlighted in red (Markowitz *et al.*, 2012); Panel B: first line of TCoffee MSA (Notredame *et al.*, 2000); Panel C: first line of Weblogo created from MSA (Crooks et al., 2004) . The programs used are described in the Methods section.

For the structure and function of b3455 and Mrub_2376, the same bioinformatics tools were used. The table below summarizes the results from the programs used, any pertinent family names and numbers, as well as E-values. It is important to analyze the E-values of each output to confirm that the gene are related and that the alignment is not due to random chance.

Bioinformatics Tool Used	***E. coli b3455***	***M. ruber Mrub_ 2376***
BLAST *E. coli* against *M. ruber*	Score: 164 bits (414) E-value: 8e-51	
CDD Data (COG category)	COG Number: COG0411 COG name: Liv G	
	E-value: 1.61e-139	E-value: 4.74e-102
Cellular Localization	Cytoplasm	
TIGRfam (protein family)	TIGR03411(urea trans UrtD)	
	E-value: 1.2e-64	E-value: 4e-49
Pfam (protein family)	PF00005 (ABC Transporter) PF12399 (Branched-chain amino acid ATP-binding cassette transporter)	
	E-value:3.1e-31 E-value: 1.9e-10	E-value: 1.7e-29 E-value: 3.2e-05
Protein Database	5L75 A protein structure	
	E-value:4.97653e-32	E-value: 1.45152e-28
KEGG Pathway Map	Prokaryotic-type ABC Transporters (02010)	

Table 2 summarizes the results from an assortment of bioinformatics tools that were used to compare *E. coli* b3455 to Mrub_ 2376. The first row of data shows the results of the initial BLAST analysis performed by BLASTing the amino acid sequence of *E. coli* b3455 to the *M. ruber* genome (Altschul *et al.*, 1990; Madden, 2002). The results yielded a low E-value of 8e-51, which means the sequences of the proteins are fairly similar to one another and the two sequences do not align because of chance. The BLAST test provides evidence the two genes might be orthologous to one another. The CDD search gave the same COG number (COG0411) and name (Liv G) for both genes (Marchler-Bauer *et al.*, 2015). The E-values were also extremely small for both genes, indicating significance. This is a strong indication the genes code for the same protein involved in branched-chain amino acid transport. All the bioinformatics tools used to analyze cellular location (TMH (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998), SignalP (Petersen *et al.*, 2011), LipoP (Juncker *et al.*, 2003), Phobius (Kall *et al.*, 2004; Kall *et al.*,

2007) and PSORT-B(Yu *et al.*, 2010)) suggested that both proteins are found in the cytoplasm. These tools also showed there an absence of a cleavage site and signal peptide sequence, which makes sense because the protein would not have to cross any membranes if it is in the cytoplasm. From the results, it is concluded the two genes have the same cellular location, which further supports the idea they are orthologs. The TIGRfam number obtained for both protein sequences was the same, TIGR03411, which is named urea trans UrtD (Haft *et al.*, 2001). The E-values for both genes were very low, indicating a strong match. When the sequences were run on Pfam, many hits were obtained (Finn *et al.*, 2014; Finn *et al.*, 2016). The top two are recorded in the table and are the same for both b3455 and Mrub_ 2376. The first Pfam hit is the ABC Transporter family (PF00005). The low E-values associated with this hit and the fact it is the ABC transporter family strongly suggests the genes are orthologs coding for proteins involved in ABC transportation. For the protein data base both genes had the hit 5L75, a protein structure, which had a relatively low E-value for both genes (Berman *et al.*, 2000; Berman *et al.*, 2000). Finally, both groups were predicted to be Prokaryotic-type ABC transporters, further suggesting the two genes are orthologs.

After confirming the correct start codon was called for Mrub_2376, a series of test were run to determine the cellular location of the proteins coded from b3455 and Mrub_2376. Figure 12 shows the results of each program for the *E. coli* sequence, b3455, and figure 13 represents the result of the same programs used but for the Mrub_2376 sequence. Looking at the TMHMM graphs, for both the b3455 and Mrub_2376, there is an absence of transmembrane helices (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998). This makes sense because b3455 is one of the NBD domains which are located on in the cytoplasm. If a protein is in the cytoplasm, it does not have transmembrane domains. If Mrub_2376 is an ortholog to b3455, it would be an NBD domain and therefore would be expected to be found in the cytoplasm. The Phobius graphs for both the genes did not have any transmembrane helices present (Kall *et al.*, 2004; Kall *et al.*, 2007). This makes sense because the NBD domains are not transmembrane proteins, therefore they should not have any transmembrane helices. For both genes, the Signal P graphs predicted that there are not any signal peptides (Petersen *et al.*, 2011). This makes sense because a signal peptide sequence is needed for proteins that cross a membrane, and because NBD domains are located in the cytoplasm, they do not cross any membranes and therefore do not need a signal peptide sequence. LipoP also predicted the absence of a signal peptide sequence for both genes (Juncker *et al.*, 2003). The PSORT-B test however did have different results for the genes (Yu *et al.*, 2010). It predicted the protein made from b3455 would be found in the cytoplasm, giving it a score of 9.12. However, for the Mrub_2376 gene, its protein is predicted to be found in the cytoplasmic membrane. This seems contradictory to the other results that predicted Mrub_2376 has 0 transmembrane helices, which would seem illogical for a cytoplasmic membrane protein to have. Due to the evidence supporting the location of the Mrub_2376 protein being in the cytoplasm despite the result from PSORT-B, it is confident to say the genes from *M. ruber* and *E. coli* are localized to the same location in the cell, which further supports the hypothesis that the two genes are orthologs.

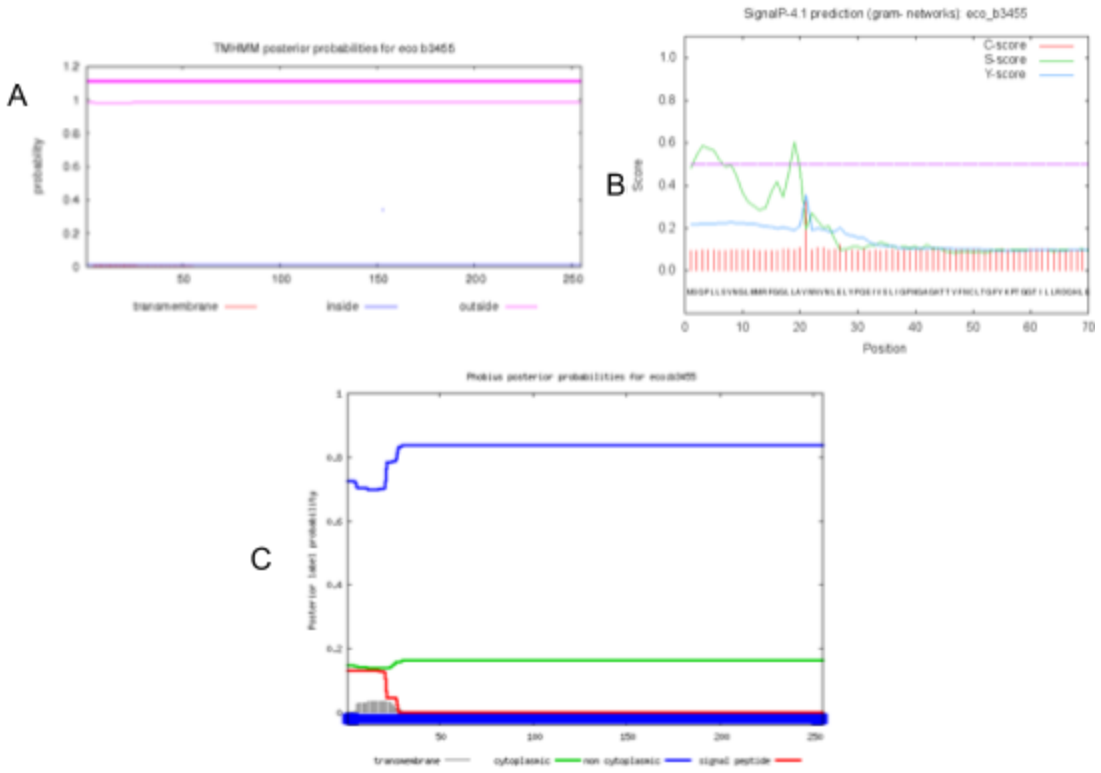


Figure 12. Cellular location determination of b3455. Panel A: TMHMM (Krogh et al., 2001; Krogh et al., 2016; Sonnhammer et al., 1998) shows the lack of transmembrane helices because there are not any red peaks present; Panel B: SignalP shows the lack of a signal peptide because the there are no peaks past the central cut-off line (Petersen et al., 2011); Panel C: Phobius shows transmembrane helices in gray peaks, where the gene is cytoplasmic in green, and where the gene is non-cytoplasmic in blue (Kall et al., 2004; Kall et al., 2007). The absence of grey peaks means there are not any transmembrane helices. The bioinformatics tools used are described in Methods.

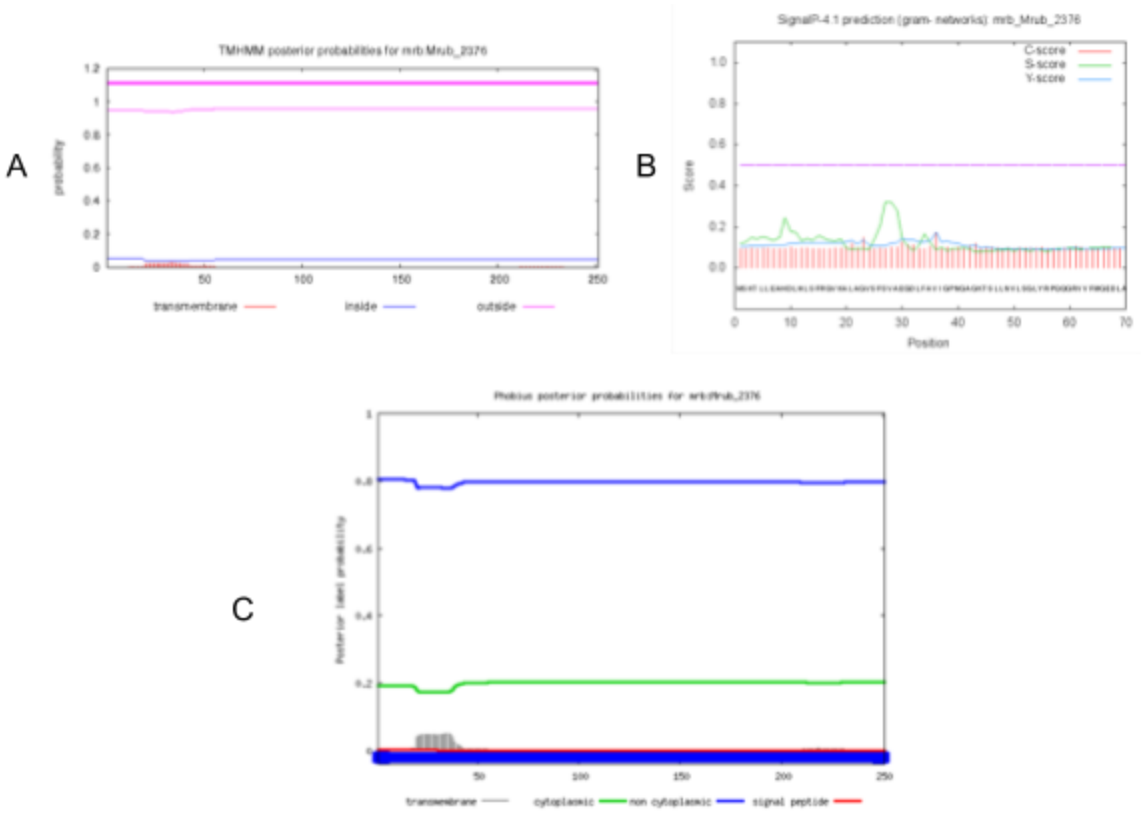


Figure 13. Cellular location determination of Mrub_2376. Panel A: TMHMM shows the lack of transmembrane helices because there are not any red peaks present (Krogh et al., 2001; Krogh et al., 2016; Sonnhammer et al., 1998); Panel B: SignalP shows the lack of a signal peptide because the there are no peaks past the central cut-off line (Petersen et al., 2011); Panel C: Phobius shows transmembrane helices in gray peaks, where the gene is cytoplasmic in green, and where the gene is non-cytoplasmic in blue (Kall et al., 2004; Kall et al., 2007). The absence of grey peaks means there are not any transmembrane helices. The bioinformatics tools used are described in Methods.

The Pfam test showed both b3455 and Mrub_2367 belong to the PF00005 (ABC Transporter) group (Finn et al., 2014; Finn et al., 2016). Alignments were also obtained from the test and are shown in figure 14. Unlike the BLAST alignment, the Pfam alignment is a pairwise alignment that compares the sequences of both b3455 and Mrub_2376 to a consensus sequence obtained from hundreds of other proteins. Both b3455 and Mrub_2376 were matched to the same consensus sequence, but the consensus sequence for the b3455 gene began 1 position after the consensus sequence for Mrub_2376. Both b3455 and Mrub_2376 have several glycine residues conserved with the consensus sequence along with a lysine residue. The commonality of shared amino acid residues with the consensus sequence between b3455 and Mrub_2376 is another piece of evidence supporting the hypothesis that the two gene are orthologous to one another.

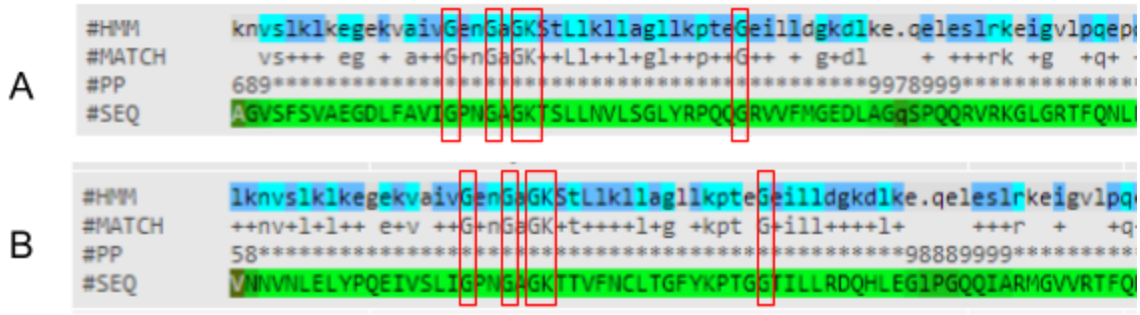


Figure 14. Panel A shows the alignment between b3455 and the consensus sequence (Finn et al., 2014; Finn et al., 2016). Panel B shows the alignment between Mrub_2376 and the consensus sequence. Conserved amino acids are written in capital letters in the "#MATCH" line. Both b3455 and Mrub_2376 have multiple glycine residues conserved with the consensus sequence as well as a lysine residue, as indicated by the red outlines on each alignment. The #HMM line is the consensus sequence and the #SEQ line is the gene being analyzed (either b3455 or Mrub_2376). The pairwise alignment was produced by the Pfam website http://pfam.sanger.ac.uk/search.

In addition to the consistency between cellular localization data, support of the orthologous relationship between b3455 and Mrub_2376 can be observed through their familial similarity exhibited by the names and E-values presented in Table 2. Further support is available by confirming that both genes are parts of operons and are involved in the same molecular pathway. Both b3455 and Mrub_2376 belong to the branched-chain amino acid category and are each part of a 5-gene operon. The IMG/M Color by Kegg feature presented clear images of each gene within its operon and in relation to the flanking regions upstream and downstream and is shown in figure 15 (Markowitz *et al.*, 2012). The presence of an operon is indicated by the same color identifier and direction of transcription. The fact that b3455 and Mrub_2376 are part of operons and within the same biochemical pathway are strong indications that the genes are orthologous

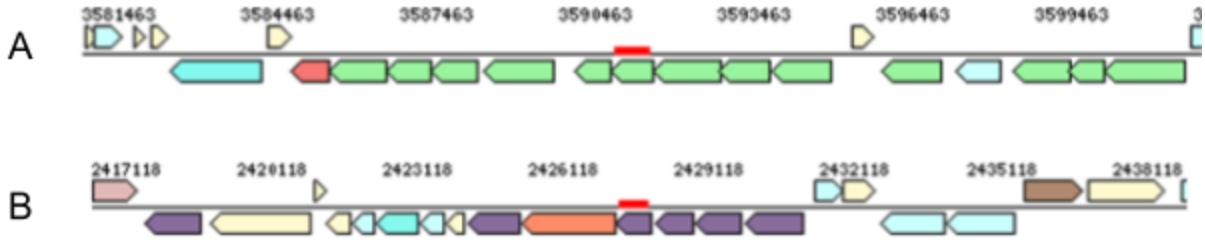


Figure 15. Both b3455 and Mrub_2376 exist as units of distinct operons. Panel A: The Color by Kegg Chromosome Map viewer (Markowitz *et al.*, 2012) of the area surrounding b3455 with the GOI indicated by the red dash; Panel B: The output of the same program for the area surrounding Mrub_2376.

E. coli* gene *b3456* and *M. ruber* gene *Mrub_2374

Figure 16 is the one of the results of the initial BLASTp search of the *E. coli* gene b3456 against the *M. ruber* genome; Mrub_2374, although not the top hit, was still identified as a possible ortholog of b3456 (Altschul *et al.*, 1990; Madden, 2002). The alignment yielded an E-value of 2e-26 and a 47% positive alignment (134/281 amino acids). This data is support that Mrub_2374 is an ortholog of b3456.

Score	Expect	Method	Identities	Positives	Gaps
106 bits(264)	2e-26	Compositional matrix adjust.	90/281(32%)	134/281(47%)	42/281(14%)

```
Query  140  LGYGGFYAIGAYTFALLNHYYGLGFWTCLPI--AGLMAAAAGFLLGFPVLRLRGDYLAIV  197
            LG   F  IGAY  + L+     G    L I   GL+AA  G +LG P LR++G YLAI
Sbjct  87   LGQAAFVGIGAYVSSHLS-----GDLAPLGILAGGLVAALIGIVLGIPSLRIKGAYLAIA  141

Query  198  TLGFGEIVRILLLNNTEITGGPNGISQIPKPTLFGLEFSRTAREGGWDTFSNFFGLKYDP  257
            TL F  +   +    T  TGG  G S   + T  GL  +
Sbjct  142  TLAFQFLADYVFKRWTAFTGGVAGRSLPAEGTFLGLPLA----------------------  180

Query  258  SDRVIFLYLVALLLVVLSLFVINRLLRMPLGRAWEALREDEIACRSLGLSPRRIKLTAFT  317
             DRV+  Y + L+  +   F   RLL    GRAW A+R+++++ +  G+   R KLTAF
Sbjct  181  DDRVV--YYLGLVFAIPMFFYAKRLLSTRAGRAWFAVRDNDLSAQLSGVDLVRSKLTAFA  238

Query  318  ISAAFAGFAGTLFAARQGFVSPESFTFAESAFVLAIVVLGGMGSQFAVILAAILLVVSRE  377
            +SA + G AG +     G V+PE+F  A S   LAIV++GG G+    +L A+ +V+  E
Sbjct  239  LSAFYGGIAGGILMHLIGAVTPENFVLAFSIQYLAIVIVGGAGTVLGAVLGALFIVLIPE  298

Query  378  LMR------------DFNEYSMLMLGGLMVLMMIWRPQGLL  406
            ++                + +  +  G L++L +I  P+GL+
Sbjct  299  ILSVIAGAFGPQYVAQLSAWRSVAFGVLILLFLILEPRGLV  339
```

Figure 16. b3456 and Mrub_2374 are possible orthologs based on the similar protein sequence. In the alignment, "query" represents the b3456 sequence and "sbjct" is the subject sequence of Mrub_2374. Analysis was performed using NCBI BLASTp program at http://www.ncbi.nlm.nih.gov (Altschul *et al.*, 1990; Madden, 2002).

Since Mrub_2374 was identified as a possible ortholog of b3456, it was important to confirm that the start codon of the *M. ruber* gene was called correctly. The determination of the correctly called start codon is to confirm that the differences observed in the BLAST alignment is due to differences in the gene sequences and not from an outside factor. The following order of bioinformatics tools was used to identify the correct start codon of Mrub_2374: IMG/M alternate ORF program (Markowitz *et al.*, 2012) to see additional start codon possibilities, BLASTp (Altschul *et al.*, 1990; Madden, 2002) to obtain amino acid sequences similar to the GOI for a multiple sequence alignment (MSA), TCoffee to create a MSA (Notredame *et al.*, 2000), and Weblogo to visually represent the MSA results (Crooks et al., 2004). The IMG/M viewer showed evidence of a correctly called start codon approximately 10 amino acids downstream from a possible Shine-Dalgarno sequence because of the existing methionine residue in the first reading frame (Markowitz *et al.*, 2012). The MSA created also supports that the start codon is accurate because all but one of the selected species (the least related) exhibited the same starting methionine residue (Notredame *et al.*, 2000); this resulted in a highly conserved M in the Weblogo (Crooks et al., 2004). The three programs indicate the correct start codon for Mrub_2374.

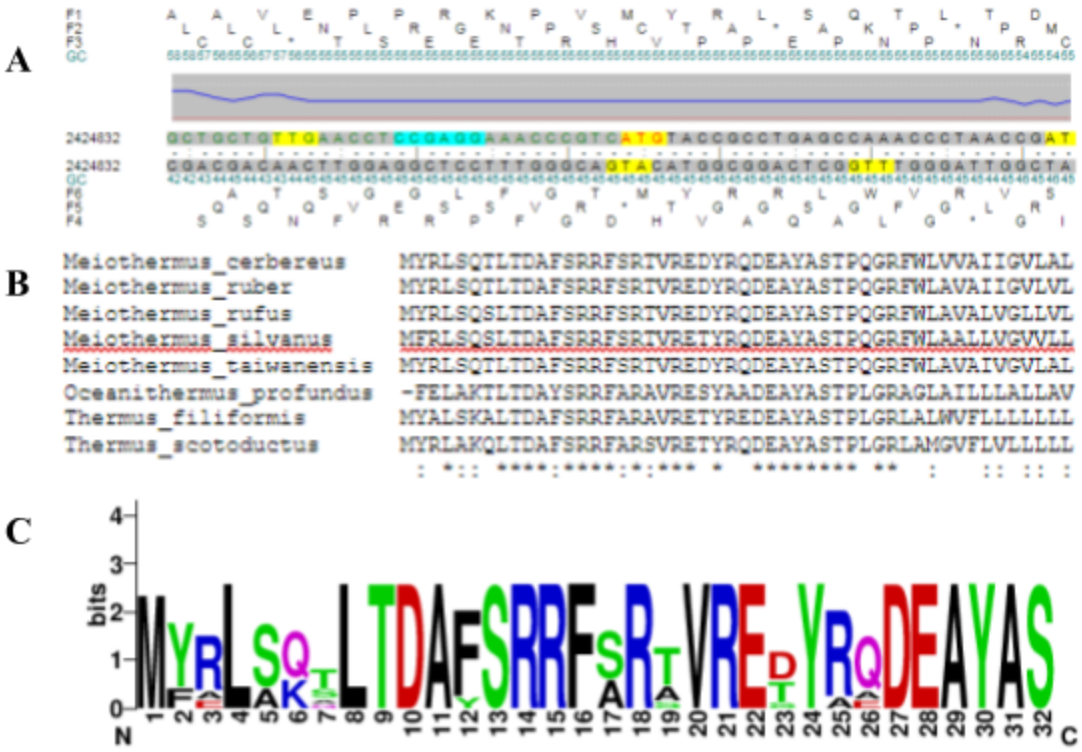


Figure 17. The start codon of Mrub_2374 appears to be correctly called. Panel A: IMG/M alternate ORF viewer with the suggested start codon highlighted and in red and the possible Shine-Dalgarno sequence highlighted in blue (Markowitz *et al.*, 2012); Panel B: first line of TCoffee MSA (Notredame *et al.*, 2000); Panel C: first line of Weblogo created from MSA (Crooks et al., 2004) . The tools used are described in the Methods section.

For the b3456 and Mrub_2374, the bioinformatics tools used provided a solid base for the analysis of the structure and function of each. The table below summarizes the results from the programs used, any pertinent family names and numbers, as well as E-values. It is important to analyze the E-values of each output to confirm that the gene are related and that the alignment is not due to random chance.

Bioinformatics Tool Used	***E. coli* b3456**	***M. ruber* Mrub_2374**
BLAST *E. coli* against *M. ruber*	Score: 106 (264) E-value: 2e-26	
CDD Data (COG category)	COG Number: COG4177 COG name: LivM	
	E-value:6.05e-78	E-value: 2.89e-52
Cellular Localization	Cytoplasmic Membrane	

TIGRfam (protein family)	TIGR03727 (Urea t UrtC arc) TIGR03408 (Urea trans UrtC)	
	E-value: 3.4e-11 E-value: 1.4e-10	E-value: 7.5e-06 E-value: 2.1e-12
Pfam (protein family)	PF02653 (BPD_transp_2) Branched-chain amino acid transport system / permease component	
	E-value:9e-60	E-value: 6.7e-37
Protein Database	Name: Structure of Oligopeptidase B from *Leishmania major*	
	E-value: N/A	E-value: 1.22862
KEGG Pathway Map	Prokaryotic-type ABC Transporters (02010)	

Table 3 is a summary of the outputs of several bioinformatics tools used to analyze the similarity between the function, location, and family of the *E. coli* gene b3457 and that of the *M. ruber* gene Mrub_2378. A BLAST analysis of the b3457 sequence against the *M. ruber* genome was performed first and the results are found in the first row of data (Altschul *et al.*, 1990; Madden, 2002). The E-value obtained is relatively low which implies that the alignment between the two sequences is not due to random chance but rather that there is an orthologous connection. The CDD from the BLAST search showed the same COG number (COG4177) and the name "LivM" for both sequence searches with very low E-values (Marchler-Bauer *et al.*, 2015). The matching COG hits indicate that the *M. ruber* and *E. coli* genes share the same function in a prokaryotic cell. Each of the programs used for analysis of the gene location in a cell exhibited that both b3457 and Mrub_2378 are found bound within the cytoplasmic membrane of the cell; data indicated that transmembrane helices are present, but no signal peptides or cleavage sites (TMH (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998), SignalP (Petersen *et al.*, 2011), LipoP (Juncker *et al.*, 2003), Phobius (Kall *et al.*, 2004; Kall *et al.*, 2007) and PSORT-B (Yu *et al.*, 2010)). A matching cellular localization also confirms the orthologous relationship between the gene pair that is present. The programs TIGRfam (Haft *et al.*, 2001) and Pfam (Finn *et al.*, 2014; Finn *et al.*, 2016) also strongly support that b3457 is an ortholog to Mrub_2378 because they yielded the same or similar top hits. The top two hits from TIGRfam detailed that the genes belong to either the urea t UrtC_arc family or the urea_trans UrtC family (Haft *et al.*, 2001). The E-values detailed in the table indicate that the first two hits for *E. coli* and *M. ruber* are the inverse of each other. The fact that the top hits are similar between species is still confirmation of orthologous relationship. One hit was obtained from Pfam for each species which is BPD_transp_2 (PF02653) for both (Finn *et al.*, 2014; Finn *et al.*, 2016). Protein Database results were unavailable for *E. coli* and one hit was available for *M. ruber* (Berman *et al.*, 2000; Berman *et al.*, 2000). The hit for Mrub_2374 is for the Structure of Oligopeptidase B from *Leishmania major*, however it has an E-value of 1.22862, which is greater than the typical cut-off value of 0.0001. This species is not related close enough, explaining the high E-value and the lack of a hit for *E. coli.* The Kegg

database confirms that b3456 and Mrub_2374 are found in the prokaryotic-type ABC transporters pathway (Kanehisa *et al.*, 2016).

The confirmation of the start codon of Mrub_2374 allows the analysis of the orthologous relationship to continue. The next piece of supporting evidence is the cellular localization of each of the genes; the location of the genes was analyzed using a series of bioinformatics programs which resulted in comparable data between Mrub_2374 and b3456. Figure 18 shows the results of each of the programs for the *E. coli* sequence b3456 and Figure 19 exhibits the results of the same programs used for the *M. ruber* sequence Mrub_2374. The TMHMM programs detailed that b3456 contains 10 transmembrane helices and Mrub_2374 contains 8 transmembrane helices (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998). PSORT-B places both genes in the cytoplasmic membrane with a maximum score of 10.00 (Yu *et al.*, 2010). For the *E. coli* GOI, SignalP (Petersen *et al.*, 2011) yielded no signal peptides present, however LipoP (Juncker *et al.*, 2003) stated that there was a signal peptide present; Phobius (Kall *et al.*, 2004; Kall *et al.*, 2007) also showed a signal peptide cleaved in the beginning of the sequence. For *M. ruber*, the remaining programs all agreed that no signal peptides were present. Although there is a slight discrepancy about the presence/lack of a signal peptide, the cellular localization data is comparable between the two species and is confirmation of an orthologous relationship.

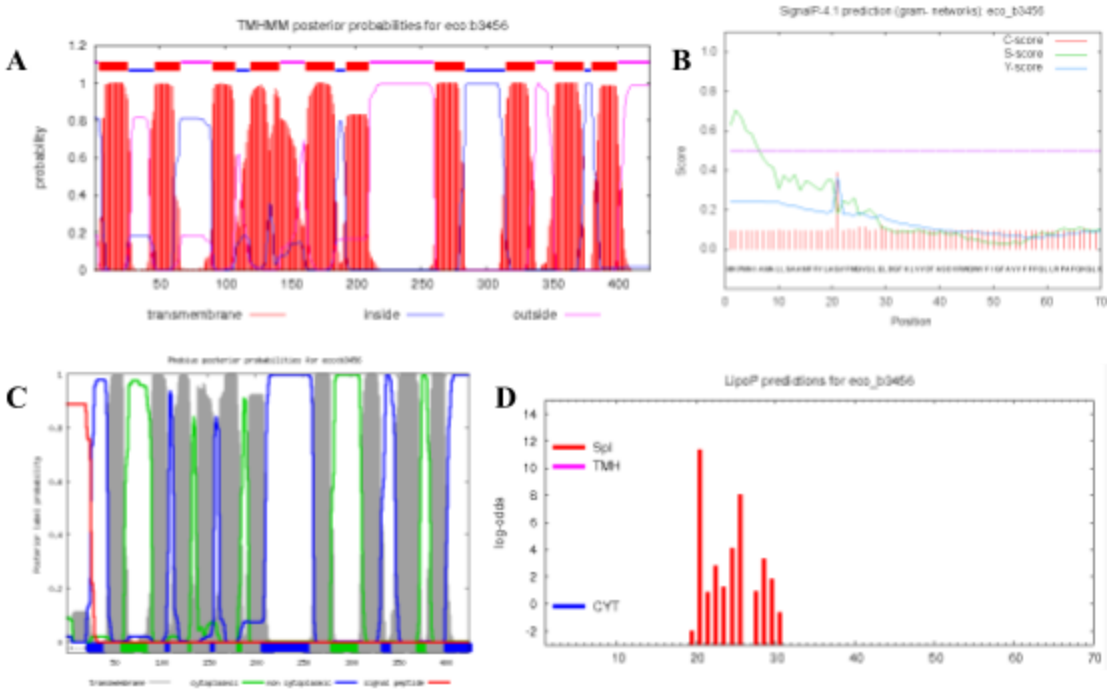


Figure 18. Cellular location determination of b3456. Panel A: TMHMM showing red peaks that represent each transmembrane helix (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); Panel B: SignalP shows the lack of a signal peptide because the there are no peaks past the central cut-off line (Petersen *et al.*, 2011); Panel C: Phobius shows transmembrane helices in gray peaks and the signal peptide in red (Kall *et al.*, 2004; Kall *et al.*, 2007); Panel D: LipoP output with the location of the signal peptide in red (Juncker *et al.*, 2003). The bioinformatics tools used are described in Methods.

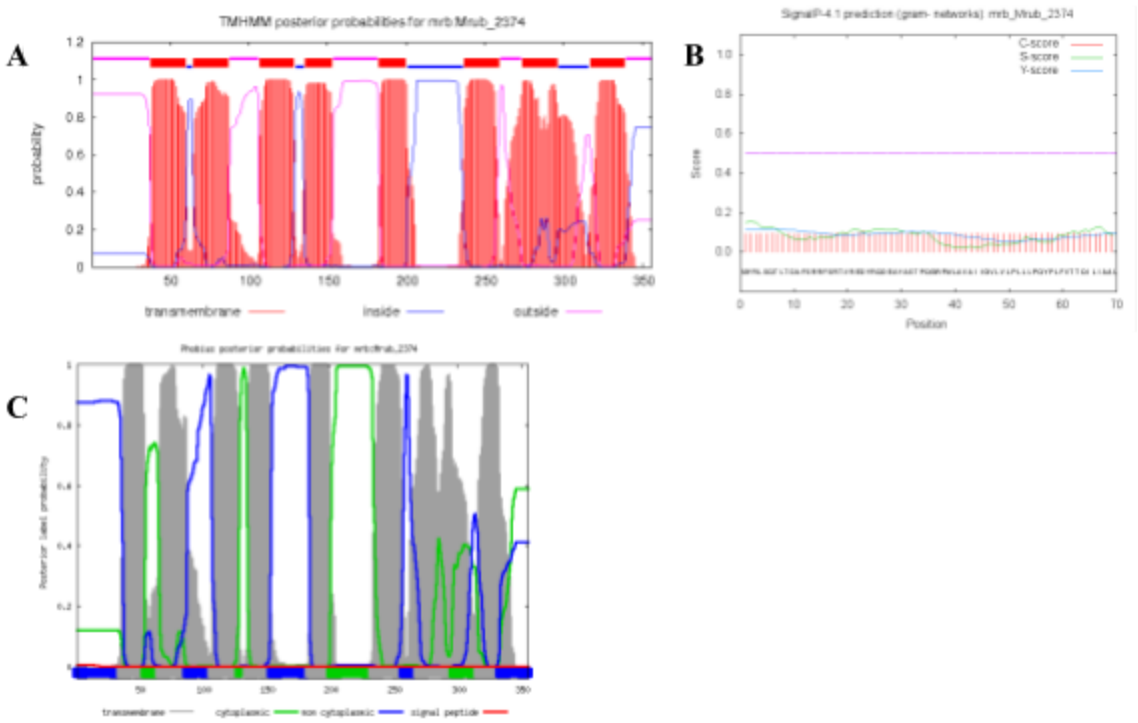


Figure 19. Cellular location determination of Mrub_2374. Panel A: TMHMM shows red peaks that represent each time the gene passes through the cytoplasmic membrane (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); Panel B: SignalP shows the lack of a signal peptide because the there are no peaks past the central cut-off line (Petersen *et al.*, 2011); Panel C: Phobius shows transmembrane helices in gray peaks, where the gene is cytoplasmic in green, and where the gene is non-cytoplasmic in blue (Kall *et al.*, 2004; Kall *et al.*, 2007). The bioinformatics tools used are described in Methods.

The Pfam program output yielded the name BPD_transp_2 (PF02653) as the name of the family that b3456 and Mrub_2374 belong to and also shows a pairwise alignment between the gene sequence and a consensus sequence which can be seen in figure 20 (Finn *et al.*, 2014; Finn *et al.*, 2016). The same consensus sequence was used for both b3456 and Mrub_2374 which indicates that if both genes align well with the consensus sequence (E-values of 9e-60 and 6.7e-37, respectively) there is further evidence of the two genes being orthologous; this results is observed so there the orthologous relationship is supported. Additionally, the number of highly conserved amino acids provides evidence for the relationship.

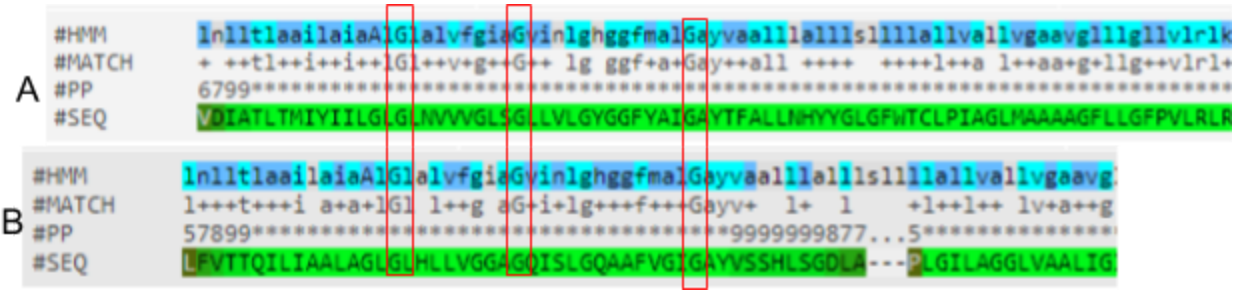


Figure 20. Consensus alignment against b3456 and Mrub_2374 (Finn *et al.*, 2014; Finn *et al.*, 2016). Panel A: The #HMM line represents the consensus sequence, #SEQ represents the b3456 sequence, #MATCH shows highly conserved amino acids; Panel B: The #HMM line represents the consensus sequence, #SEQ represents the Mrub_2374 sequence, #MATCH shows highly conserved amino acids. The red boxes surround the amino acids that are conserved in both sequences.

The shared families and domains identified (Table 1), the BLASTp query search (Figure 16), and similar cellular localization interpretations (Figures 18 and 19) are substantial evidence that b3456 and Mrub_2374 are orthologs to one another. The respective operons that b3456 and Mrub_2374 are each a part of are additional pieces of evidence that the two genes of interest are an orthologous pair. b3456 and Mrub_2374 and their flanking upstream and downstream regions can be observed using the IMG/M Color by Kegg feature to identify the operons (Markowitz *et al.*, 2012). For Mrub_2374, the gene is in an operon with the 4 purple-colored genes downstream from it; the salmon-colored gene interrupting the operon was found to be of functional importance and will be discussed further in the conclusion. For b3456, the gene of interest is found in the middle of the operon. The location of the gene within the operon is a slight discrepancy between the two genes, but the common operon distinction is still evidence that Mrub_2374 is an ortholog to b3456.

Figure 21. Both b3456 and Mrub_2374 exist as units of distinct operons. Panel A: The Color by Kegg Chromosome Map viewer (Markowitz *et al.*, 2012) of the area surrounding b3456 with the GOI indicated by the red dash; Panel B: The output of the same program for the area surrounding Mrub_2374.

E. coli* gene *b3457* and *M. ruber* gene *Mrub_2378

Figure 22 shows the output of the initial BLASTp search of b3457 against the *M. ruber* genome, selecting *Mrub_2378* as the one of the top results (Altschul *et al.*, 1990; Madden, 2002). This search was performed before the rest of the bioinformatics programs were used to confirm that b3457 and Mrub_2378 are orthologs. The low E-value of 6e-17 and the 154 matching amino acids between the two sequences are support of the relationship. The BLAST results represent the first piece of support that there are structural and functional similarities between the gene found in *E. coli* and *M. ruber*.

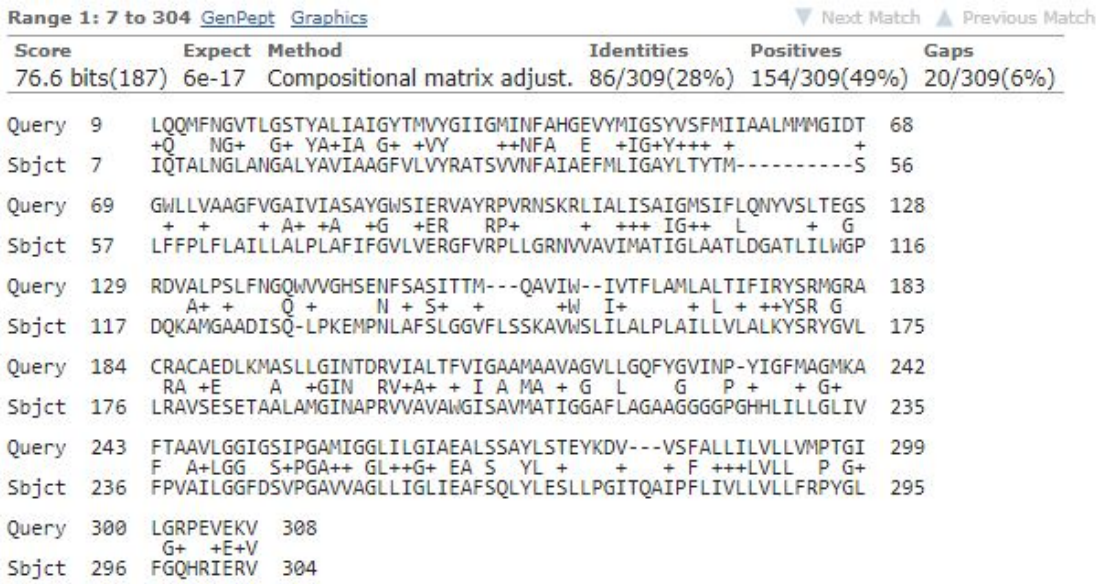


Figure 22. Mrub_2378 and b3457 have a similar protein sequence. In the sequence alignment, "query" represents is b3457 and "sbjct" represents the subject sequence of Mrub_2378. Analysis was performed using NCBI BLASTp program at http://www.ncbi.nlm.nih.gov (Altschul *et al.*, 1990; Madden, 2002).

Before continuing with additional bioinformatics tools, it was necessary to confirm that the start codon of Mrub_2378 was correctly called. This is necessary because an incorrectly called start codon may lead to an inaccurate alignment between the *E. coli* and *M. ruber* sequences. Figure 23 shows the various programs used for to obtain this result. The IMG/M alternate ORF program did not suggest any viable options for an alternate start codon (Markowitz *et al.*, 2012). Using similar sequences obtained from a BLAST search, the multiple alignment sequence (Notredame *et al.*, 2000) showed that the start codon was highly conserved among the *Meiothermus* genus but not among the other related organisms selected. The Weblogo output showed that the start codon was only moderately conserved because of the multiple sequence alignment entered (Crooks et al., 2004). The fact that the start codon was not conserved among all of the selected sequences is not enough support to claim that the start codon was incorrectly called; the methionine was conserved among *Meiothermus* species, which includes our gene of interest, and there were no other suggested start codons in the alternate ORF viewer.

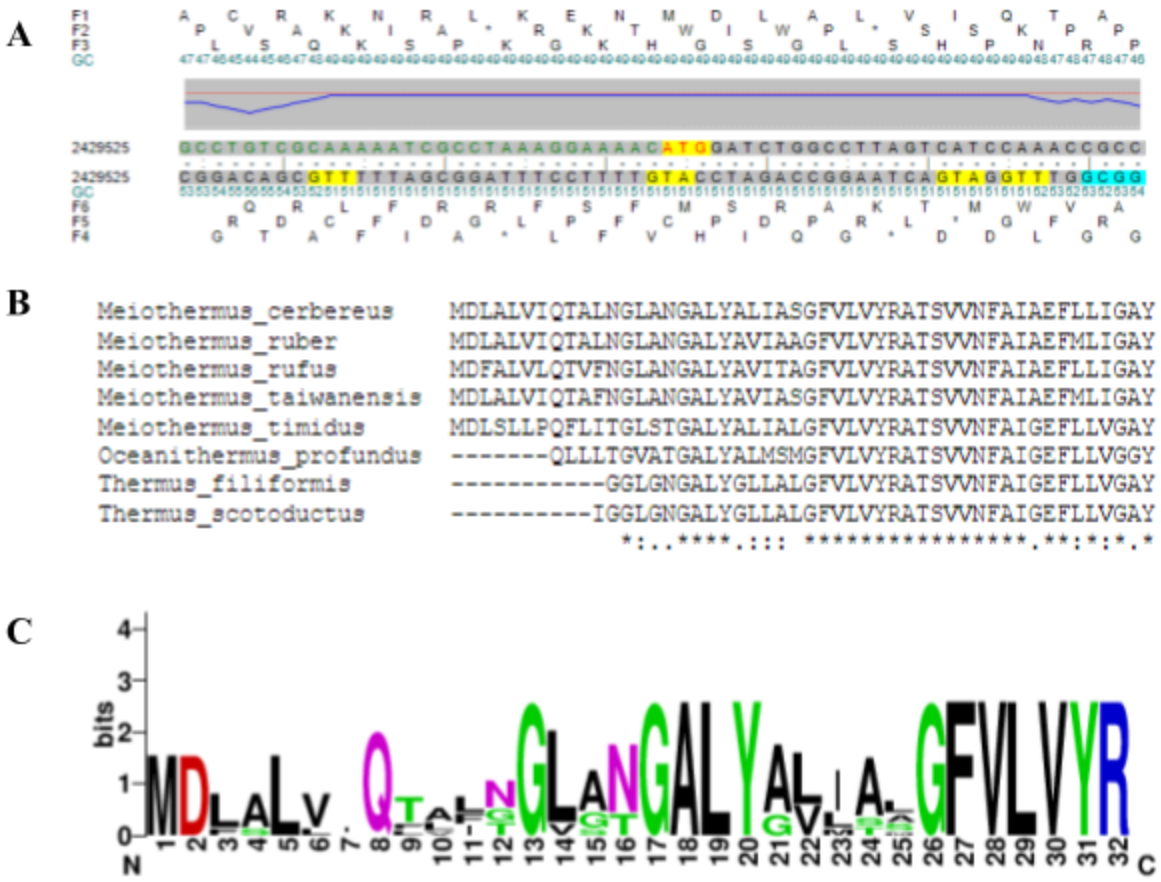


Figure 23. The start codon of Mrub_2378 is correctly called. Panel A: IMG/M alternate ORF viewer with the suggested start codon highlighted in red (Markowitz *et al.*, 2012); Panel B: first line of TCoffee MSA with dashes representing unmatched amino acids (Notredame *et al.*, 2000); Panel C: first line of Weblogo created from MSA (Crooks et al., 2004). The programs used are described in the Methods section.

To analyze the relationship between b3457 and Mrub_2374, various bioinformatics tools were used to provide information for the analysis of the structure and function of each. The table below summarizes the results from the programs used, including any pertinent family names and numbers and E-values. It is important to analyze the E-values of each output to confirm that the gene are related and that the alignment is not due to random chance and that there is a relationship present.

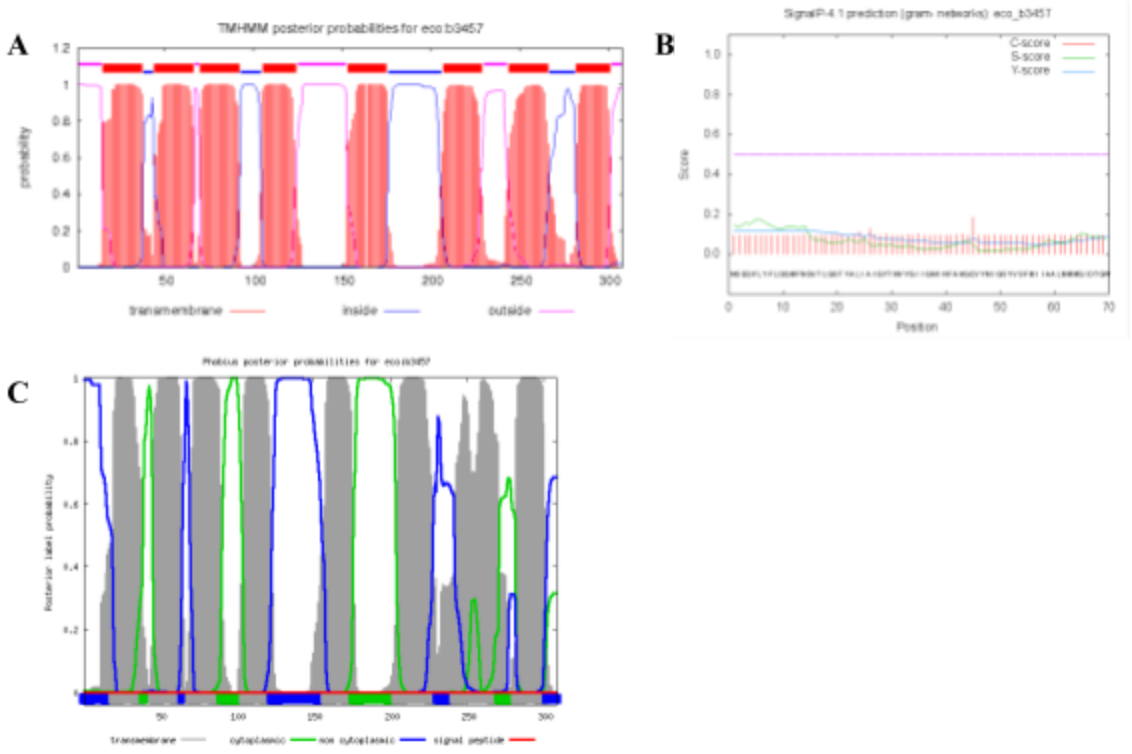
Bioinformatics Tool Used	***E. coli* b3457**	***M. ruber* Mrub_2378**
BLAST *E. coli* against *M. ruber*	Score: 76.6 bits(187) E-value: 6e-17	
CDD Data (COG category)	COG Number: COG0559	

	COG name: LivH	
	E-value: 4.54e-82	E-value: 3.24e-38
Cellular Localization	Cytoplasmic Membrane	
TIGRfam (protein family)	TIGR03409 (urea trans UrtB) TIGR03622 (urea t UrtB)	
	E-value: 9.23-15 E-value: 1.2e-07	E-value: 2.3e-10 E-value: 3.5e-07
Pfam (protein family)	PF02653 (Branched-chain amino acid transport system / permease component) Branched-chain amino acid transport system / permease component	
	E-value:6.7e-71	E-value: 9e-36
Protein Database	Name: N/A	
	E-value: N/A	E-value: N/A
KEGG Pathway Map	Prokaryotic-type ABC Transporters (02010)	

Table 4 presents the summarized results obtained from various bioinformatics programs that were utilized in the comparison of the function, location, and family of the *E. coli* gene b3457 and the *M. ruber* gene Mrub_2378. The first BLAST analysis performed of the b3457 sequence against the *M. ruber* genome is shown in the first row of data (Markowitz *et al.*, 2012). The E-value obtained is relatively low, indicating that the alignment between the two sequences is not due to chance alone and that there is an orthologous connection present. The conserved domain data pulled from the BLAST search resulted in the same COG number and name (COG0559, LivH) with separately low E-values for both species (Marchler-Bauer *et al.*, 2015). The matching COG hits indicate that the *M. ruber* and *E. coli* genes share the same function in a prokaryotic cell. The programs used for cellular location determination agreed that both b3457 and Mrub_2378 are identified winding through the cytoplasmic membrane of the cell; transmembrane helices are found and further data indicated that no signal peptides or cleavage sites are present (TMH (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998), SignalP (Petersen *et al.*, 2011), LipoP (Juncker *et al.*, 2003), Phobius (Kall *et al.*, 2004; Kall *et al.*, 2007) and PSORT-B(Yu *et al.*, 2010)). The matching cellular localization analysis further confirms the orthologous relationship between the gene pair. The protein families identified from TIGRfam (Haft *et al.*, 2001) and Pfam (Finn *et al.*, 2014; Finn *et al.*, 2016) also strongly support that b3457 is an ortholog to Mrub_2378. The top hit from TIGRfam (Haft *et al.*, 2001) placed both genes in the urea trans UrtB family (TIGR03409) and the top hit from Pfam (Finn *et al.*, 2014; Finn *et al.*, 2016) assigned both genes to the Branched-chain amino acid transport system / permease component family (PF02653). The E-values obtained for both the *E. coli* and *M. ruber* genes

from both programs are close to zero. Information from PDB was unavailable for both genes in the pair and therefore could not add further confirmation to the relationship between the two (Berman *et al.*, 2000; Berman *et al.*, 2000). Using Kegg, it is known that b3457 and Mrub_2378 are found in the prokaryotic-type ABC transporters pathway (Kanehisa *et al.*, 2016).

With the confirmation of the correctly called start codon of Mrub_2378, analysis of the similarity between b3457 and Mrub_2378, continued with cellular localization programs. Figure 24 shows the results of each of the programs for the *E. coli* sequence b3457 and Figure 25 exhibits the results of the same programs used for the *M. ruber* sequence Mrub_2378. There is consistency between each panel of the two figures which supports that the two sequences of interest are orthologs. It can be observed that both genes of interest contain 8 transmembrane helices (TMHMM (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998) and Phobius (Kall *et al.*, 2004; Kall *et al.*, 2007)), contain no signal peptides (SignalP (Petersen *et al.*, 2011) and LipoP- no plot (Juncker *et al.*, 2003)), and are found within the cytoplasmic membrane (PSORT-B (Yu *et al.*, 2010)). In addition to the plots, PSORT-B assigned both b3457 and Mrub_2378 the maximum score of 10.00 in the cytoplasmic membrane location category.



\

Figure 24. Cellular location determination of b3457. Panel A: TMHMM showing red peaks that represent each transmembrane helix(Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); Panel B: SignalP shows the lack of a signal peptide because the there are no peaks past the central cut-off line (Petersen *et al.*, 2011); Panel C: Phobius shows transmembrane helices in gray peaks, where the gene is cytoplasmic in green, and where the gene is non-cytoplasmic in blue (Kall *et al.*, 2004; Kall *et al.*, 2007). The bioinformatics tools used are described in Methods.

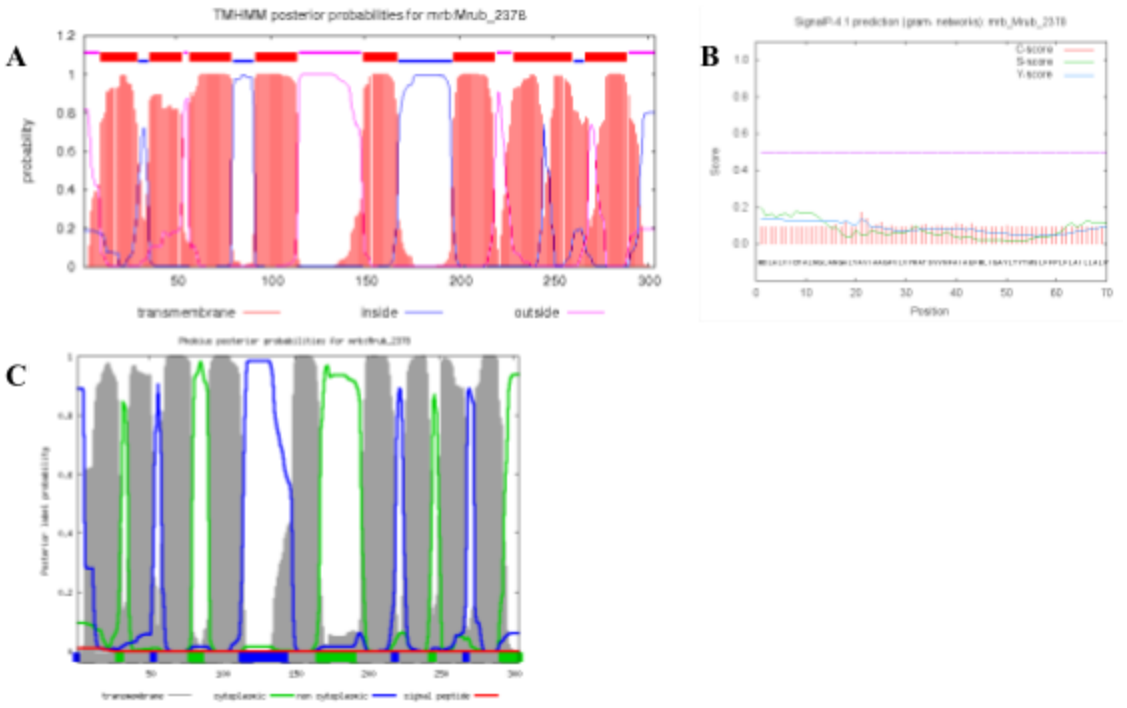


Figure 25. Cellular location determination of Mrub_2378. Panel A: TMHMM shows red peaks that represent each time the gene passes through the cytoplasmic membrane (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); Panel B: SignalP shows the lack of a signal peptide because the there are no peaks past the central cut-off line (Petersen *et al.*, 2011); Panel C: Phobius shows transmembrane helices in gray peaks, where the gene is cytoplasmic in green, and where the gene is non-cytoplasmic in blue (Kall *et al.*, 2004; Kall *et al.*, 2007). The bioinformatics tools used are described in Methods.

The output of the Pfam program provided the name of the family that b3457 and Mrub_2378 belong to Branched-chain amino acid transport system / permease component as well as a pairwise alignment between the gene sequence and a consensus sequence which can be seen in figure 26 (Finn *et al.*, 2014; Finn *et al.*, 2016). The same consensus sequence was used for both b3457 and Mrub_2379 so when both genes align well with the consensus sequence (E-values of 6.7e-71 and 9e-36, respectively) there is further evidence of the two genes being orthologous. The number of highly conserved amino acids is also evidence.

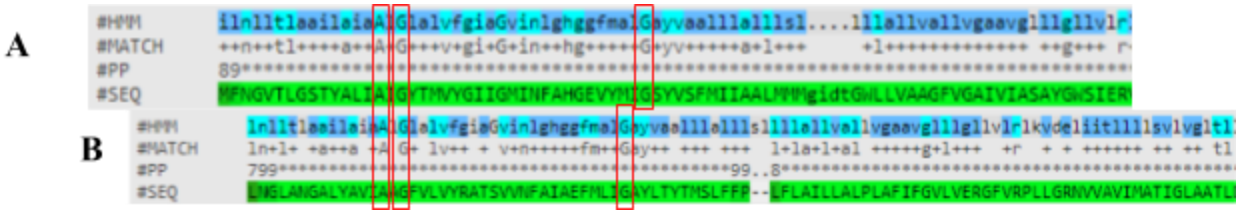


Figure 26. Consensus alignment against b3457 and Mrub_2378 (Finn *et al.*, 2014; Finn *et al.*, 2016). Panel A: The #HMM line represents the consensus sequence, #SEQ represents the b3457 sequence, #MATCH shows highly conserved amino acids; Panel B: The #HMM line represents the consensus sequence, #SEQ represents the Mrub_2378 sequence, #MATCH shows highly conserved amino acids.

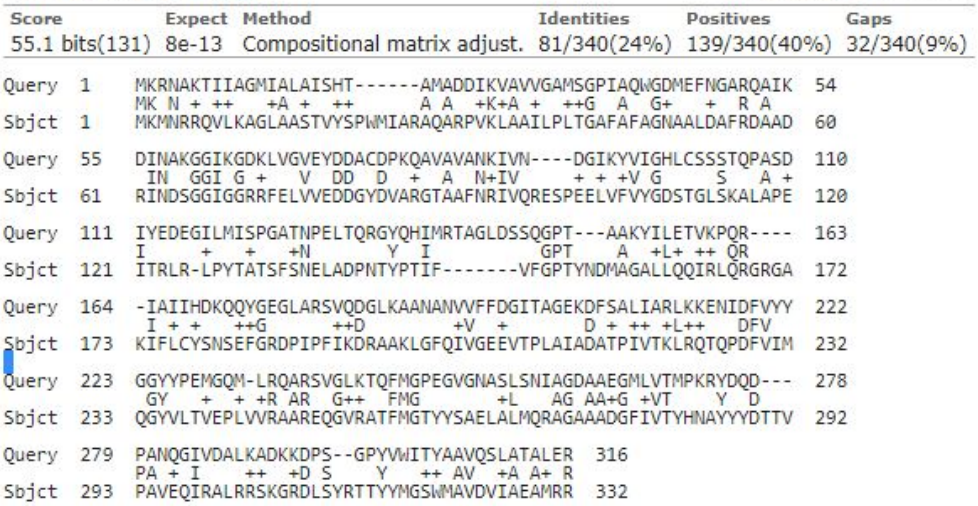
Highly conserved amino acids between both sequences are outlined in red; for both sequences, select glycine and alanine residues are highly conserved.

In addition to the consistency between cellular localization data, support of the orthologous relationship between b3457 and Mrub_2378 can be observed through their familial similarity exhibited by the names and E-values presented in Table 1. Further support is available by confirming that both genes are parts of operons and are involved in the same molecular pathway. Both b3457 and Mrub_2378 belong to the branched-chain amino acid category and are each part of a 5-gene operon. The IMG/M Color by Kegg feature presented clear images of each gene within its operon and in relation to the flanking regions upstream and downstream (Markowitz *et al.*, 2012). The presence of an operon is indicated by the same color identifier and direction of transcription. The fact that b3457 and Mrub_2378 are part of operons and within the same biochemical pathway are strong indications that the genes are orthologous.

Figure 27. Both b3457 and Mrub_2378 exist as units of distinct operons (Markowitz *et al.*, 2012). Panel A: The Color by Kegg Chromosome Map viewer of the area surrounding b3457 with the GOI indicated by the red dash; Panel B: The output of the same program for the area surrounding Mrub_2378.

E. coli* gene *b3458* and *M. ruber* gene *Mrub_2379

Figure 28 is the output yielded from the initial BLASTp search of the b3458 gene from *E. coli* against the *M. ruber* genome to search for the Mrub_2379 ortholog (Altschul *et al.*, 1990; Madden, 2002). The Mrub_2379 ortholog was found as one of the top hits on the BLASTp search. The low E-value of 8e-13 and the moderate positive alignment between the two sequences (139/340 amino acids aligned) are indications of the orthologous relationship present between the two genes of interest.



Score	Expect	Method	Identities	Positives	Gaps
55.1 bits(131)	8e-13	Compositional matrix adjust.	81/340(24%)	139/340(40%)	32/340(9%)

```
Query  1    MKRNAKTIIAGMIALAISHT------AMADDIKVAVVGAMSGPIAQWGDMEFNGARQAIK  54
            MK N + ++   +A +  ++      A A  +K+A +  ++G  A  G+   +  R A
Sbjct  1    MKMNRRQVLKAGLAASTVYSPWMIARAQARPVKLAAILPLTGAFAFAGNAALDAFRDAAD  60

Query  55   DINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVN----DGIKYVIGHLCSSSTQPASD  110
             IN  GGI G +   V  DD  D  +  A  N+IV     + + +V G     S   A +
Sbjct  61   RINDSGGIGGRRFELVVEDDGYDVARGTAAFNRIVQRESPEELVFVYGDSTGLSKALAPE  120

Query  111  IYEDEGILMISPGATNPELTQRGYQHIMRTAGLDSSQGPT---AAKYILETVKPQR----  163
            I     +    +N       Y  I          GPT    A  +L+ ++ QR
Sbjct  121  ITRLR-LPYTATSFSNELADPNTYPTIF-------VFGPTYNDMAGALLQQIRLQRGRGA  172

Query  164  -IAIIHDKQQYGEGLARSVQDGLKAANANVVFFDGITAGEKDFSALIARLKKENIDFVYY  222
             I + +   ++G      ++D        +V  +       D + ++ +L++   DFV
Sbjct  173  KIFLCYSNSEFGRDPIPFIKDRAAKLGFQIVGEEVTPLAIADATPIVTKLRQTQPDFVIM  232

Query  223  GGYYPEMGQM-LRQARSVGLKTQFMGPEGVGNASLSNIAGDAAEGMLVTMPKRYDQD---  278
             GY   +  + +R AR  G++  FMG        +L   AG AA+G +VT    Y  D
Sbjct  233  QGYVLTVEPLVVRAAREQGVRATFMGTYYSAELALMQRAGAAADGFIVTYHNAYYYDTTV  292

Query  279  PANQGIVDALKADKKDPS--GPYVWITYAAVQSLATALER  316
            PA + I    ++  +D S    Y   ++ AV  +A A+ R
Sbjct  293  PAVEQIRALRRSKGRDLSYRTTYYMGSWMAVDVIAEAMRR  332
```

Figure 28. b3458 and Mrub_2379 are possible orthologs based on the similar protein sequence. “Query” represents the b3458 sequence and “sbjct” is the subject sequence of *Mrub_2379*. Analysis was performed using NCBI BLASTp program at http://www.ncbi.nlm.nih.gov (Altschul *et al.*, 1990; Madden, 2002).

With an orthologous relationship suspected between b3458 and Mrub_2379 it is important to confirm that the *M. ruber* genome has the correctly called starting codon. The importance of knowing the correct starting codon is to properly align the sequences of the two genes of interest. Figure 29 shows the various programs used for to confirm the start codon result. The IMG/M alternate ORF program suggests several possible start codons near one another so further confirmation is required (Markowitz *et al.*, 2012). Using similar sequences obtained from a BLAST search, the multiple alignment sequence created from TCoffee showed that the original codon is conserved among each of the selected sequences (Notredame *et al.*, 2000). The MSA exhibited a highly conserved methionine residue in the Weblogo (Crooks et al., 2004). This is all support the the start codon of Mrub_2379 was called correctly.

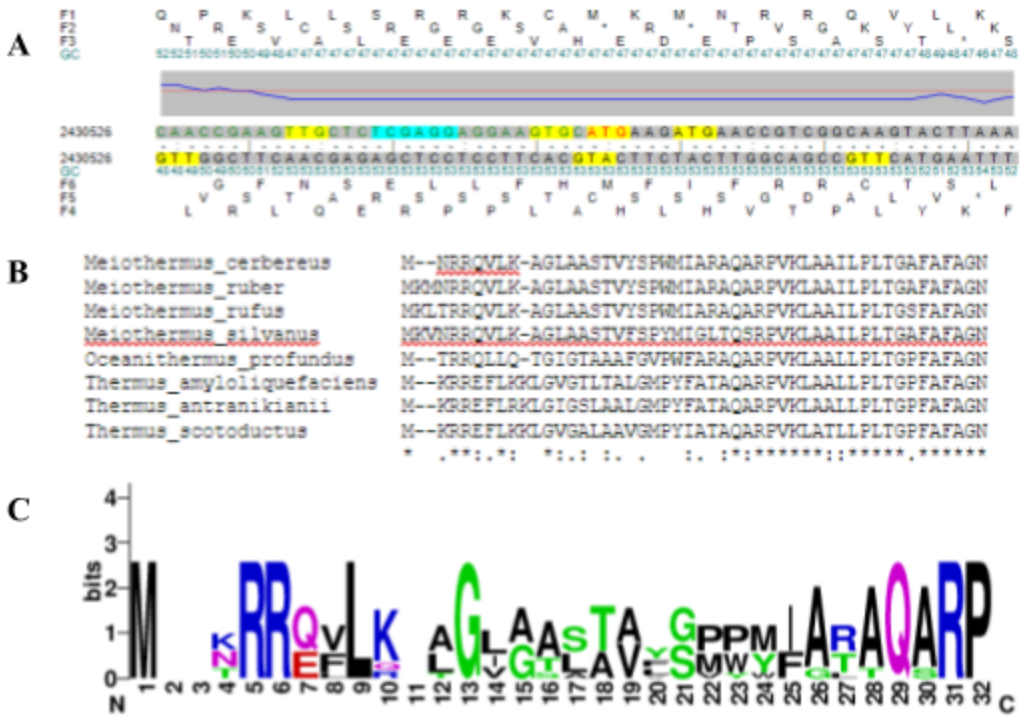


Figure 29. The start codon of Mrub_2379 appears to be correctly called. Panel A: IMG/M alternate ORF viewer with the original start codon highlighted and in red, suggested alternate start codons in plain highlight, and the possible Shine-Dalgarno sequence highlighted in blue (Markowitz *et al.*, 2012); Panel B: first line of TCoffee MSA (Notredame *et al.*, 2000); Panel C: first line of Weblogo created from MSA (Crooks et al., 2004). The tools used are described in the Methods section.

For the b3458 and Mrub_2379, the structure of function of each gene was analyzed by using different bioinformatics tools. The table below summarizes the outputs of the programs used, and provides the pertinent family names and numbers, as well as E-values. It is important to analyze the E-values of each output to confirm that the gene are related and that the alignment is not due to random chance.

Bioinformatics Tool Used	***E. coli* b3458**	***M. ruber* Mrub_2379**
BLAST *E. coli* against *M. ruber*	Score: 55.1 bits (131) E-value: 8e-13	
CDD Data (COG category)	COG Number: COG0683 COG name: Liv K	
	E-value: 5.31e-154	E-value: 9.88e-40
Cellular Localization	Periplasmic Space	

TIGRfam (protein family)	TIGR03407 (urea_ABC_UrtA) TIGR03669 (urea_ABC_arch)	
	E-value: 0.0014 E-value: N/A	E-value: 1.1e-05 E-value: 1.6e-05
Pfam (protein family)	PF13458 (Periplasmic Binding Protein)	
	E-value: 1.7e-61	E-value: 4e-70
Protein Database	Name 1USG, L-leucine-binding protein	
	E-value: 0.0	E-value: N/A
KEGG Pathway Map	Prokaryotic-type ABC Transporters (02010)	

Table 5 is a summary of the several bioinformatics tools used to analyze the similarity between the function, location, and family of the *E. coli* gene b3458 and that of the *M. ruber* gene Mrub_2379 and the outputs yielded. An initial BLAST analysis of the b3458 sequence against the *M. ruber* genome was performed and the results are found in the first row of data (Altschul *et al.*, 1990; Madden, 2002). The relatively low E-value obtained of 8e-13 implies that the alignment between the two sequences is not due to random chance but rather that the sequences align relatively well and an orthologous connection is present. The CDD from the BLAST search showed the same COG number (COG0683) and the name "LivK" for both sequence searches with very low E-values (Marchler-Bauer *et al.*, 2015). The matching COG hits indicate that the *M. ruber* and *E. coli* genes share the same function in a prokaryotic cell. Each of the programs used for analysis of the gene location in a cell exhibited that both b3458 and Mrub_2379 are found floating within the periplasmic space of the cell; data indicated that transmembrane helices are not present, but one signal peptide (and associated cleavage site) is found (TMH (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998), SignalP (Petersen *et al.*, 2011), LipoP (Juncker *et al.*, 2003), Phobius (Kall *et al.*, 2004; Kall *et al.*, 2007) and PSORT-B(Yu *et al.*, 2010)). A matching cellular localization also confirms the orthologous relationship between the gene pair that is present. The programs TIGRfam (Haft *et al.*, 2001) and Pfam (Finn *et al.*, 2014; Finn *et al.*, 2016) also strongly support that b3458 is an ortholog to Mrub_2379 because they yielded the same or similar top hits. Two top hits from TIGRfam (Haft *et al.*, 2001) were provided for Mrub_2379 which detailed that the genes belongs to either the urea ABC UrtA family or the urea ABC arch family. The *E. coli* ortholog only provided one hit and it matched the top hit from Mrub_2379. The E-values detailed in the table are somewhat low and indicate that the first two hits for *E. coli* and *M. ruber* are the same. One hit was obtained from Pfam for each species which is Periplasmic Binding Protein (PF13458) for both, confirming that the gene is found in the periplasmic space as the substrate-binding domain of the ABC transporter (Finn *et al.*, 2014; Finn *et al.*, 2016). PDB showed one hit for b3458 and results were unavailable for the *M. ruber* gene (Berman *et al.*, 2000; Berman *et al.*, 2000). The protein that matched b3458 is an L-leucine binding protein which matches the predicted function of the gene. The Kegg

database confirms that b3458 and Mrub_2379 are found in the prokaryotic-type ABC transporters pathway (Kanehisa *et al.*, 2016).

Further analysis of the possible orthologous relationship between b3458 and Mrub_2379 was performed by utilizing several cellular localization bioinformatics tools. The tools were used for each gene, which can be seen in Figure 30 (b3458) and Figure 31 (Mrub_2379). Neither gene contains transmembrane helices (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); this is expected for the substrate-binding protein of the transporter complex. SignalP showed that a signal peptide is present in b3458 with the cleavage site after position 23 and no signal peptide in Mrub_2379 (Petersen *et al.*, 2011). LipoP stated the same information as SignalP for the *E. coli* gene and that there is also a signal peptide present present in the *M. ruber* gene with the cleavage site after position 27 (Juncker *et al.*, 2003). PSORT-B identified both genes of interest in the periplasmic membrane with the scores being 10.00 for *E. coli* and 9.76 for *M. ruber* (Yu *et al.*, 2010). Finally, Phobius confirmed the signal peptide seen in each gene near the beginning of the amino acid sequence (Kall *et al.*, 2004; Kall *et al.*, 2007). The near-identical results between the two genes is strong confirmation that Mrub_2379 is an ortholog of b3458. The fact that SignalP did not identify a signal peptide for the *M. ruber* gene is not sufficient evidence to discredit the relationship between the two genes because of the consistency observed among the remainder of the programs.

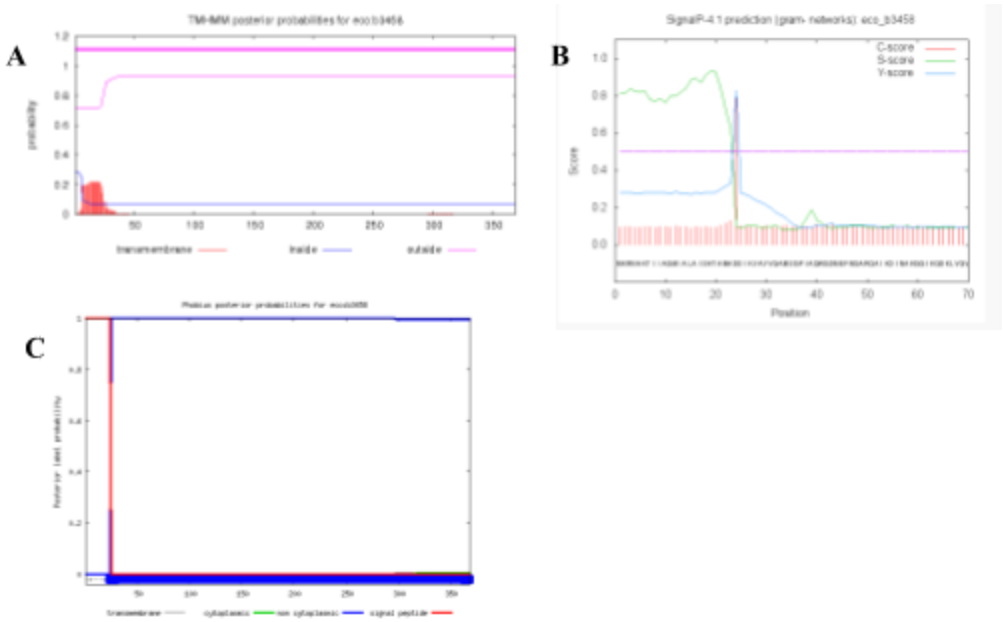


Figure 30. Cellular location analysis of b3458. Panel A: TMHMM showing no transmembrane helices, but a small peak that may indicate a signal peptide (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); Panel B: SignalP shows a noticeable peak past the central cut-off line but the numerical data does not show a signal peptide (Petersen *et al.*, 2011); Panel C: Phobius shows the signal peptide in red, no transmembrane helices, and a blue line indicating the location is non-cytoplasmic (Kall *et al.*, 2004; Kall *et al.*, 2007). The bioinformatics tools used are described in Methods.

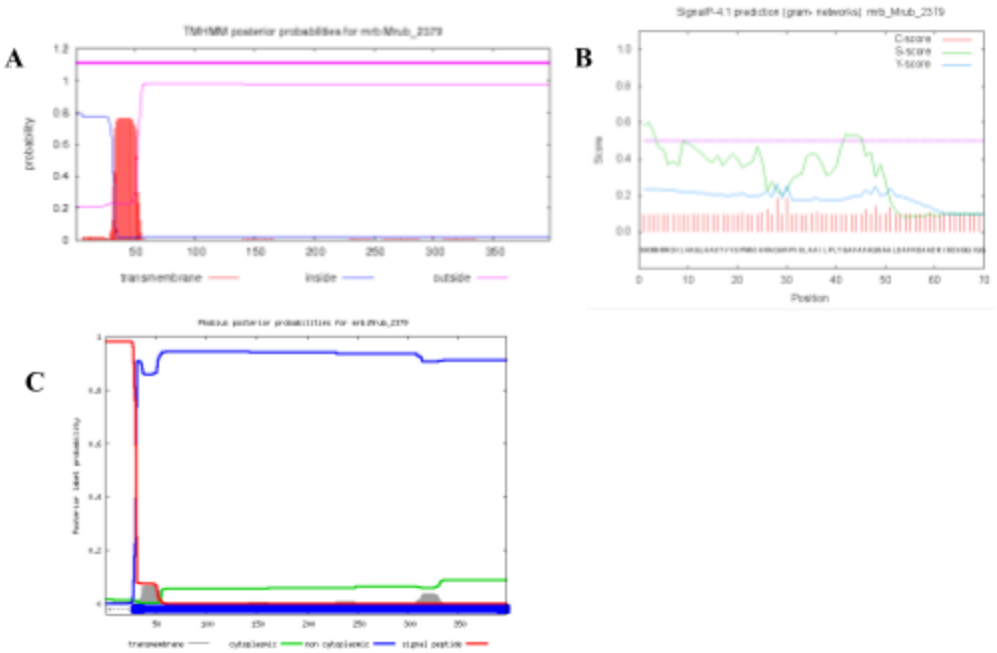


Figure 31. Determination of the cellular location of Mrub_2379. Panel A: TMHMM shows a red peak that implies a signal peptide may be present (Krogh *et al.*, 2001; Krogh *et al.*, 2016; Sonnhammer *et al.*, 1998); Panel B: SignalP shows evidence of a signal peptide because of the peak passing the central cut-off line (Petersen *et al.*, 2011); Panel C: Phobius shows the signal peptide in red, no transmembrane helices, and a blue line indicating the location is non-cytoplasmic (Kall *et al.*, 2004; Kall *et al.*, 2007). The bioinformatics tools used are described in Methods.

The program Pfam was used to obtain the name of the family that n3458 and Mrub_2379 belonged to (Finn *et al.*, 2014; Finn *et al.*, 2016); the result was that both are a part of the periplasmic binding protein family (PF13458). Pfam also provided a pairwise alignment between the gene sequence and a consensus sequence which can be seen in figure 32. The consensus sequence used was the same for both b3458 and Mrub_2379 so since both genes align well with the consensus sequence (E-values of 1.7e-61 and 4e-70, respectively) there is further evidence of the two genes being related as orthologs. The number of highly conserved amino acids is also evidence.

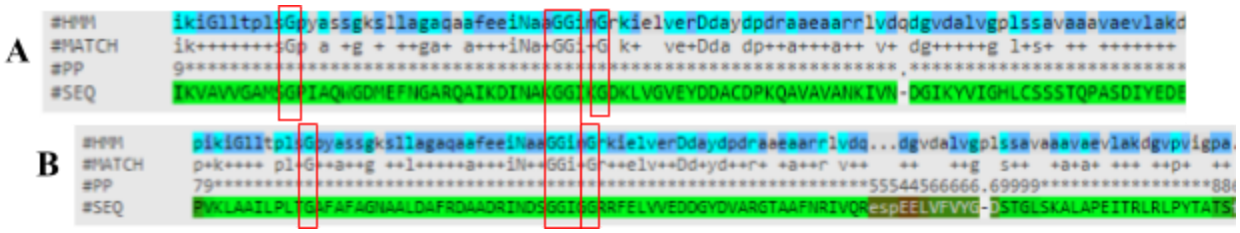


Figure 32. Consensus alignment against b3458 and Mrub_2379 (Finn *et al.*, 2014; Finn *et al.*, 2016). Panel A: The #HMM line represents the consensus sequence, #SEQ represents the b3458 sequence, #MATCH shows highly conserved amino acids; Panel B: The #HMM line represents the consensus sequence, #SEQ represents the Mrub_2379 sequence, #MATCH shows highly conserved amino acids. The highly conserved amino acids between each sequence are outlined in red; glycine residues are seen as highly conserved the most often.

The common family and domain names identified (Table 1), the initial BLASTp search (Figure 28), and similar cellular localization interpretations (Figures 30 and 31) are each significant pieces of evidence that b3458 and Mrub_2379 are orthologs. Additional support for the relationship can be derived from the fact that each gene is part of an operon and the position of each gene in the operon matches one another (Markowitz *et al.*, 2012). Both b3458 and Mrub_2379 can be found as the most downstream gene in the operon set. The information about the operon relationships is additional support of the orthologous relationship.

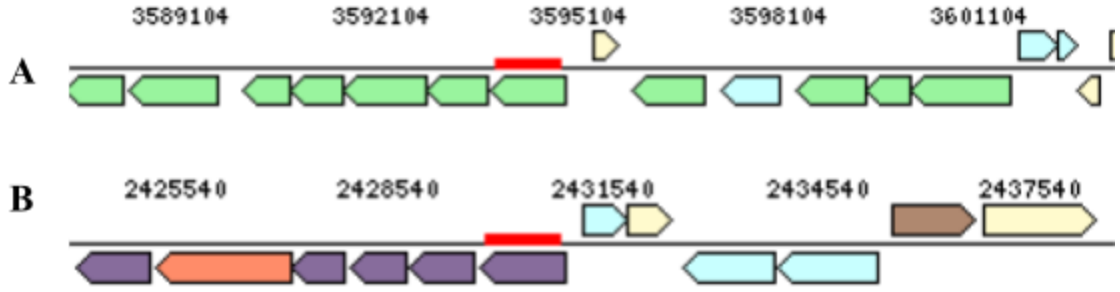


Figure 33. Visual representation of b3458 and Mrub_2379 within their operons (Markowitz *et al.*, 2012). Panel A: The Color by Kegg Chromosome Map viewer of the area surrounding b3458 with the GOI indicated by the red dash; Panel B: The output of the same program for the area surrounding Mrub_2379.

CONCLUSION

The results obtained and analyzed from the various bioinformatics tools support that the *E. coli* gene b3454 and the *M. ruber* gene Mrub_2377 are orthologs of one another. This means the organisms share a common ancestor and are related (Koonin, 2005). The first piece of evidence was obtained through the BLAST search between b3454 and the *M. ruber* genome, which pulled up Mrub_2377. After this, TMHMM, SignalP, LipoP, Phobius and PSORT-B were used to determine the cellular location of both the genes annotated. All test except for one agreed both the proteins encoded from b3454 and Mrub_2377 were in the cytoplasm. However, PSORT-B concluded Mrub_2377 was in the cytoplasmic membrane. Due to the evidence given by the other bioinformatics tools, we believe the location of the Mrub_2377 protein is in the cytoplasm, and not in the cytoplasmic membrane. With this being said, b3454 and Mrub_2377 have the same cellular location, the cytoplasm, and this is further evidence to support the claim the genes are orthologous. Additional evidence of b3454 and Mrub_2377 being orthologs is the fact that the family names produced in the top results of Pfam, TIGRfam, and CDD databases were the same for each gene, indicating functional similarity. Color by Kegg showed the genes to be in operons consisting of other genes coding for ABC transporter proteins. The only results the two genes differed at was the PSORT-B test, and as stated before, the amount of evidence contradicting this test was substantial enough to believe the PSORT-B test was inaccurate. From the substantial

amount of data obtained from a variety of bioinformatics tools, we conclude b3454 and Mrub_2377 are orthologous to one another.

The results obtained and analyzed from the various bioinformatics tools support that the *E. coli* gene b3456, and the *M. ruber* gene Mrub_2374 are orthologs of one another. The first piece of evidence was obtained through a BLAST search between b3455 and the *M. ruber* genome, which pulled up Mrub_2376. After this, TMHMM, SignalP, LipoP, Phobius and PSORT-B were used to determine the cellular location of both the genes annotated. All test except for one agreed both the proteins encoded from b3455 and Mrub_2376 were in the cytoplasm. However, PSORT-B concluded b3455 was in the cytoplasm and Mrub_2376 was in the cytoplasmic membrane. Due to the evidence given by the other bioinformatics tools, we believe the location of the Mrub_2376 protein is in the cytoplasm, and not in the cytoplasmic membrane. With this being said, b3455 and Mrub_2376 have the same cellular location, the cytoplasm, and this is further evidence to support the claim the genes are orthologous. Additional evidence of b3455 and Mrub_2376 being orthologs is the fact that the family names produced in the top results of Pfam, TIGRfam, and CDD databases were the same for each gene, indicating functional similarity. Color by Kegg showed the genes to be in operons consisting of other genes coding for ABC transporter proteins. The only results the two genes differed at was the PSORT-B test, and as stated before, the amount of evidence contradicting this test was substantial enough to believe the PSORT-B test was inaccurate. From the substantial amount of data obtained from a variety of bioinformatics tools, we conclude b3455 and Mrub_2376 are orthologous to one another.

The results obtained and analyzed from the various bioinformatics tools support that the *E. coli* gene b3456 and the *M. ruber* gene Mrub_2374 are orthologs of one another. The first piece of evidence that allowed us to draw this conclusion was the BLAST alignment that compared the amino acid sequences of the genes of interest; the low E-values and the relatively high similarity lead us to continue analysis. The summarized results of TMHMM, SignalP, LipoP, PSORT-B, and Phobius indicated that the genes were each found in the cytoplasmic membrane with several transmembrane helices passing through the membrane. Additional evidence of b3456 and Mrub_2374 being orthologs is the fact that the family names produced in the top results of Pfam, TIGRfam, and CDD databases were the same for each gene, indicating functional similarity. Although the genes were found at slightly different positions in their operons, the consistency of them each being a part of an operon is support of the orthologous relationship.

The salmon-colored gene interrupting the *M. ruber* 5-gene operon is identified as Mrub_2375. Although it is a different color in the Kegg map (indicating that it has a different function), it is transcribed in the same direction as the rest of the known operon. Additionally, the IMG/M Chromosome viewer for sequences with the same top COG hits shows that this gene is consistently found interrupting the operon at the same position across species. This is strong

evidence that Mrub_2375 may actually serve a functional purpose in the operon. The unexpected gene was identified as a "AMP-dependent synthetase and ligase" which is not usually found in an ABC transporter system. Although the result was unexpected, the fact that the gene and its location in the operon are conserved throughout recent evolutionary history (i.e. observed in closely related species) that it serves a functional purpose that may not be fully known at this time.

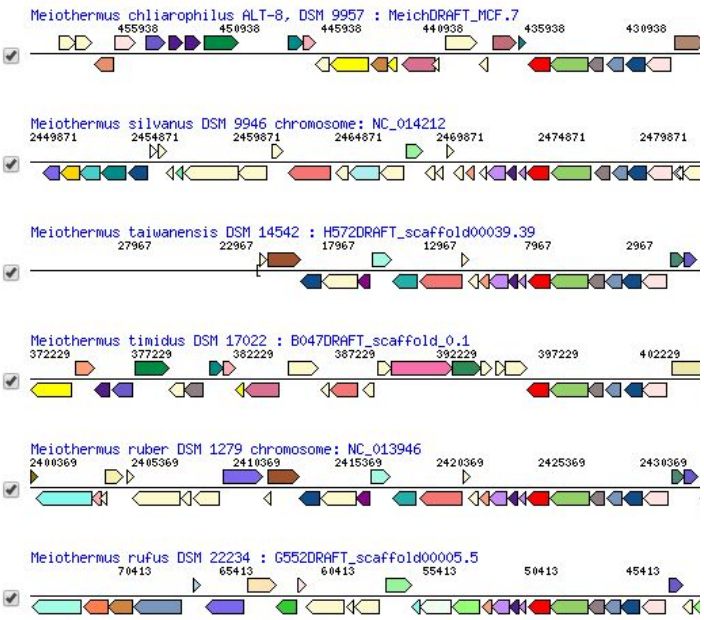


Figure 34. The IMG/M Chromosome Viewer for the areas immediately surrounding Mrub_2374 for several species sharing the same top COG hit as the gene of interest (Markowitz et al., 2012). Mrub_2374 is the red gene in each sequence, the green gene is Mrub_2375, the following 4 genes are known to be part of the expected ABC transporter system.

The various bioinformatics tools and the outputs collected support that the *E. coli* gene b3457 and the *M. ruber* gene Mrub_23748 have an orthologous relationship. First, evidence of the relationship was identified through the BLAST alignment that compared the amino acid sequences of the genes of interest and yielding a low E-value and relatively high similarity. The combined results of cellular localization tools TMHMM, SignalP, LipoP, PSORT-B, and Phobius agree that the genes are found in the cytoplasmic membrane with several transmembrane helices each. Additional evidence of b3457 and Mrub_2378 being orthologs is the fact that the family names produced in the top results of Pfam, TIGRfam, and CDD databases were the identical for the genes, even though the E-values varied slightly. The genes are also found at the same position in their respective operons that code for the ABC transporter complex which is further support that Mrub_2378 is an ortholog of b3457.

The results obtained from the a variety of bioinformatics programs described above are each important pieces of evidence that the *M. ruber* gene Mrub_2379 is an ortholog to the *E. coli* gene b3458. The initial BLASTp query sequence alignment was our first indication that the two genes were an orthologous pair. Further confirmation of the relationship between b3458 and Mrub_2379 was obtained by identifying the cellular location of each; TMHMM, SignalP, LipoP, PSORT-B, and Phobius were used to determine that both genes are found in the periplasmic space of a cell. In addition, the family and domain names of the top hits yielded from Pfam, TIGRfam and CDD are consistent between b3458 and Mrub_2379 which confirms that they belong to the same families and have similar functions. The position of each gene in its respective operon is also confirmation of an orthologous relationship. There were no major deviations in data for this set of genes.

For our site-directed mutagenesis, we chose to mutate an amino acid from the gene Mrub_2376. This gene was chosen after looking at each of the *M. ruber* genes' HMM logo (Finn *et al.*, 2014; Finn *et al.*, 2016); the goal was to identify a gene that had a highly conserved amino acid, ideally near other highly conserved amino acids for ease of identification in the protein sequence. Three highly conserved glycine residues were found near each other in the beginning of the amino acid sequence of Mrub_2376. We chose to mutate the glycine (G) residue immediately preceding the moderately conserved lysine (K) residue. The position in the HMM logo was used to identify the position in the original amino acid sequence and the position in the nucleotide sequence. The position was entered into the NEBaseChanger tool; the GGC codon was changed to GCC to convert the desired glycine to an alanine. The forward and reverse primers that would code for the new strand were provided: Q5SDM_2/9/2018_F GGGCCAATCACCGCGAACAAATC and Q5SDM_2/9/2018_R GGGCCAATCACCGCGAACAAATC.

In the proposed mutation, a glycine residue located at position 43 is replaced with a alanine amino acid. Glycine is a unique amino acid because it has a hydrogen as its side change instead of a carbon (which is the case for the other amino acids) (Betts and Russell, 2003). The small, hydrogen side chain gives the glycine amino acid a lot more flexibility than the other amino acids. The flexibility of glycine allows it to reside in parts of protein structures that are forbidden to all other amino acids. A specific example of this is glycine is usually found in tight turns in the protein. Alanine on the other hand is described as "the dullest amino acid". It does not have the flexibility glycine has, and therefore substituting an alanine for a glycine could change the structure of the protein, and therefore its function. If a conserved glycine is replaced with an alanine, this could alter the protein's shape enough so it does not function properly.

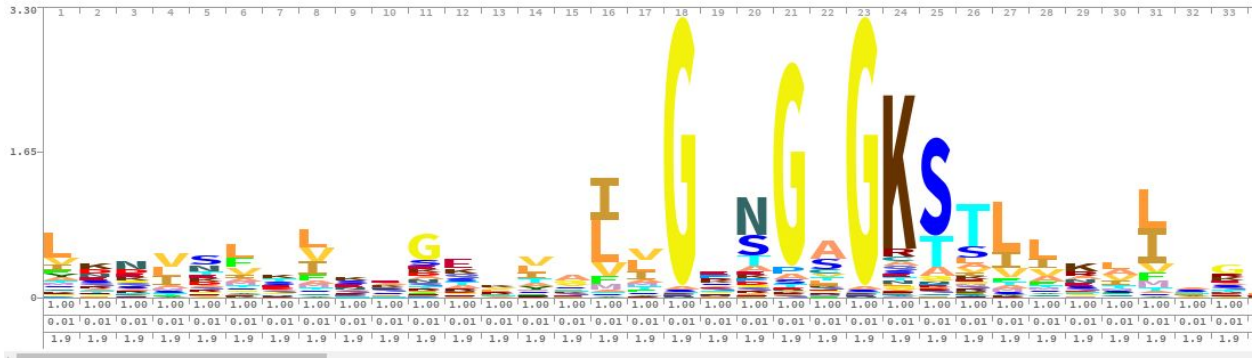


Figure 35. The first line of the HMM Logo produced from the Mrub_2376 amino acid sequence against a consensus sequence (Finn *et al.*, 2014; Finn *et al.*, 2016). The large letters indicate that the residue is highly conserved.

Through the use of numerous bioinformatics tools and the consistency seen in the data, we conclude that our hypothesis of the *M. ruber* genes Mrub_2377, Mrub_2376, Mrub_2374, Mrub_2378, and Mrub_2379 being orthologs to the respective *E. coli* genes b3454, b3455, b3456, b3457, and b3458 is correct.

References:

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403-410.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. [Internet]. The Protein Data Bank; [cited 2018 Feb 6]. Available from: http://www.rcsb.org/.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank Nucleic Acids Research, 28: 235-242.

Betts MJ and Russell RB. 2003. Amino-Acid Properties and Consequences of Substitutions. Bioinformatics for Geneticists. 311–342.

Biolabs, N. E. Home - NEB | New England Biolabs. Home - NEB | New England Biolabs. Available from: https://www.neb.com/.

Cooper GM. 2000. The Cell: A Molecular Approach. 2nd edition. Sunderland, MA: Sinauer Associates; 17.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator, Genome Research. 14:1188-1190.

Finn RD, Bateman A, Clements J, et al. 2014. Pfam: the protein families database. Nucleic Acids Research [Internet]. [cited 2018 Feb 6]. 42 (Database issue):D222-D230. Available from http://pfam.xfam.org/

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future: Nucleic Acids Res. [Internet]. [cited 2018 Feb 6] 44:D279-D285. 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.

Juncker A, Willenbrock H, von Heijne G, Nielsen H, Brunak S and Krogh A. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. 12(8):1652-62. Available at: http://www.cbs.dtu.dk/services/LipoP/.

Kall L, Krough A, Sonnhammer E. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 338(5):1027-1036.

Kall L, Krogh A, Erik LL. Sonnhammer E. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res. 35:W429-32.

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

Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, 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 RP, Paulsen I, Karp PD. 2013. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. 41:D605-612.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. 39th edition. Bethesda, Maryland: PubMed; 309.

Krogh A, Larsson B, von Heijne G, Sonnhammer E. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol, 305(3):567-580.

Krogh A, Rapacki K. 2016. TMHMM Server, v. 2.0. Cbs.dtu.dk. [Internet]. Denmark: Technical University of Denmark. [cited 2018 Feb 6]. Available from http://www.cbs.dtu.dk/services/TMHMM/

Madden T. 2002 [Updated 2003 Aug 13]. The BLAST Sequence Analysis Tool [Internet] In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information. Available from http://www.ncbi.nlm.nih.gov/books/NBK21097/ BLAST tool: BLASTp tool from https://blast.ncbi.nlm.nih.gov/Blast.cgi

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. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. [Internet]. [cited 2018 Feb 6] 43(Database issue):D222-2. Available from https://www.ncbi.nlm.nih.gov/pubmed/25414356

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 Res. 40(D1):D115-22. Available from: http://nar.oxfordjournals.org/content/40/D1/D115.full.

Moussatova A, Kandt C, O'Mara M, Tieleman P. 2008. ATP-binding cassette transporters in Escherichia coli. Calgary, Alberta: Elsevier: 1757-1771.

Notredame, Higgins, Heringa. T-Coffee: A novel method for multiple sequence alignments. 2000. JMB, 302 (205-217).

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

Phylogenetic Diversity. Joint Genome Institute. [accessed 2018 Feb 6]. https://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/

"L. Scott, Personal Communication"

Sonnhammer E, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. In J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff, and C. Sensen, editors, Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology. Menlo Park, CA: AAAI Press. p. 175-182.

Tindall BJ, Sikorski J, Lucas S, Goltsman E, Copeland A, Del Rio TG, Lapidus A. 2010. Complete genome sequence of Meiothermus ruber type strain (21T). Standards in Genomic Sciences, 3(1):26–36.

Quay SC, Dick TE and Oxender, DL. 1977. Role of transport systems in amino acid metabolism: leucine toxicity and the branched-chain amino acid transport systems. Journal of Bacteriology, 129(3):1257-1265.

Wilkens S. Structure and mechanism of ABC transporters. F1000Prime Reports. 2015 [accessed 2018 Feb 6];7.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FSL. 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics. 26(13):1608-1615.