2018

Mrub_2120, Mrub_2121, Mrub_2122, Mrub_2123 and Mrub_2124 are orthologs of *E. coli* genes b3458, b3457, b3456, b3455 and b3454, respectively, and make up an operon that codes for the branched-chain amino acid ABC transporter in *Meiothermus ruber* DSM 1279

Aaron Jones
Augustana College, Rock Island Illinois

Madelyn Huber
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_2120, Mrub_2121, Mrub_2122, Mrub_2123 and Mrub_2124 are orthologs of *E. coli* genes b3458, b3457, b3456, b3455 and b3454, respectively, and make up an operon that codes for the branched-chain amino acid ABC transporter in *Meiothermus ruber* DSM 1279

Authors: Madelyn Huber, Aaron Jones and Lori Scott

Abstract:

In this project we investigated the biological function of the genes Mrub_2120, Mrub_2121, Mrub 2122, Mrub_2123 and Mrub_2124 (KEGG map number 02010). We predict these genes encode components of a branched-chain amino acid ATP Binding Cassette (ABC) transporter: 1) Mrub_2120 (DNA coordinates 2169247-2170416 on the reverse strand) encodes the branched-chain amino acid binding protein that is localized to the periplasm; 2) Mrub_2121 (DNA coordinates 2170433..2171353 on the reverse strand) encodes the first TMD; 3) Mrub_2122 (DNA coordinates 2171365..2172279 on the reverse strand) encodes the second TMD; 4) Mrub_2123 (DNA coordinates 2172276..2173028 on the reverse strand) encodes the first NBD; 5) Mrub_2124 (DNA coordinates 2173025…2173735 on the reverse strand) encodes the second NBD. This branched-chain amino acid transport system has been found in *E. coli* K-12 MG1655 which was used as the model organism in this study. The predicted homologs of Mrub_2120, Mrub_2121, Mrub_2122, Mrub_2123 and Mrub_2124, are *livK, livH, livM, livG* and *livF,* respectively. Together, these genes form an operon encoding for an ABC transporter that selectively transports branched-chain amino acids across the intracellular plasma membrane of bacteria. This project is part of the *Meiothermus ruber* genome analysis project, which predicts gene function using the bioinformatics tools collected under the umbrella of the Guiding Education through Novel Investigation–Annotation Collaboration Toolkit (GENI-ACT).

Key Words: *Meiothermus ruber*, genome, bioinformatics, annotation, GENI-ACT, ABC transporter, branched-chain amino acid transport, permease, Mrub_2120, Mrub_2121, Mrub_2122, Mrub_2123, Mrub_2124, ATP-binding, periplasmic-binding protein

Introduction:

Transporters are a diverse and vital component of membrane transport because they work to bring things in and out of a cell across a membrane. Their diversity is indicative of many different types and functions within the class of transporters. ATP Binding Cassette (ABC) transporters have come to be known as the largest and oldest family of transporters to be identified yet. ABC transporters utilize the energy of ATP binding and hydrolysis to transport various substrates across cellular membranes (Vasilis *et al.*, 2009). As previously stated, ABC transporters are representative of the largest protein family to be identified thus far, and this is mainly because they can be located in both bacteria and eukaryotes. However, in bacteria the ABC transporters work as importers and exporters while in eukaryotes they function as strictly exporters. ABC transporters are characterized by an ATP binding subunit. The typical structure of ABC transporters is made up of 4 functional units: two nucleotide-binding domains (NBD) and two transmembrane domains (TMD). These functional units can be represented as NBD1, NBD2, TMD1 and TMD2. Within the large category of ABC transporters there are many classifications of transport systems, however, a particular one of interest for this study is the branched-chain amino acid transport system. Within humans, there are 48 different ABC transporters that can be linked to disease in humans (Wilkens, 2015).

In the case of this particular research, the system of interest is a specific type of ABC transporter classified as a branched-chain amino acid transporter. Leucine, isoleucine and valine are all classified as branched-chain amino acids. The branched-chain amino acid ABC

transporter of interest in this study is encoded by the *liv* genes. These genes enable the transfer of branched-chain amino acids from the periplasm across the intracellular plasma membrane into the cytoplasm. There are several *E.coli* locus tags associated with branched-chain amino acid transport as displayed in Figure 1: b3458, b3457, b3456, b3455 and b3454. The genes that are associated with this specific system include *livF*, *livH*, *livK*, *livM* and *livG* while the proteins associated include LivF, LivH, LivK, LivM and LivG, respectively (Quay *et al.*, 1977).

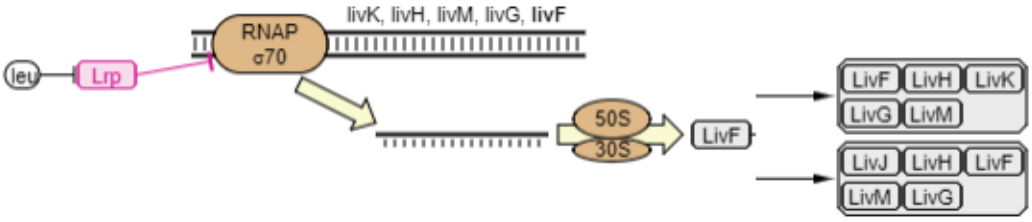


Figure 1. The branched-chain amino acid ABC transporter pathway of *E. coli.* RNA polymerase binds to the promoter region to begin the transcription of *livK, livH, livM, livG and livF* genes. DNA is then transcribed into mRNA, followed by translation of the mRNA transcript by 50S and 30S ribosomes. The result of the translation is the produced proteins LivF, LivH, LivK, LivM and LivG (Keseler *et al.*, 2013).

The organism of study used in this research is *Meiothermus ruber*. *E. coli* acts as the model organism because it has been widely studied and much is already known about it. The name *Meiothermus ruber* has been derived from the Greek words “meion” meaning lesser, “thermus” meaning hot and “ruber” meaning red. *M. ruber* is a Gram-negative, non-motile, non-spore-forming, red-pigmented thermophilic species that thrives in high temperature ares between 35-70℃, stemming from the genus of *Meiothermus* (*Tindall et al.*, 2010). The *Meiothermus ruber DSM 1279* genome is approximately 3,097,457 bp in length and includes 3,052 protein-coding genes (*Tindall et al*., 2010). It will be used to compare the *M. ruber* genome to the *E.*

coli genome in order to find orthologs within the branched-chain amino acid transporter system (Tindall *et al.*, 2010).

The *M. ruber* locus tags of interest include: Mrub_2120, Mrub_2121, Mrub_2122, Mrub_2123 and Mrub_2124. These genes of interest (GOI's) were identified using the KEGG database (Kanehisa *et al.*, 2016). An operon is known as a unit made up of linked genes that is thought to regulate other genes typically responsible for protein synthesis. In the case of *M. ruber*, it is suspected that these genes form an operon. Using all five locus tags from our *E. coli* genes (b3454-b3458), it can be shown that our model organism's genes of interest (GOI's) are all part of an operon as shown in Figure 2. Figure 2 displays the locus tags of the *E. coli* GOI's starting at the red underline with b3456 and then b3454, b3455, b3457 and b3458 on adjacent sides. All of the genes are highlighted in green, which is indicative of their similar function: membrane transport. Each gene of interest is indeed part of an operon because the four genes in sequence with the marker gene are the same color and all have similar names. Each locus tag is indicative of a different liv gene each with a different function. The genes are as follows: *livK, livH, livM, livG* and *livF,* and include the functions of branched-chain amino acid transport system substrate-binding protein, branched-chain amino acid transport system permease protein and branched-chain amino acid transport system ATP-binding protein (Kanehisa *et al*. 2016). Since the *Meiothermus ruber* GOI's are suspected to be orthologous to the model organism's, it is predicted that the *M. ruber* genes will form an operon as well.

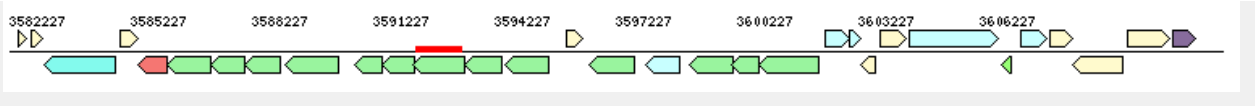


Figure 2. A segment of genes within the *Escherichia coli* K-12 genome. The genes of interest begin at amino acid position 3590747 and are transcribed in the reverse direction. Image taken from IMG/M (Markowitz *et al*., 2012).

The purpose of this study is to identify orthologs of *E. coli* and *M. ruber* within an ABC transport system. It has been suggested that the Mrub_2120 is not an ortholog of *E. coli* b3458, Mrub_2121 is an ortholog of *E. coli* b3457, Mrub_2122 is an ortholog of the *E. coli* b3456, Mrub_2123 is an ortholog of *E. coli* b3455 and Mrub_2124 is an ortholog of *E. coli* b3457 which together make up the various subunits needed for ABC transport.

Materials and Methods:

Identifying Meiothermus ruber orthologs in the E. coli K-12 genome

Nucleotide and amino acid sequences for the genes of interest (GOI) were obtained from gene details pages in the KEGG database (Kanehisa *et al.,* 2016). The ABC transporter pathway for the branched-chain amino acid transporter was obtained from KEGG for *M. ruber*, and from both KEGG (Kanehisa et al., 2016) and EcoCyc for *E. coli* (Keseler *et al*. 2013). To identify the GOIs as orthologs, each amino acid sequence was BLASTed (Altschul *et al.*, 1990; Madden, 2012) against the opposite species's genome – *M. ruber* genes were BLASTed as the query sequence against the *E. coli* K-12 genome and vice versa. The top BLAST hit for each gene was recorded as well as the hit for the suspected corresponding ortholog of interest. Significant bit scores and E-values were recorded as well. If the respective ortholog of interest was not among the list of hits (or didn't have a low enough E-value), the genes were suspected not to be orthologs.

Alternative Open Reading Frame and Clustal Alignment Analysis

To verify the correct start codon/starting point in the gene alignment, an IMG/M alternative reading frame map was utilized. The GOI was found by searching the IMG database (Markowitz *et al*., 2012) for the locus tag . Once the correct gene was identified, the alternative reading frame of amino acids was generated to look for a Shine-Delgarno sequence or other

potential start codons in the same reading frame. This would verify that the predicted start codon is correct if no other potential start codons could be identified. To generate a clustal alignment of sequences taken from multiple closely related species, T-Coffee was used (Notredame *et al.*, 2000). The sequences used for T-coffee were taken from a list of BLAST hits. The GOI's were BLASTed against the entire GEN database. Closely related species (about 8-10) were taken from the list and plugged into T-coffee along with the GOI. This generated a clustal sequence alignment that was subsequently used to create a Weblogo Panel (Crooks *et al.*, 2004). Highly conserved amino acids were identified based on the Weblogo and recorded in the lab notebook. The purpose of these two tools was to see how well the start codons lined up across multiple related species as well as determine how conserved the starting methionine was in these species. This analysis was done for *M. ruber* genes only. The start codons for *E. coli* are assumed to be accepted (control).

Paralog Identification

Paralogs were identified using the gene details pages located in the KEGG database (Kanehisa *et al.*, 2016). These were the same gene details pages used to obtain the FASTA sequences used for all the bioinformatics tools. "Paralog" was selected from the SSDB information, and the number of paralogs of each GOI as well as the names of some of the top hits were recorded in the lab notebook.

Cell Localization Information

The bioinformatics tool PSORT-B (Yu *et al.*, 2010) was used to determine the cell localization of each *M. ruber* gene and its corresponding ortholog in *E. coli*. Amino acid sequences of the genes were plugged into the PSORT-B query, and localization scores were assigned to each of several locations in the cell: cytoplasmic membrane, cytoplasm, periplasm,

outer membrane and extracellular based on a scale of 10. The highest score indicated the predicted location for the protein product of each gene.

To determine the number of transmembrane helices, TMHMM software was utilized (Krogh *et al*., 2001; Krogh and Rapacki, 2016; Sonnhammer *et al.*, 1998). Amino acid sequences for each gene were plugged into the query box, and a TMHMM graph outlying the number of transmembrane helices was generated. SignalP (Petersen *et al.*, 2011) and LipoP (Juncker *et al.*, 2003) were used to predict the presence of a signal peptide using probabilities and identify the type of signal peptide. Amino acid sequences were plugged into the query box of each tool and probability plots were generated. Gram-negative bacteria was selected as the query type. Phobius was also utilized in this same manner (Kall *et al.*, 2004; Kall *et al*., 2007). If signal peptidases were present, the cleavage sites were also generated by these data.

Determining Structural and Functional Features

To determine the functional and structural characteristics of *M. ruber* and *E. coli* orthologs, CDD, TIGRFAM, Pfam and PDB bioinformatics tools were utilized. CDD (Marchler Bauer *et al*., 2015) was accessed by clicking the superfamily name at the top of the previously generated BLAST outputs. COG names and numbers were recorded for the top two hits of each gene as well as the bit scores and E-values. The COG's of the orthologs were then compared for functional similarities.

For TIGRFAM (Haft *et al.*, 2001), amino acid sequences were plugged into the query and submitted for analysis. The data was filtered to show only hits below an E-value of 0.01 (significance cut-off). Similar to CDD, TIGRFAM numbers and names were recorded and compared to the suspected corresponding orthologs for functional similarities. Bit scores and E-values were recorded as well.

For Pfam (Finn *et al.*, 2014; Finn *et al.*, 2016), amino acid sequences were plugged into the query box analyzed to determine the protein family and clan. Family names and numbers, and clan names and numbers were recorded along with their bit scores and E-values. Additionally, HMM Weblogos generated on Pfam were catalogued showing highly conserved amino acids in each of the known proteins. This logo contained only genes matching the Pfam hits, not the direct GOI plugged into the query.

PDB (Berman *et al*., 2000 [Internet]; Berman *et al*., 2000) was used to determine the structural properties of the proteins encoded by the GOI's. Amino acid sequences were plugged into the query box and submitted for analysis. A positive hit is indicative of a protein that has been successfully crystallized and catalogued. No hit indicates that the protein's structure has yet to be verified via crystallography.

Operon Verification

To verify that the GOI's operate as part of an operon, IMG/M (Markowitz *et al*., 2012) was used once again. As previously done, the gene detail pages were found by searching for the locus tags of the individual genes. A chromosome map was then generated using the "color by KEGG" option. This produces a map of the entire chromosome detailed with the functional properties of each individual gene on the chromosome. Operons were identified by looking at the colors of neighboring genes to see if they were the same color and were being transcribed in the same direction. After this, another chromosome map was generated. This time, however, the GOI's were compared to the homologs of other related species. Operons were identified on this map by how well the suspected operon stayed conserved across the different species (Markowitz *et al*., 2012).

Site-Directed Mutagenesis

The last activity done in this study was a site-directed mutagenesis of a one of the GOI's (both *M. ruber* and *E. coli* orthologs). The site-directed mutagenesis was performed on Mrub_2122 and b3456. Using NEBaseChanger (Biolabs; Betts and Russell, 2003), a highly conserved glycine residue at amino acid position 50 (Mrub) and 51 (*E. coli*) was deleted from the sequence. This mutation will be carried out in later experiments using wet-lab techniques, and functional genomics will be performed to see how the mutation affects the degree to which *M. ruber* and *E. coli* can transport branched-chain amino acids with a mutated ABC transporter.

Results:

Table 1. *E.coli* b3458 and Mrub_2120 are orthologs

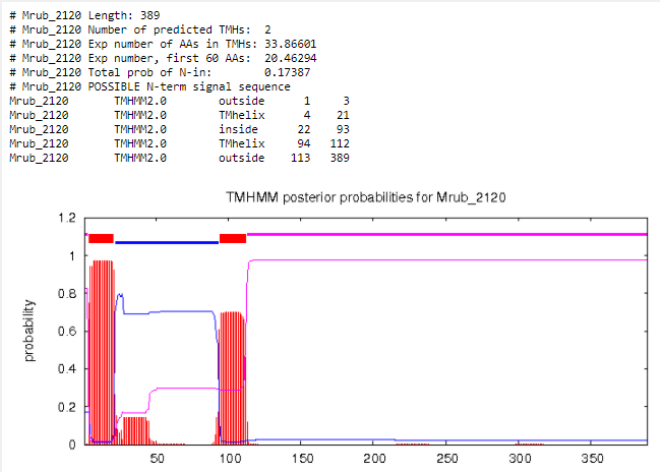
Bioinformatics Tool Used	**E.coli Gene b3458**	**Mrub_2120**
BLAST against opposite genome	Branched-chain amino acid ABC transporter substrate-binding protein [Meiothermus ruber] E-value: 5e-51	
CDD Data	LivK (COG0683)	
	E-value: 2.40e-102	E-value: 8.79e-59
Cellular Localization	Anchored to the periplasmic side of the inner membrane	
TIGRfam- Protein Family	urea_ABC_UrtA: urea ABC transporter (TIGR03407)	
	E-value: 4.2e-4	E-value: 7.8e-8
Pfam- Protein Family	Periplasmic Binding Protein (PF13458)	
	E-value: 1.7e-61	E-value: 1.6e-71

Protein Database	1USG L-leucine-binding protein E-value: 0.00	4EVQ Crystal structure of ABC transporter from R. palustris - solute binding protein (RPA0668) in complex with 4-hydroxybenzoate E-value: 8.70812E-42
KEGG Pathway Map	ABC Transporters KEGG Number: 02010	

Table 1 is a summary of all of the bioinformatics data collected in order to compare the orthologs *E. coli* b3458 gene to Mrub_2120. The first row of data is the comparison of results after running a BLAST against each gene. In the case of these two genes there was in fact a significant hit when the sequences were run against each other. The branched-chain amino acid ABC transporter substrate-binding protein [Meiothermus ruber] had an e-value of about 5e-51, showing its significance. The next row shows the data collected from the CDD when the BLAST was run. The CDD found the same name and COG number when each of the gene sequences was run through the database. The COG number and name were found to be COG0683 and LivK, respectively. Both genes also had very small E-values, showing the significance of the results. Bioinformatics tools such as SignalP, TMHMM, LipoP and PSORT-B suggested that both genes are located on the periplasmic side of the inner membrane. The TIGRFAM database also pulled up the same TIGRFAM name and number for both genes of TIGR03407 and urea ABC transporter. Furthermore, the Pfam database also pulled up the same Pfam name and number for both genes: PF13458 and periplasmic binding protein. Both E-values were low, indicating significance. Finally, both genes are involved in the same ABC transporter pathway. Lastly, when run through the PDB database each gene had a hit. *E. coli* b3458 had a hit labeled 1US L-

leucine-binding protein which gave an e-value of exactly 0.00 while Mrub_2120 had a hit labeled 4EVQ Crystal structure of ABC transporter from R. palustris - solute binding protein. Both hits for *E.coli* b3458 and Mrub_2120 were indicative of their individual functions and roles they play within the transport system.

Panel A



Panel B

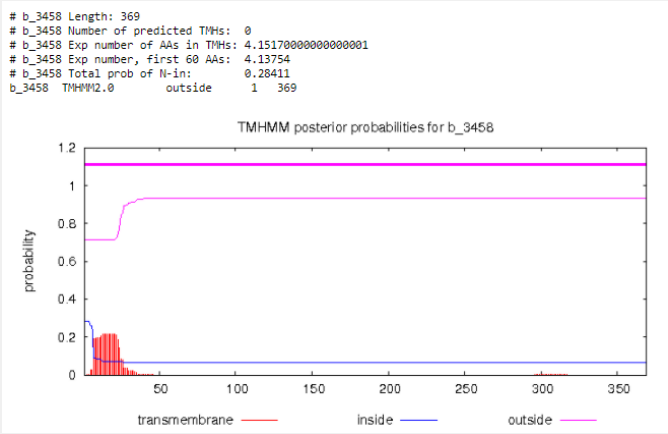
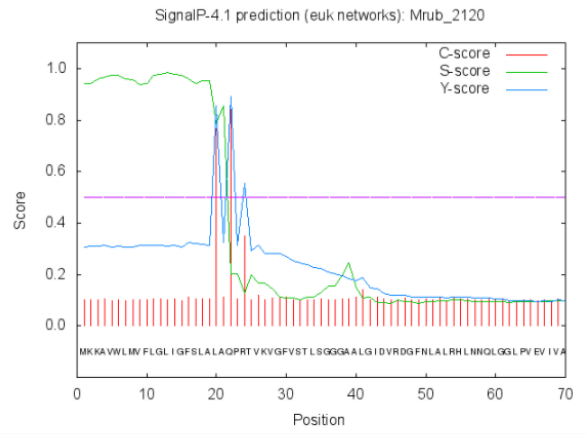


Figure 2. Mrub_2120 and *E.coli* b3458 both contain slightly different THM regions. Panel A is the TMHMM for Mrub_2120. Panel B is the TMHMM for *E.coli* b3458. The predicted locations for the two are that they are anchored to the periplasmic side of the inner membrane.

The plots shown in figure 2 are the hydropathy plots for both Mrub_2120 and *E.coli* b3458. The red peaks on each plot are indicative of transmembrane helices. Mrub_2120 (Panel A) contains two transmembrane helices while *E.coli* b3458 contains zero (Panel B), demonstrating that the proteins could have different structures. The presence of transmembrane

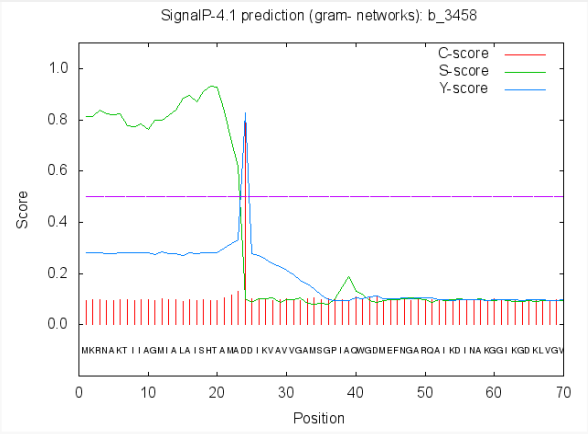
helices suggests that the protein coded by Mrub_2120 may be embedded in or anchored to the inner plasma membrane. The protein coded by b3458 is likely anchored to the periplasmic side of the inner membrane because there was a slight TMH signal. Specifically, these genes are believed to code for substrate binding proteins located in the periplasmic space.

Panel A




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    22       0.842
  max. Y    22       0.893
  max. S    13       0.985
  mean S     1-21    0.947
       D     1-21    0.922   0.450   YES
Name=Mrub_2120  SP='YES' Cleavage site between pos. 21 and 22: ALA-QP D=0.922 D-cutoff=0.450 Networks=SignalP-noTM
# data
```

Panel B

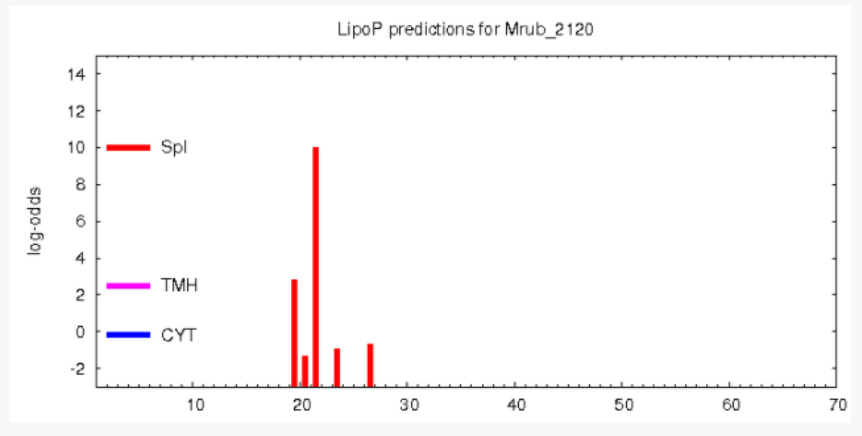



```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    24       0.817
  max. Y    24       0.827
  max. S    19       0.933
  mean S     1-23    0.820
       D     1-23    0.823   0.570   YES
Name=b_3458     SP='YES' Cleavage site between pos. 23 and 24: AMA-DD D=0.823 D-cutoff=0.570 Networks=SignalP-noTM
```

Figure 3. SignalP Data. Both the Mrub_2120 and *E.coli* b3458 contain a signal peptide and a cleavage site. Panel A shows the plot for Mrub_2120. Panel B shows the plot for *E.coli* b3458.

SignalP graphs are displayed in figure 3 for both Mrub_2120 and *E.coli* b3458. The cleavage sight for both were predicted. Mrub_2120 had a cleavage sight after amino acid 21 while *E.coli* b3456 had a cleavage sight after amino acid 23. The gathered information tells us that both genes of interest contain a signal peptide further confirming the localization of both Mrub_2120 and *E.coli* b3458 to the inner membrane.

Panel A



Panel B

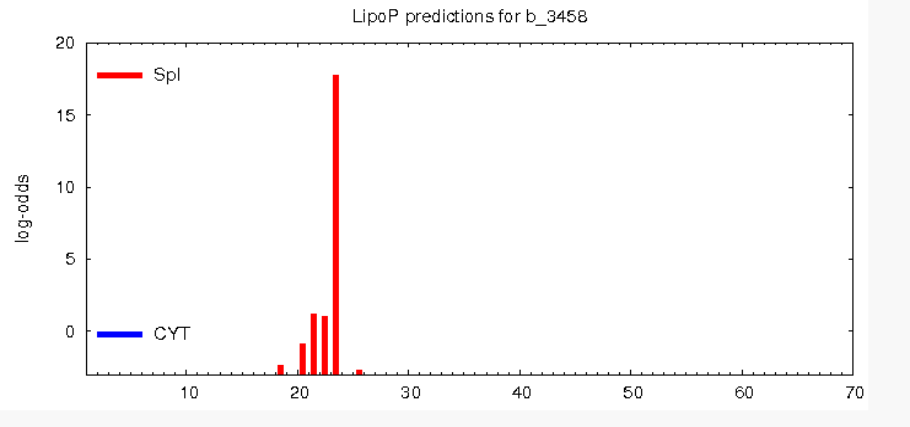
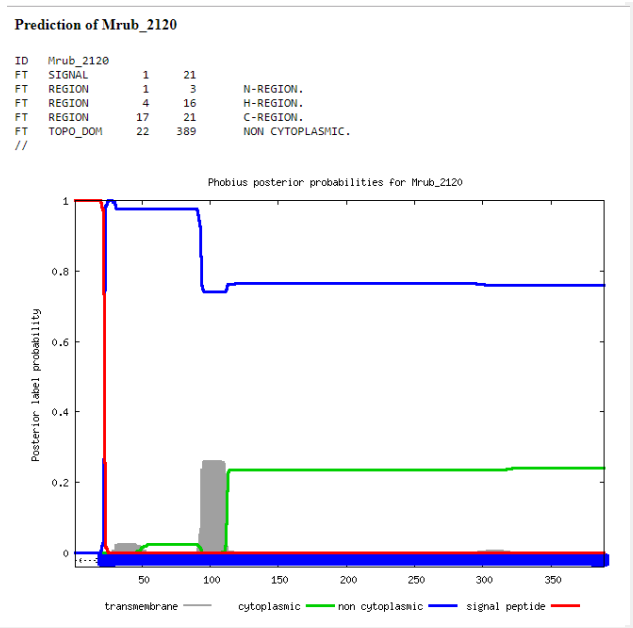


Figure 4. The given plots show the amino acid where the cleavage site is located, thus confirming the presence of a signal peptide. Panel A show the LipoP prediction for Mrub_2120. Panel B shows the LipoP prediction for E.coli b3458.

LipoP graphs are displayed in figure 4 for both Mrub_2120 and *E.coli* b3458. Mrub_2120 was found to have a cleavage sight after amino acid 21 while *E.coli* b3458 one after

amino acid 23. The graph for Mrub_2120 also further confirmed the presence of transmembrane helices while the graph for *E.coli* b3458 confirms the lack of transmembrane helices.

Panel A



Panel B

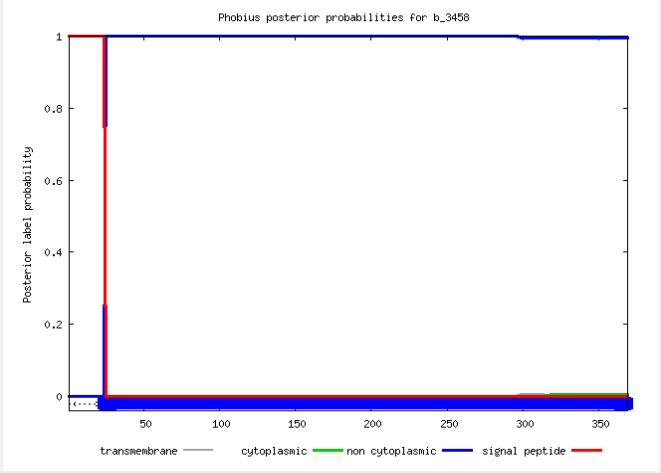
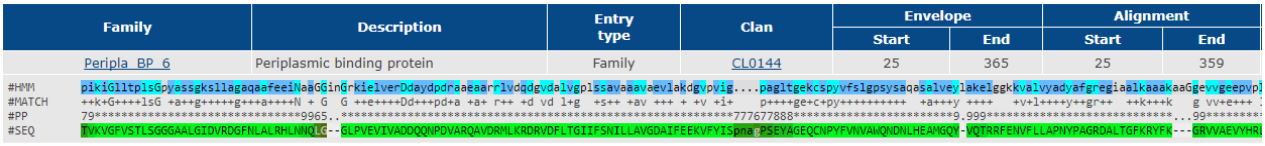


Figure 5. The given graphs display the possible locations of the given genes of interest. Panel A show the Phobius graph for Mrub_2120. Panel B shows the Phonius graph for *E.coli* b3458.

The Phobius graphs are displayed in figure 5 for both Mrub_2120 and *E.coli* b3458. Mrub_2120 was found to fluctuate from the cytoplasm and non-cytoplasm as any transporter would however, it is not localized in the cytoplasm. The *E.coli* b3458 Phobius graph displays

that it is strictly non-cytoplasmic. Both Phobius graphs also further confirm the presence of signal peptides in both Mrub_2120 and *E.coli* b3458.

Panel A



Panel B

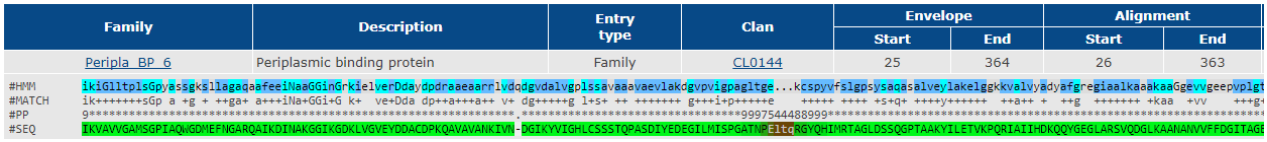
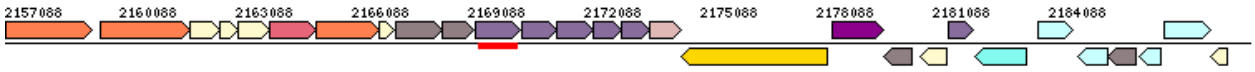


Figure 6. Mrub_2120 and b3458 contain most of the same highly conserved amino acids. Both genes code for the same domain Peripla_BP_6. Panel A shows the pairwise alignment for Mrub_2120. Panel B shows the pairwise alignment for *E.coli* b3458.

The pairwise alignments shown above in figure 6 several similar highly conserved amino acids. Both Mrub_2120 and *E.coli* b3458 contain a few highly conserved glycines toward the beginning of their protein sequences. The fact that these two protein sequences share conserved residues is yet another reason the two are orthologous to each other.

Panel A



Panel B

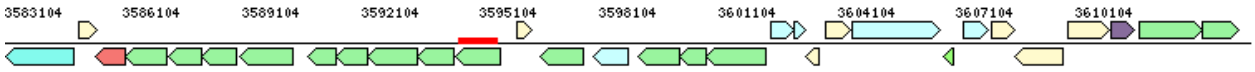
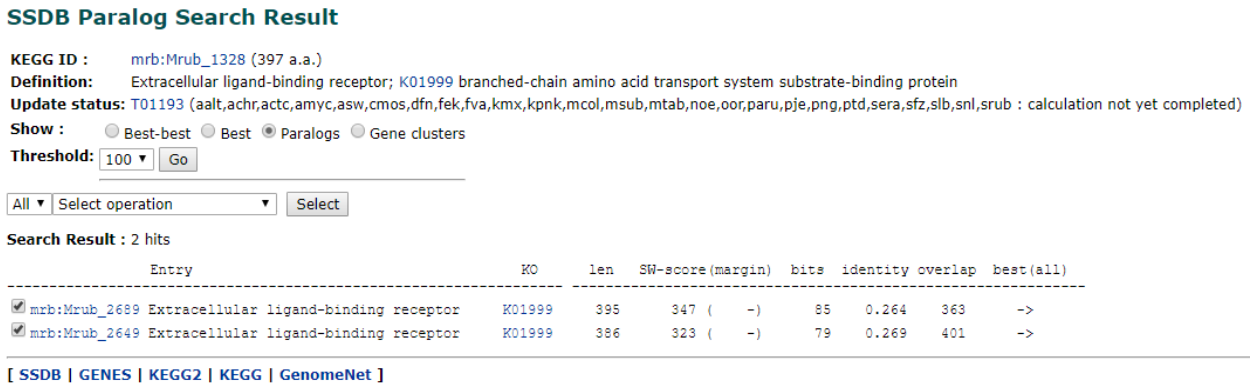


Figure 7. Mrub_2120 and *E.coli* b3458 are both a part of an operon. Panel A is the Mrub_2120 gene sequence. Panel B is the *E.coli* b3458 gene sequence.

The gene sequences shown in figure 7 display that both Mrub_2120 and *E.coli* b3458 are a part of an operon. The similar colors upstream and downstream are indicative of an the same function and possibly pathway, indicating an operon is present. Mrub_2120 is in a light purple shade indicative of membrane transport while *E.coli* b3458 is a light green indicative of its involvement in amino acid metabolism. The evidence shown here further supports that these two genes are orthologs.

Panel A



SSDB Paralog Search Result

KEGG ID : mrb:Mrub_1328 (397 a.a.)
Definition: Extracellular ligand-binding receptor; K01999 branched-chain amino acid transport system substrate-binding protein
Update status: T01193 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculation not yet completed)
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

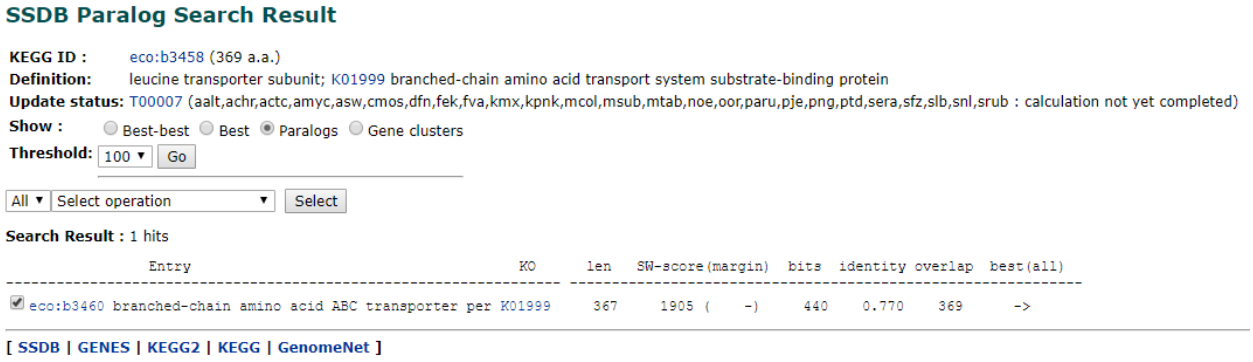
All Select operation Select

Search Result : 2 hits

Entry		KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
mrb:Mrub_2689	Extracellular ligand-binding receptor	K01999	395	347	(-)	85	0.264	363	->
mrb:Mrub_2649	Extracellular ligand-binding receptor	K01999	386	323	(-)	79	0.269	401	->

[SSDB | GENES | KEGG2 | KEGG | GenomeNet]

Panel B



SSDB Paralog Search Result

KEGG ID : eco:b3458 (369 a.a.)
Definition: leucine transporter subunit; K01999 branched-chain amino acid transport system substrate-binding protein
Update status: T00007 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculation not yet completed)
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

All Select operation Select

Search Result : 1 hits

Entry		KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
eco:b3460	branched-chain amino acid ABC transporter per	K01999	367	1905	(-)	440	0.770	369	->

[SSDB | GENES | KEGG2 | KEGG | GenomeNet]

Figure 8. Paralogs of Mrub_2120 and b3458. Panel A shows the results of a DB paralog search through the KEGG database for Mrub_2120. Panel B is the results of the same search using b3458.

The paralog data present in figure 8 depicts the known paralogs of the genes of interest. Mrub_2120 was found to have two paralogs present while b3458 only had one. The relatively close number of paralogs between these two genes is further evidence indicating that the two are orthologous to each other.

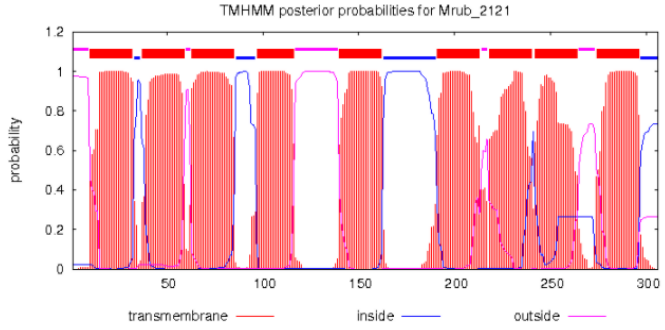
Table 2. *E.coli* b3457 and Mrub_2121 are orthologs

Bioinformatics Tool Used	**E.coli Gene b3457**	**Mrub_2121**
BLAST against opposite genome	Mrub_2121 Score: 92 E-value: 3e-22	Two results found but no b3457
CDD Data	LivH (COG0559)	
	E-value: 4.91e-87	E-value:1.08e-52
Cell Localization	Embedded in the Cell Membrane	
TIGRfam- Protein Family	urea_t_UrtC_arc: urea ABC transporter, perm (TIGR03409)	
	E-value: 2.4e-15	E-value: 3.1e-31
Pfam- Protein Family	BPD_transp_2 (PF02653)	
	E-value: 6.7e-71	E-value: 1e-40
Protein Database	No hits	
KEGG Pathway Map	ABC Transporters KEGG Number: 02010	

Table 2 is a summary of all of the bioinformatics data collected in order to compare the orthologs *E.coli* b3457 and Mrub_2121. The first row of data is the comparison of results after running a BLAST against each gene. While the *E.coli* sequence BLAST against *M. ruber* had a

hit on the gene Mrub_2121 with a fairly low e-value, the BLAST of the Mrub_2121 sequence show several hits but none that matched that of b3457. The next row shows the data collected from the CDD when the BLAST was run. The CDD found the same name and COG number when each of the gene sequences was run through the database. The COG number and name were found to be COG0559 and LivH, respectively. Both genes also had very small e-values, showing the significance of the results. Bioinformatics tools such as SignalP, THM, LipoP and PSORT-B suggested that both genes are embedded in the cell membrane. The TIGRfam database also pulled up the same TIGRfam name and number for both genes of TIGR03409 and urea ABC transporter permease. Furthermore, the pfam database also pulled up the same pfam name and number for both genes PF02653 and BPD_transp_2. Both e-values were low indicative of significance. Finally, both genes are involved in the same ABC transporter pathway.

Panel A



Panel B

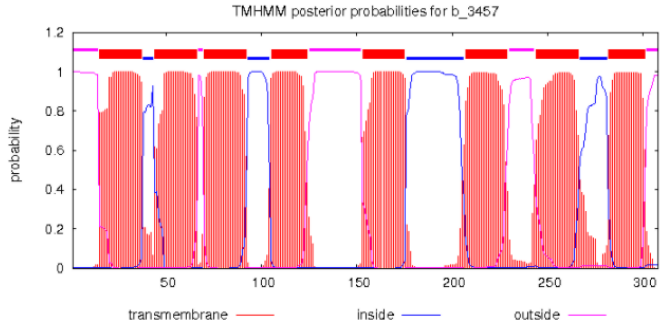
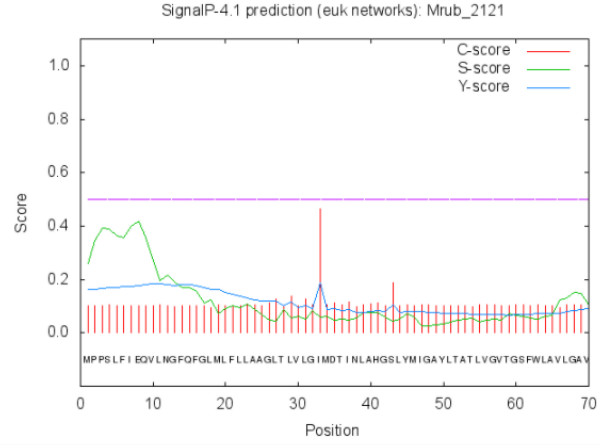


Figure 9. Mrub_2121 and *E.coli* b3457 both contain similar THM regions. Panel A is the TMHMM for Mrub_2121. Panel B is the TMHMM for *E.coli* b3457. The predicted locations for the two are that they are embedded in the cell membrane.

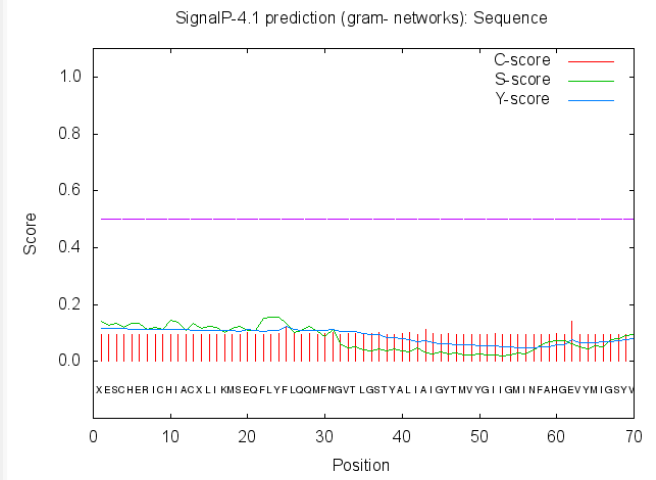
The plots shown in figure 9 are the hydropathy plots for both Mrub_2121 and *E.coli* b3457. The red peaks on each plot are indicative of transmembrane helices. Mrub_2121 (Panel A) contains nine transmembrane helices while *E.coli* b3457 contains eight (Panel B), demonstrating that the proteins are of similar structure. The presence of transmembrane helices suggests that the proteins coded by these genes are localized/embedded in the cell membrane. Specifically, these genes are believed to code for one of the TMD's of the ABC transporter.

Panel A




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    33       0.466
  max. Y    33       0.187
  max. S     8       0.416
  mean S     1-32    0.185
       D     1-32    0.186   0.500   NO
Name=Mrub_2121  SP='NO' D=0.186 D-cutoff=0.500 Networks=SignalP-TM
```

Panel B




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    62       0.142
  max. Y    25       0.123
  max. S    24       0.157
  mean S     1-24    0.127
       D     1-24    0.124   0.510   NO
Name=Sequence   SP='NO' D=0.124 D-cutoff=0.510 Networks=SignalP-TM
```

Figure 10. Both the Mrub_2121 and *E.coli* b3457 both do not contain a cleavage site and therefore also do not contain a signal peptide. Panel A shows the plot for Mrub_2121. Panel B shows the plot for *E.coli* b3457.

SignalP graphs are displayed in figure 10 for both Mrub_2121 and *E.coli* b3457. There was no cleavage sight predicted for Mrub_2121 or *E.coli* b3457. The gathered information tells us that both genes of interest do not contain a signal peptide further confirming the localization of both Mrub_2121 and *E.coli* b3457 to the cell membrane.

Panel A

```
# Mrub_2121 TMH score=14.1447 margin=14.345613
# Cut-off=-3
Mrub_2121       LipoP1.0:Best   TMH     1       1       14.1447
Mrub_2121       LipoP1.0:Margin TMH     1       1       14.345613
Mrub_2121       LipoP1.0:Class  CYT     1       1       -0.200913


# NO PLOT made - less than 4 putative cleavage sites predicted
```

Panel B

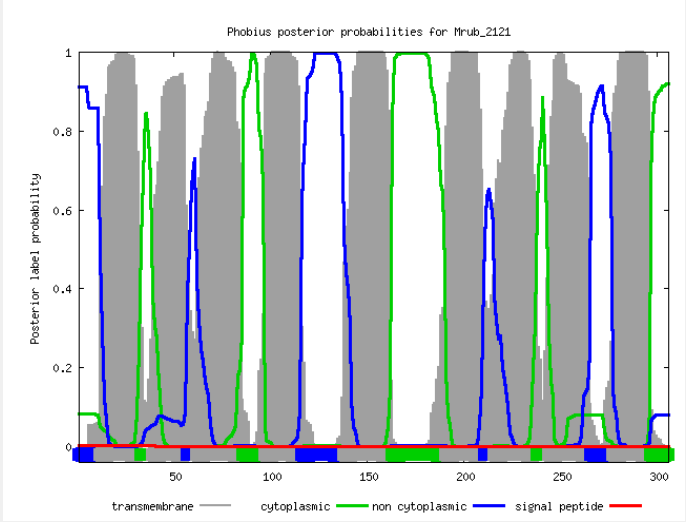
```
# b_3457 TMH score=16.6878 margin=16.888713
# Cut-off=-3
b_3457  LipoP1.0:Best   TMH     1       1       16.6878
b_3457  LipoP1.0:Margin TMH     1       1       16.888713
b_3457  LipoP1.0:Class  CYT     1       1       -0.200913


# NO PLOT made - less than 4 putative cleavage sites predicted
```

Figure 11. The images produced from LipoP display that there are less than four putative cleavage sites. Panel A show the LipoP prediction for Mrub_2121. Panel B shows the LipoP prediction for *E.coli* b3457.

LipoP graphs are displayed in figure 11 for both Mrub_2121 and *E.coli* b3457. Neither Mrub_2121 nor *E.coli* b3457 were shown to have signal peptides. The lack of a signal peptide was indicated by the images produced by LipoP showing less than four putative cleavage sites for both genes of interest.

Panel A



Panel B

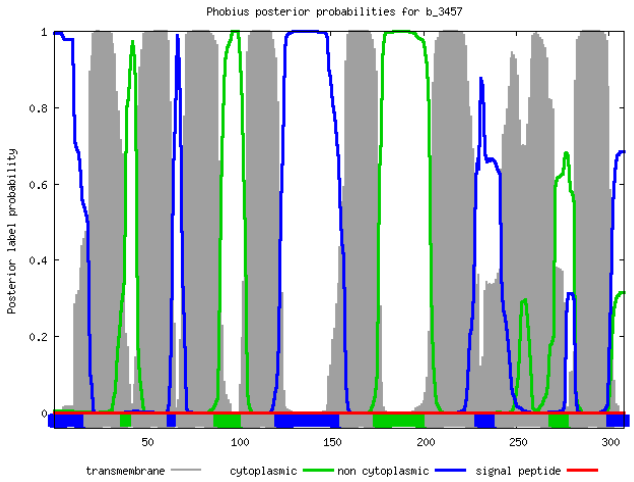
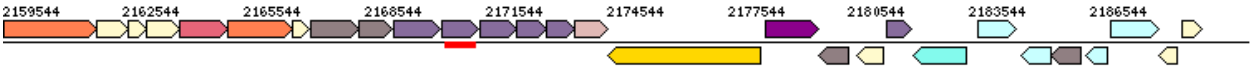


Figure 12. The given graphs display the possible locations of the given genes of interest. Panel A show the Phobius graph for Mrub_2121. Panel B shows the Phonius graph for *E.coli* b3457.

The Phobius graphs displayed in figure 12 for both Mrub_2121 and *E.coli* b3457 are very similar. Both Mrub_2121 and *E.coli* b3457 were found to fluctuate from the cytoplasm and non-cytoplasm as any transporter would however, it is not localized in the cytoplasm. Both Phobius graphs also further confirm the lack of signal peptides in both Mrub_2121 and *E.coli* b3457.

Panel A



Panel B

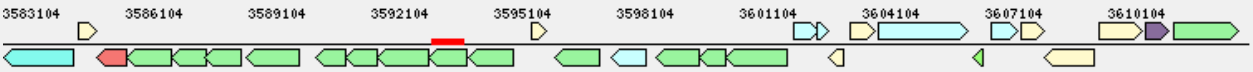
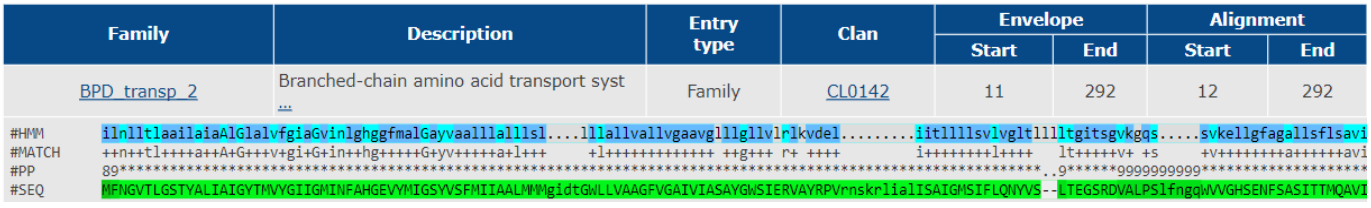


Figure 13. Mrub_2121 and *E.coli* b3457 are both a part of an operon. Panel A is the Mrub_2121 gene sequence. Panel B is the *E.coli* b3457 gene sequence.

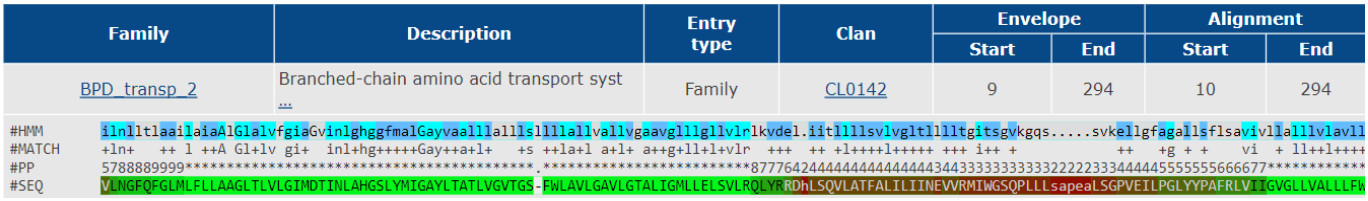
The gene sequences shown in figure 13 display that both Mrub_2121 and *E.coli* b3457 are a part of an operon. The similar colors upstream and downstream are indicative of the same function and possibly pathway, indicating an operon is present. Mrub_2121 is in a light purple shade indicative of membrane transport while *E.coli* b3457 is a light green indicative of its involvement in amino acid metabolism. The evidence shown here further supports that these two genes are orthologs.

Panel A



Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
BPD_transp_2	Branched-chain amino acid transport syst ...	Family	CL0142	11	292	12	292

Panel B

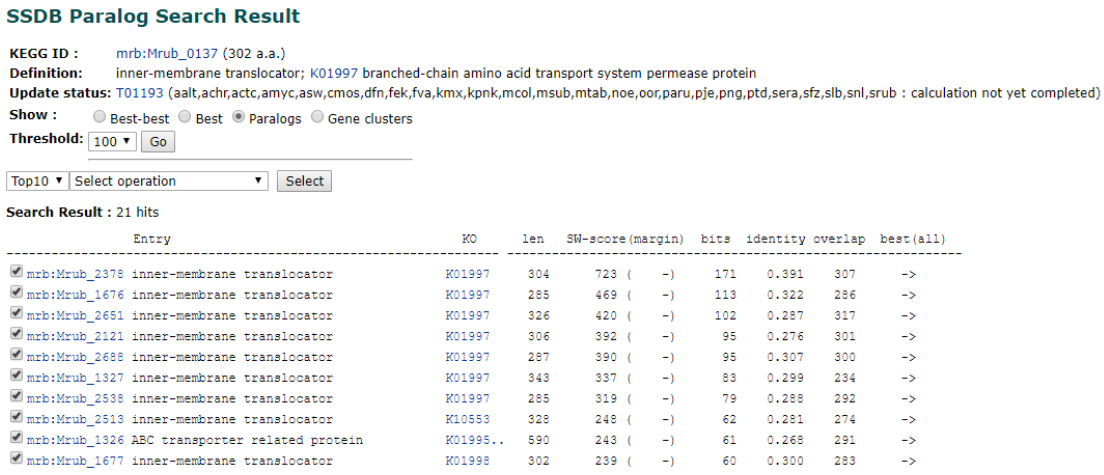


Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
BPD_transp_2	Branched-chain amino acid transport syst ...	Family	CL0142	9	294	10	294

Figure 14. Mrub_2121 and b3457 contain most of the same highly conserved amino acids. Both genes code for the same domain BPD_transp_2. Panel A shows the pairwise alignment for Mrub_2121. Panel B shows the pairwise alignment for *E.coli* b3457.

The pairwise alignments shown above in figure 14 contain a lot of the same highly conserved amino acids. Both Mrub_2121 and *E.coli* b3457 contain highly conserved alanine and glycine toward the beginning of their protein sequences. The fact that these two protein sequences share conserved residues is yet another reason the two are orthologous to each other.

Panel A



SSDB Paralog Search Result

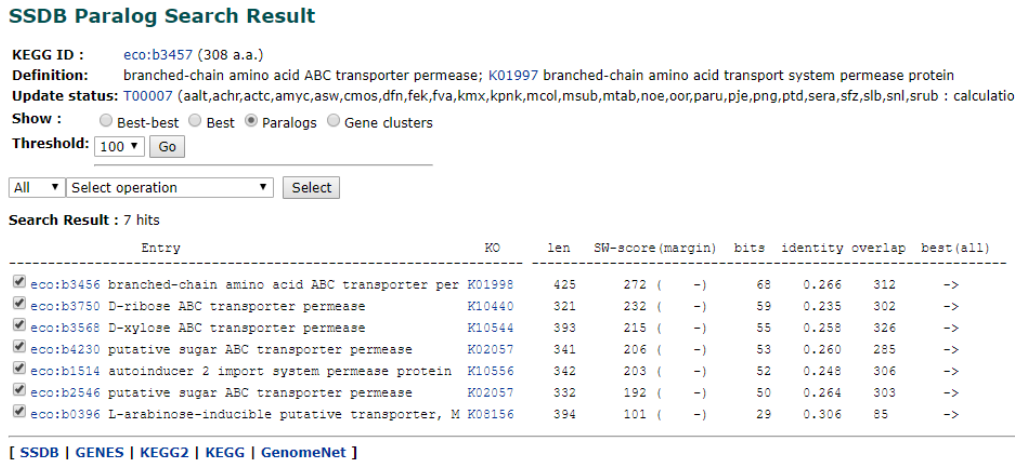
KEGG ID : mrb:Mrub_0137 (302 a.a.)
Definition: inner-membrane translocator; K01997 branched-chain amino acid transport system permease protein
Update status: T01193 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculation not yet completed)
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

Top10 Select operation Select

Search Result : 21 hits

Entry	KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
mrb:Mrub_2378 inner-membrane translocator	K01997	304	723	(-)	171	0.391	307	->
mrb:Mrub_1676 inner-membrane translocator	K01997	285	469	(-)	113	0.322	286	->
mrb:Mrub_2651 inner-membrane translocator	K01997	326	420	(-)	102	0.287	317	->
mrb:Mrub_2121 inner-membrane translocator	K01997	306	392	(-)	95	0.276	301	->
mrb:Mrub_2688 inner-membrane translocator	K01997	287	390	(-)	95	0.307	300	->
mrb:Mrub_1327 inner-membrane translocator	K01997	343	337	(-)	83	0.299	234	->
mrb:Mrub_2538 inner-membrane translocator	K01997	285	319	(-)	79	0.288	292	->
mrb:Mrub_2513 inner-membrane translocator	K10553	328	248	(-)	62	0.281	274	->
mrb:Mrub_1326 ABC transporter related protein	K01995..	590	243	(-)	61	0.268	291	->
mrb:Mrub_1677 inner-membrane translocator	K01998	302	239	(-)	60	0.300	283	->

Panel B



SSDB Paralog Search Result

KEGG ID : eco:b3457 (308 a.a.)
Definition: branched-chain amino acid ABC transporter permease; K01997 branched-chain amino acid transport system permease protein
Update status: T00007 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculatio
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

All Select operation Select

Search Result : 7 hits

Entry	KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
eco:b3456 branched-chain amino acid ABC transporter per	K01998	425	272	(-)	68	0.266	312	->
eco:b3750 D-ribose ABC transporter permease	K10440	321	232	(-)	59	0.235	302	->
eco:b3568 D-xylose ABC transporter permease	K10544	393	215	(-)	55	0.258	326	->
eco:b4230 putative sugar ABC transporter permease	K02057	341	206	(-)	53	0.260	285	->
eco:b1514 autoinducer 2 import system permease protein	K10556	342	203	(-)	52	0.248	306	->
eco:b2546 putative sugar ABC transporter permease	K02057	332	192	(-)	50	0.264	303	->
eco:b0396 L-arabinose-inducible putative transporter, M	K08156	394	101	(-)	29	0.306	85	->

[SSDB | GENES | KEGG2 | KEGG | GenomeNet]

Figure 15. Paralogs of Mrub_2121 and b3457. Panel A shows the results of a DB paralog search through the KEGG database for Mrub_2121. Panel B is the results of the same search using b3457.

The paralog data present in figure 15 depicts the known paralogs of the genes of interest. Mrub_2121 was found to have twenty one paralogs present while b3457 only had seven. While the number of paralogs between these two genes is quite different, these two genes are still thought to be orthologous to each other given the other bioinformatics evidence.

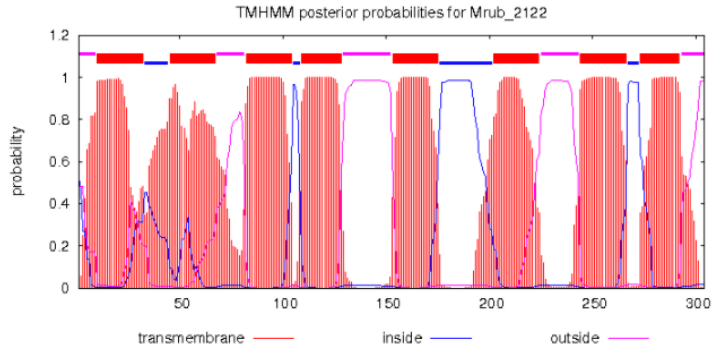
Table 3. *E.coli* b3456 and Mrub_2122 are orthologs

Bioinformatics Tool Used	**E.coli Gene b3456**	**Mrub_2122**
BLAST against opposite genome	Mrub_2122 Score: 82.4 E-value: 2e-18	b3456 Score: 77.8 E-value: 1e-16
CDD Data	LivM (COG4177)	
	E-value: 5.12e-53	E-value: 6.05e-78
Cell Localization	Embedded in the Cell Membrane	
TIGRfam- Protein Family	urea_t_UrtC_arc: urea ABC transporter, perm (TIGR03727)	
	E-value: 1e-11	E-value: 2.1e-09
Pfam- Protein Family	BPD_transp_2 (PF02653)	
	E-value: 9e-60	E-value: 9.5e-42
Protein Database	No hits	
KEGG Pathway Map	ABC Transporters KEGG Number: 02010	

Table 3 is a summary of all of the bioinformatics data collected in order to compare the orthologs *E.coli* b3456 and Mrub_2122. The first row of data is the comparison of results after running a BLAST against each gene. The *E.coli* sequence BLAST run against *M.ruber* had a hit on the gene Mrub_2122 with a fairly low e-value and the BLAST of the Mrub_2122 sequence against *E.coli* had a hit on the *E.coli* gene b3456. The next row shows the data collected from the CDD when the BLAST was run. The CDD found the same name and COG number when each of the gene sequences was run through the database. The COG number and name were found to be COG4177 and LivM, respectively. Both genes also had very small e-values, showing the

significance of the results. Bioinformatics tools such as SignalP, TMHMM, LipoP and PSORT-B suggested that both genes are embedded in the cell membrane. The TIGRfam database also pulled up the same TIGRfam name and number for both genes of TIGR03727 and urea ABC transporter permease. Furthermore, the Pfam database also pulled up the same Pfam name and number for both genes PF02653 and BPD_transp_2. Both e-values were low indicative of significance. Finally, both genes are involved in the same ABC transporter pathway (KEGG 02010).

Panel A



Panel B

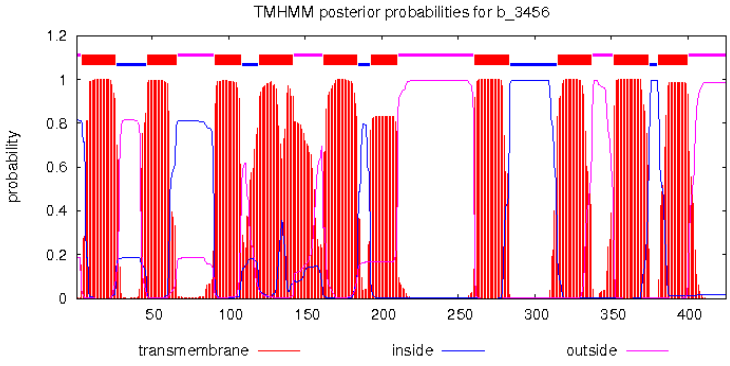
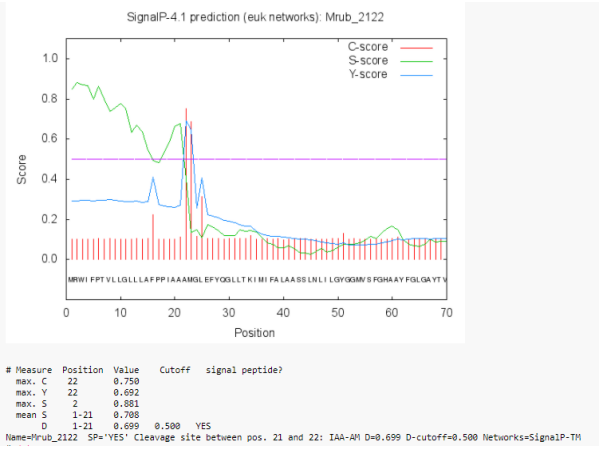


Figure 16. Mrub_2122 and *E.coli* b3456 both contain similar THM regions. Panel A is the TMHMM for Mrub_2122. Panel B is the TMHMM for *E.coli* b3456. The predicted location for the two are that they are embedded in the cell membrane.

The plots shown in figure 16 are the hydropathy plots for both Mrub_2122 and *E.coli* b3456. The red peaks on each plot are indicative of transmembrane helices. Mrub_2122 contains eight transmembrane helices while *E.coli* b3456 contains ten. The structures are slightly different but are believed to be functionally similar. Both plots are indicative of the protein coded by these genes being found in the cell membrane. Specifically, these two genes are believed to make up the second TMD of the branched-chain amino acid ABC transporter.

Panel A



Panel B

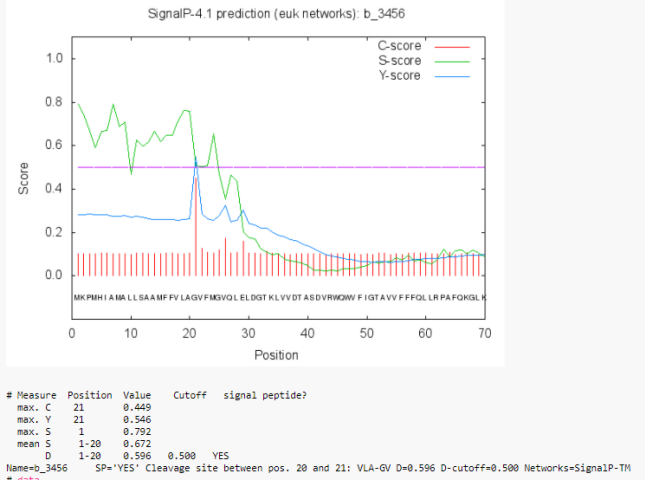
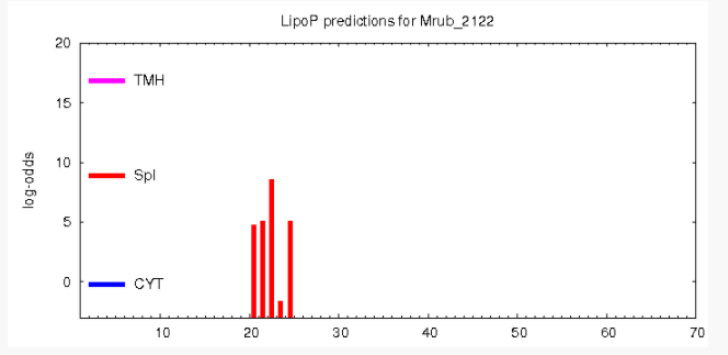


Figure 17. Both the Mrub_2122 and *E.coli* b3456 contain a signal peptide and a cleavage site. Panel A shows the plot for Mrub_2122. Panel B shows the plot for *E.coli* b3456.

SignalP graphs are displayed in figure 17 for both Mrub_2122 and *E.coli* b3456. The cleavage sight for both was predicted. Mrub_2122 had a cleavage sight after amino acid 20 while *E.coli* b3456 had a cleavage sight after amino acid 21. The gathered information tells us that both genes of interest contain a signal peptide further confirming the localization of both Mrub_2122 and *E.coli* b3456 to the cell membrane.

Panel A



Panel B

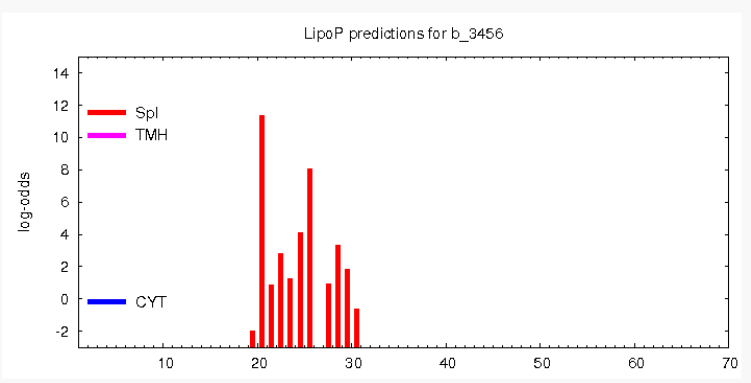
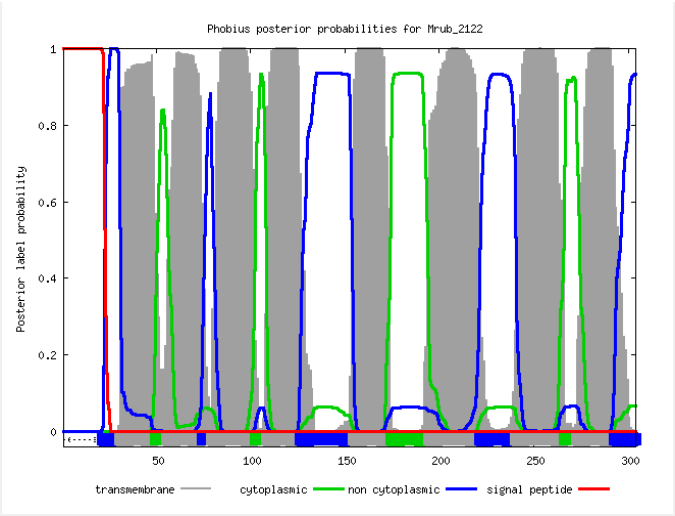


Figure 18. The given plots show the amino acid where the cleavage site is located, thus confirming the presence of a signal peptide. Panel A show the LipoP prediction for Mrub_2122. Panel B shows the LipoP prediction for *E.coli* b3456.

LipoP graphs are displayed in figure 18 for both Mrub_2122 and *E.coli* b3456. Mrub_2122 was found to have a cleavage sight after amino acid 20 while *E.coli* b3456 has one after amino acid 21. The graph for Mrub_2122 also further confirmed the presence of

transmembrane helices while the graph for *E.coli* b3456 confirms the lack of transmembrane helices.

Panel A



Panel B

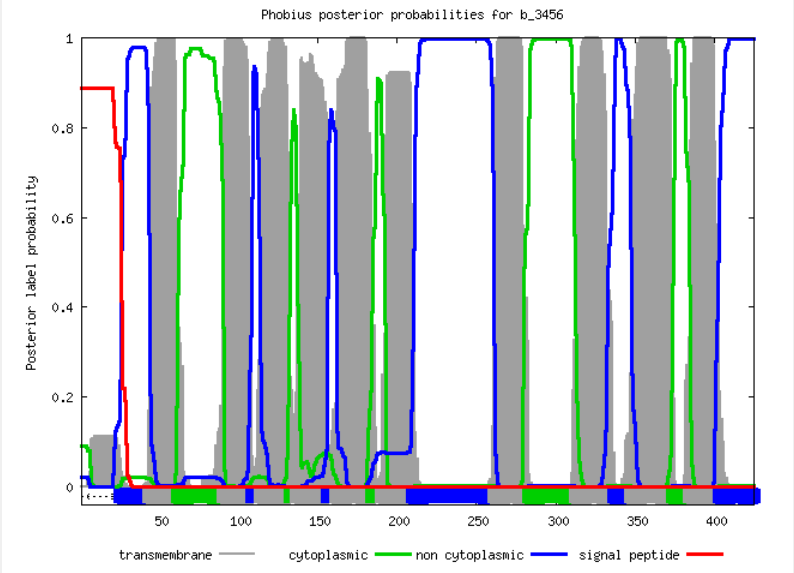
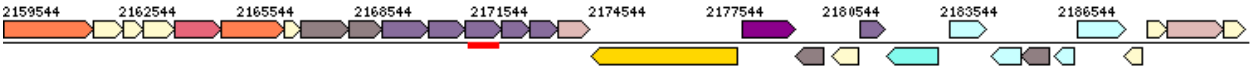


Figure 19. The given graphs display the possible locations of the given genes of interest. Panel A show the Phobius graph for Mrub_2122. Panel B shows the Phonius graph for *E.coli* b3456.

The Phobius graphs displayed in figure 19 for both Mrub_2122 and *E.coli* b3456 are very similar. Both Mrub_2122 and *E.coli* b3456 were found to fluctuate from the cytoplasm and non-

cytoplasm as any transporter would however, it is not localized in the cytoplasm. Both Phobius graphs also further confirm the presence of signal peptides in both Mrub_2122 and *E.coli* b3456.

Panel A



Panel B

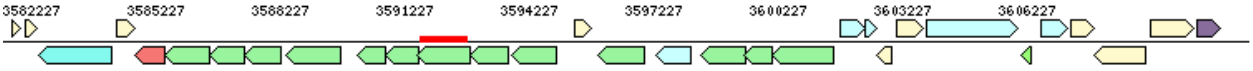


Figure 20. Mrub_2122 and *E.coli* b3456 are both a part of an operon. Panel A is the Mrub_2122 gene sequence. Panel B is the *E.coli* b3456 gene sequence.

The gene sequences in figure 20 shows that both Mrub_2122 and *E.coli* b_456 are both a part of an operon. The similar colors upstream and downstream are indicative of the same function, indicating an operon is present. Mrub_2122 is in a light purple shade indicative of membrane transport while E.coli b3456 is a light green indicative of its involvement in amino acid metabolism. The evidence shown here supports the notion that these two genes are orthologs.

Panel A

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
BPD_transp_2	Branched-chain amino acid transport syst ...	Family	CL0142	27	292	30	289

```
#HMM      nlltlaailaiaAlGlalvfgiaGvinlghggfmalGayvaalllalllslllallvallvgaavgllllglllvlrlkvdeliitllllsvlvgltlllltgitsgvkgqssvkellgfagallsflsavivllalllvla
#MATCH     llt+++i+a+aA +l+l+ g++G++++gh++ ++lGay++a+l+  +  + +++ +va++++a+++ll+g+++lr++  ++l+ +l++  ++++++ + +++ +g+ g +   + +g +    ++l   ++l +ll++l
#PP       689*****************************************************************************999998888888887.67777777777766666655555555553.345888888888
#SEQ       GLLTKIMIFALAASSLNLILGYGGMVSFGHAAYFGLGAYTVAILMREGVTSGWVIWAVAVGLSALLALLIGAISLRTRGIYFIMITLAFAQMIYYLV-VSFKQYGGEDGLRARRPEFGLFELNDTTLY-YLTLGVLLAALV
```

Panel B

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
DUF3382	Domain of unknown function (DUF3382)	Family	n/a	4	103	6	103
BPD_transp_2	Branched-chain amino acid transport syst ...	Family	CL0142	110	398	112	398

```
#HMM      lnlltlaailaiaAlGlalvfgiaGvinlghggfmalGayvaalllalllslllallvallvgaavgllllglllvlrlkvdeliitllllsvlvglttllltgitsgvkgqssvkellgfagallsflsaviv.....................
#MATCH    + ++tl++i++i++lGl++v+g++G++ lg ggf+a+Gay++all ++++  ++++l++a l++aa+g+llg++vlrl+ d+l+i++l++ ++v+++ll++t+it+g++g+s++++++ f  +++ ++ +  +
#PP       6799*****************************************************************************************************************************************
#SEQ       VDIATLTMIYIILGLGLNVVVGLSGLLVLGYGGFYAIGAYTFALLNHYYGLGFWTCLPIAGLMAAAAGFLLGFPVLRLRGDYLAIVTLGFGEIVRILLLNNTEITGGPNGISQIPKPTLFGLEFSRTAREGGWdtfsnffglkydpsdrviflv
```

Figure 21. Mrub_2122 and b3456 contain most of the same highly conserved amino acids. Both genes code for the same domain BPD_transp_2. Panel A shows the pairwise alignment for Mrub_2122. Panel B shows the pairwise alignment for *E.coli* b3456.

The pairwise alignments shown above in figure 21 contain a variety of the same highly conserved amino acids. Both Mrub_2122 and *E.coli* b3456 contain a highly conserved alanine toward the beginning of their protein sequences. The fact that these two protein sequences share conserved residues is yet another reason the two are orthologous to each other.

Panel A

SSDB Paralog Search Result

KEGG ID : mrb:Mrub_0138 (358 a.a.)
Definition: inner-membrane translocator; K01998 branched-chain amino acid transport system permease protein
Update status: T01193 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculatic
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

Top10 Select operation Select

Search Result : 19 hits

Entry	KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
mrb:Mrub_2374 inner-membrane translocator	K01998	356	786	(-)	185	0.439	351	->
mrb:Mrub_1326 ABC transporter related protein	K01995..	590	505	(-)	121	0.335	334	->
mrb:Mrub_2652 inner-membrane translocator	K01998	477	481	(-)	115	0.314	370	->
mrb:Mrub_2539 ABC transporter related protein	K01995..	559	480	(-)	115	0.349	255	->
mrb:Mrub_1677 inner-membrane translocator	K01998	302	393	(-)	95	0.305	315	->
mrb:Mrub_2122 inner-membrane translocator	K01998	304	391	(-)	95	0.296	328	->
mrb:Mrub_2687 inner-membrane translocator	K01998	318	361	(-)	88	0.283	293	->
mrb:Mrub_2688 inner-membrane translocator	K01997	287	263	(-)	66	0.266	289	->
mrb:Mrub_2121 inner-membrane translocator	K01997	306	261	(-)	65	0.259	316	->
mrb:Mrub_2378 inner-membrane translocator	K01997	304	257	(-)	64	0.256	305	->

Panel B

SSDB Paralog Search Result

KEGG ID : eco:b3456 (425 a.a.)
Definition: branched-chain amino acid ABC transporter permease; K01998 branched-chain amino acid transport system permease protein
Update status: T00007 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculatic
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

All Select operation Select

Search Result : 6 hits

Entry	KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
eco:b3457 branched-chain amino acid ABC transporter per	K01997	308	272	(-)	68	0.266	312	->
eco:b1514 autoinducer 2 import system permease protein	K10556	342	222	(-)	56	0.264	296	->
eco:b4230 putative sugar ABC transporter permease	K02057	341	161	(-)	43	0.260	289	->
eco:b0657 apolipoprotein N-acyltransferase	K03820	512	107	(-)	30	0.304	115	->
eco:b2740 putative transporter	K03299	454	105	(-)	30	0.311	119	->
eco:b2975 glycolate transporter	K02550	560	105	(-)	30	0.331	124	->

[SSDB | GENES | KEGG2 | KEGG | GenomeNet]

Figure 22. Paralogs of Mrub_2122 and b3456. Panel A shows the results of a DB paralog search through the KEGG database for Mrub_2122. Panel B is the results of the same search using b3456.

The paralog data present in figure 22 depicts the known paralogs of the genes of interest. Mrub_2122 was found to have nineteen paralogs present while b3456 only had six. While the number of paralogs between these two genes is quite different, these two genes are still thought to be orthologous to each other given the other bioinformatics evidence.

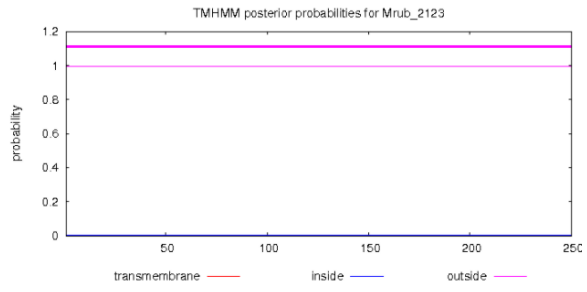
Table 4. *E.coli* b3455 and Mrub_2123 are orthologs

Bioinformatics tool used	**E. coli b3455 gene**	**Mrub_2123 gene**
BLAST against opposite genome	Mrub_2123 Score: 176 E-value: 8e-56	b3455 Score: 69.3 E-value: 2e-15
CDD Data	LivG (COG0411)	
	E-value: 1.61e-139	E-value: 1.43e-107
Cell localization	Cytoplasm	
TIGRfam - Protein family	urea_trans_UrtD: urea ABC transporter (TIGR03411)	
	E-value: 3.7e-65	E-value: 4.4e-71
Pfam - Protein family	ABC Transporter (PF0005)	
	E-value: 3.1e-32	E-value: 2.1e-12
Protein Database	No hits	
KEGG Pathway Map	ABC Transporters KEGG Number: 02010	

Table 4 is a summary of all of the bioinformatics data collected in order the orthologs *E.coli* b3455 and Mrub_2123. The first row of data is the comparison of results after running a BLAST against each gene. The *E.coli* sequence BLAST run against *M.ruber* had a hit on the gene Mrub_2123 with a fairly low e-value and the BLAST of the Mrub_2123 sequence against *E.coli* had a hit on the *E.coli* gene b3455. The next row shows the data collected from the CDD when the BLAST was run. The CDD found the same name and COG number when each of the gene sequences was run through the database. The COG number and name were found to be COG0411 and LivG, respectively. Both genes also had very small e-values, showing the significance of the results. Bioinformatics tools such as SignalP, TMHMM, LipoP and PSORT-B suggested that both *E.coli* b3455 and Mrub_2123 are localized in the cytoplasm. The TIGRfam database also pulled up the same TIGRfam name and number for both genes of TIGR03411 and urea ABC transporter. Furthermore, the Pfam database also pulled up the same Pfam name and number for both genes PF0005 and ABC transporter. Both e-values were low indicative of significance. Finally, both genes are involved in the same ABC transporter pathway (KEGG 02010).

Panel A

```
# Mrub_2123 Length: 250
# Mrub_2123 Number of predicted TMHs:  0
# Mrub_2123 Exp number of AAs in TMHs: 0.01809
# Mrub_2123 Exp number, first 60 AAs:  0.01285
# Mrub_2123 Total prob of N-in:        0.00334
Mrub_2123       TMHMM2.0        outside     1   250
```



Panel B

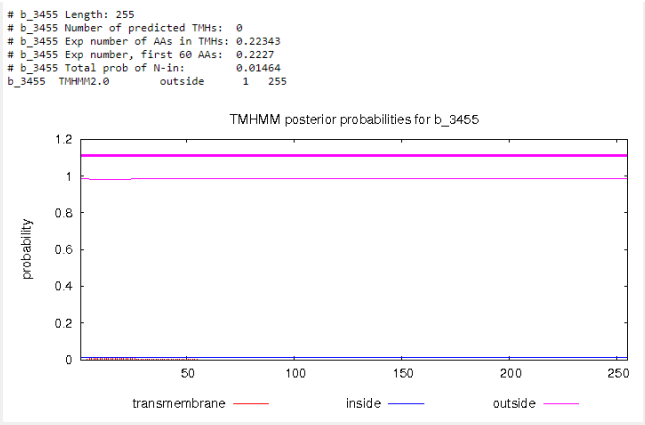
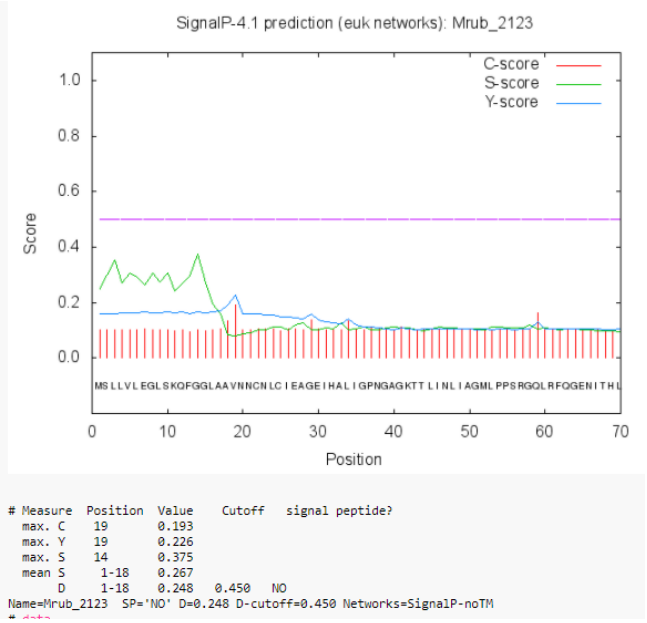


Figure 23. Mrub_2123 and *E.coli* b3455 both contain similar TMH regions. Panel A is the TMHMM for Mrub_2123. Panel B is the TMHMM for *E.coli* b3455. The predicted location for the two are located in the cytoplasm.

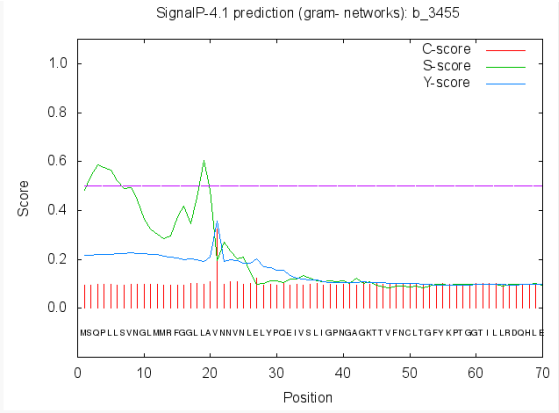
The plots in figure 23 are hydropathy plots for both Mrub_2123 and *E.col*i b3455. The red peaks on each plot are indicative of transmembrane helices. Both Mrub_2123 and *E.coli* b3455 contain no transmembrane helices. Because both plots have zero TMHs it is indicative of the protein coded by these genes being found in the cell membrane because it does not move in and out of the cell. Specifically, it is believed that both of these genes are anchored to the intracellular side of the inner plasma membrane. They are believed to encode one of the NBD's of the ABC transporter.

Panel A




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    19        0.193
  max. Y    19        0.226
  max. S    14        0.375
  mean S     1-18     0.267
       D     1-18     0.248   0.450   NO
Name=Mrub_2123  SP='NO' D=0.248 D-cutoff=0.450 Networks=SignalP-noTM
```

Panel B



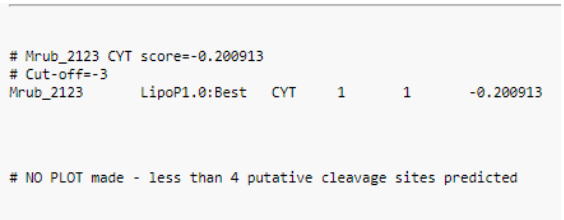

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    21        0.325
  max. Y    21        0.357
  max. S    19        0.606
  mean S     1-20     0.448
       D     1-20     0.400   0.570   NO
Name=b_3455     SP='NO' D=0.400 D-cutoff=0.570 Networks=SignalP-noTM
```

Figure 24. Both the Mrub_2123 and *E.coli* b3455 both do not contain a cleavage site and therfore also do not contain a signal peptide. Panel A shows the plot for Mrub_2123. Panel B shows the plot for *E.coli* b3455.

SignalP graphs are displayed in figure 24 for both Mrub_2123 and *E.coli* b3455. There was no cleavage sight predicted for Mrub_2123 or *E.coli* b3455 had a cleavage. The gathered

information tells us that both genes of interest do not contain a signal peptide further confirming the localization of both Mrub_2123 and *E.coli* b3455 to the cytoplasm.

Panel A



Panel B

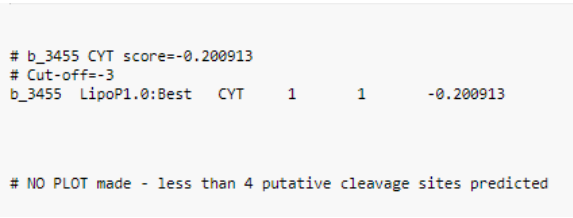
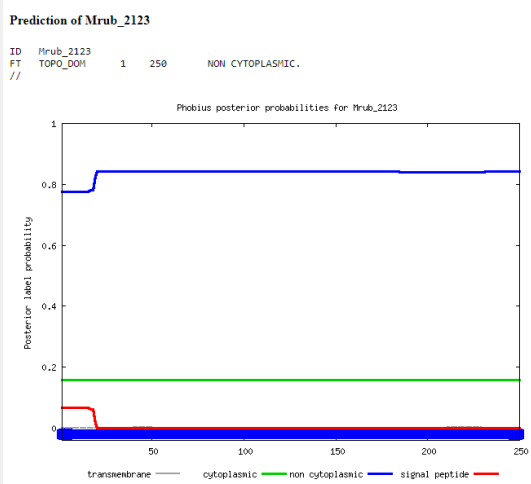


Figure 25. The images produced from LipoP display that there are less than four putative cleavage sites. Panel A show the LipoP prediction for Mrub_2123. Panel B shows the LipoP prediction for *E.coli* b3455.

LipoP graphs are displayed in figure 25 for both Mrub_2123 and *E.coli* b3455. Neither Mrub_2123 nor *E.coli* b3455 were shown to have signal peptides. The lack of a signal peptide was indicated by the images produced by LipoP showing less than four putative cleavage sites for both genes of interest.

Panel A



Panel B

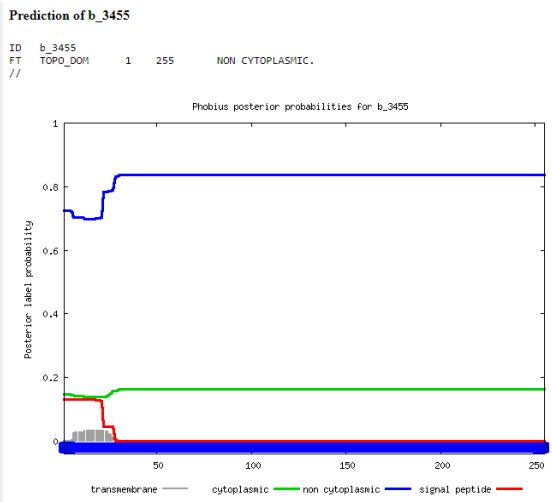
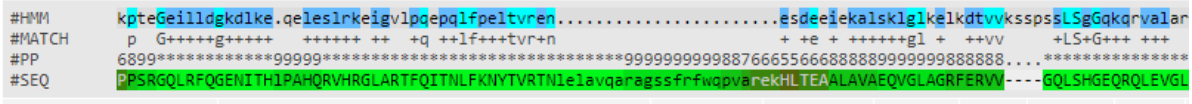


Figure 26. The given graphs display the possible locations of the given genes of interest. Panel A show the Phobius graph for Mrub_2123. Panel B shows the Phonius graph for *E.coli* b3455.

The Phobius graphs displayed in figure 26 for both Mrub_2123 and *E.coli* b3455 are very similar. Both Mrub_2123 and *E.coli* b3455 were found to fluctuate from the cytoplasm and non-cytoplasm as any transporter would, however the graphs show us both are localized in the cytoplasm. Both Phobius graphs also further confirm the lack of signal peptides in both Mrub_2123 and *E.coli* b3455.

Panel A

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
ABC_tran	ABC transporter	Domain	CL0023	1	124	3	124



Panel B

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
ABC_tran	ABC transporter	Domain	CL0023	21	182	21	181

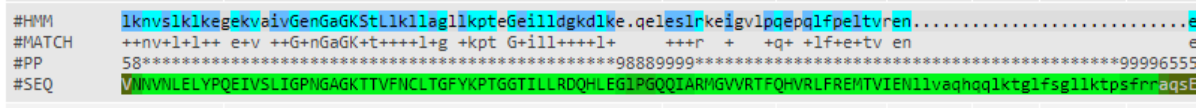
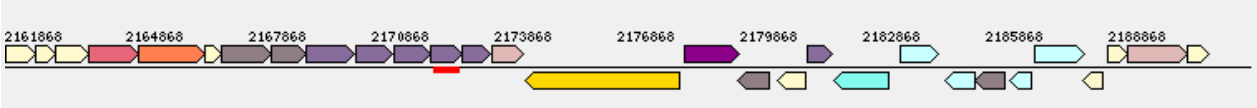


Figure 27. Mrub_2123 and b3455 contain most of the same highly conserved amino acids. Both genes code for the same domain ABC transporter. Panel A shows the pairwise alignment for Mrub_2123. Panel B shows the pairwise alignment for *E.coli* b3455.

The pairwise alignments shown above in figure 27 contain a variety of the same highly conserved amino acids. Both Mrub_2123 and *E.coli* b3457 contain a highly conserved glycine toward the beginning of where the protein sequences align. The fact that these two protein sequences share a conserved residue is yet another reason the two are orthologous to each other.

Panel A



Panel B

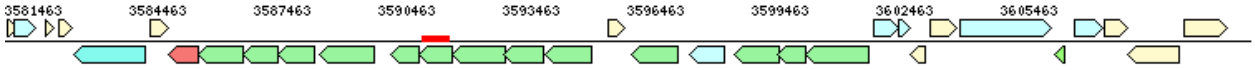
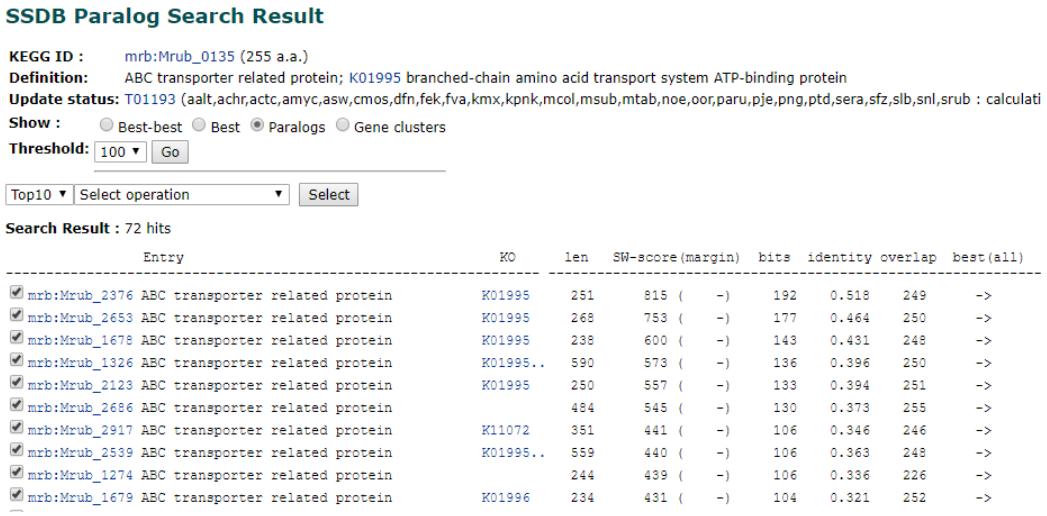


Figure 28. The Mrub_2123 and *E.coli* b3455 genes are a part of an operon. Panel A is the Mrub_2123 gene sequence. Panel B is the *E.coli* b3455 gene sequence.

The two genome sequences in figure 28 show that both Mrub_2123 and *E.coli* b3455 are both a part of an operon. The similar colors upstream and downstream are indicative of the same

function, indicating an operon is present. Mrub_2123 is in a light purple shade indicative of membrane transport while *E.coli* b3455 is a light green indicative of its involvement in amino acid metabolism. The evidence shown here adds to the confirmation that these two genes are orthologs.

Panel A



SSDB Paralog Search Result

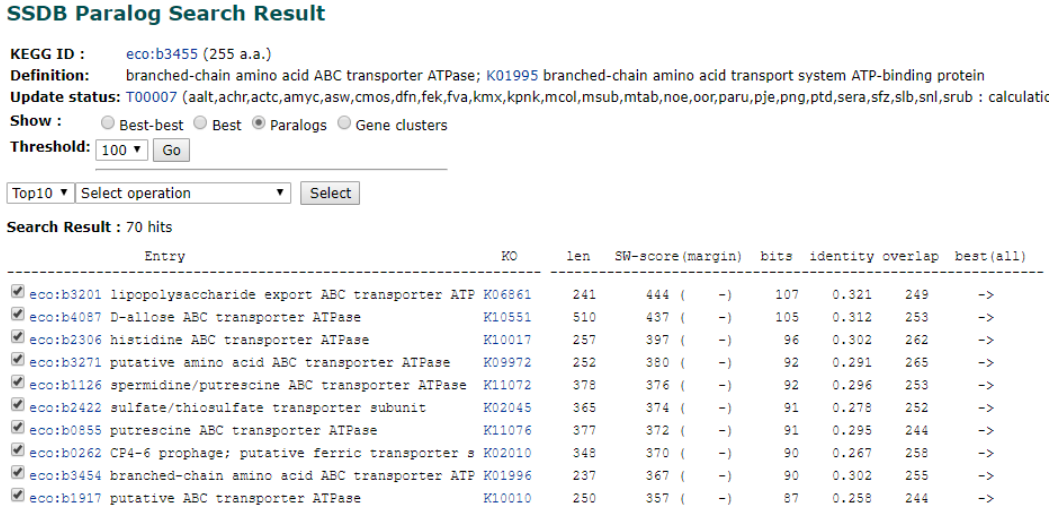
KEGG ID : mrb:Mrub_0135 (255 a.a.)
Definition: ABC transporter related protein; K01995 branched-chain amino acid transport system ATP-binding protein
Update status: T01193 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculati
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

Top10 Select operation Select

Search Result : 72 hits

Entry	KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
mrb:Mrub_2376 ABC transporter related protein	K01995	251	815	(-)	192	0.518	249	->
mrb:Mrub_2653 ABC transporter related protein	K01995	268	753	(-)	177	0.464	250	->
mrb:Mrub_1678 ABC transporter related protein	K01995	238	600	(-)	143	0.431	248	->
mrb:Mrub_1326 ABC transporter related protein	K01995..	590	573	(-)	136	0.396	250	->
mrb:Mrub_2123 ABC transporter related protein	K01995	250	557	(-)	133	0.394	251	->
mrb:Mrub_2686 ABC transporter related protein		484	545	(-)	130	0.373	255	->
mrb:Mrub_2917 ABC transporter related protein	K11072	351	441	(-)	106	0.346	246	->
mrb:Mrub_2539 ABC transporter related protein	K01995..	559	440	(-)	106	0.363	248	->
mrb:Mrub_1274 ABC transporter related protein		244	439	(-)	106	0.336	226	->
mrb:Mrub_1679 ABC transporter related protein	K01996	234	431	(-)	104	0.321	252	->

Panel B



SSDB Paralog Search Result

KEGG ID : eco:b3455 (255 a.a.)
Definition: branched-chain amino acid ABC transporter ATPase; K01995 branched-chain amino acid transport system ATP-binding protein
Update status: T00007 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculatic
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

Top10 Select operation Select

Search Result : 70 hits

Entry	KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
eco:b3201 lipopolysaccharide export ABC transporter ATP	K06861	241	444	(-)	107	0.321	249	->
eco:b4087 D-allose ABC transporter ATPase	K10551	510	437	(-)	105	0.312	253	->
eco:b2306 histidine ABC transporter ATPase	K10017	257	397	(-)	96	0.302	262	->
eco:b3271 putative amino acid ABC transporter ATPase	K09972	252	380	(-)	92	0.291	265	->
eco:b1126 spermidine/putrescine ABC transporter ATPase	K11072	378	376	(-)	92	0.296	253	->
eco:b2422 sulfate/thiosulfate transporter subunit	K02045	365	374	(-)	91	0.278	252	->
eco:b0855 putrescine ABC transporter ATPase	K11076	377	372	(-)	91	0.295	244	->
eco:b0262 CP4-6 prophage; putative ferric transporter s	K02010	348	370	(-)	90	0.267	258	->
eco:b3454 branched-chain amino acid ABC transporter ATP	K01996	237	367	(-)	90	0.302	255	->
eco:b1917 putative ABC transporter ATPase	K10010	250	357	(-)	87	0.258	244	->

Figure 29. Paralogs of Mrub_2123 and b3455. Panel A shows the results of a DB paralog search through the KEGG database for Mrub_2123. Panel B is the results of the same search using b3455.

The paralog data present in figure 29 depicts the known paralogs of the genes of interest. Mrub_2123 was found to have seventy-two paralogs present while b3455 had seventy. The relatively close number of paralogs between these two genes is further evidence indicating that the two are orthologous to each other.

Table 5. *E.coli* b3454 and Mrub_2124 are orthologs

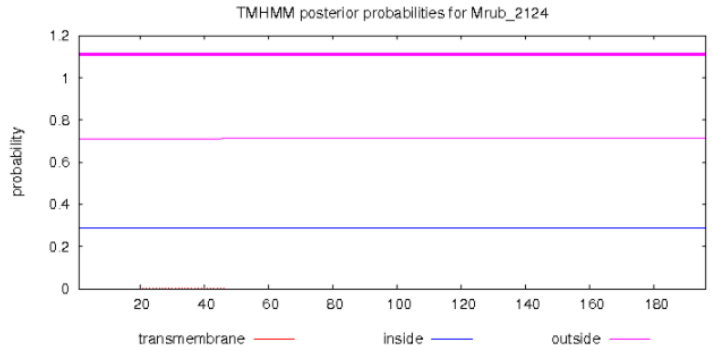
Bioinformatics tool used	**E. coli b_3454 gene**	**Mrub_2124 gene**
BLAST against opposite genome	Mrub_2124 Score: 173 E-value: 4e-55	b3454 Score: 143 E-value: 1e-43
CDD Data	LivF (COG0410)	
	E-value: 5.11e-136	E-value: 7.80e-84
Cell localization	Anchored to cytoplasmic membrane	
TIGRfam - Protein family	urea_trans_UrtE: urea ABC transporter (TIGR03410)	
	E-value: 4.7e-60	E-value: 2.5e-32
Pfam - Protein family	ABC Transporter (PF0005)	
	E-value: 3.3e-33	E-value: 7.4e-20
Protein Database	No hits	
KEGG Pathway Map	ABC Transporters KEGG Number: 02010	

Table 5 is a summary of all of the bioinformatics data collected in order to compare the orthologs *E.coli* b3454 and Mrub_2124. The first row of data is the comparison of results after running a BLAST against each gene. The *E.coli* sequence BLAST run against M.ruber had a hit

on the gene Mrub_2124 with a fairly low e-value and the BLAST of the Mrub_2124 sequence against *E.coli* had a hit on the *E.coli* gene b3454. The next row shows the data collected from the CDD when the BLAST was run. The CDD found the same name and COG number when each of the gene sequences was run through the database. The COG number and name were found to be COG0410 and LivF, respectively. Both genes also had very small e-values, showing the significance of the results. Bioinformatics tools such as SignalP, THM, LipoP and PSORT-B suggested that both genes are anchored to the cytoplasmic membrane. The TIGRfam database also pulled up the same TIGRfam name and number for both genes of TIGR03410 and urea ABC transporter. Furthermore, the pfam database also pulled up the same pfam name and number for both genes PF0005 and ABC transporter. Both e-values were low indicative of significance. Finally, both genes are involved in the same ABC transporter pathway.

Panel A

```
# Mrub_2124 Length: 196
# Mrub_2124 Number of predicted TMHs:  0
# Mrub_2124 Exp number of AAs in TMHs: 0.04111
# Mrub_2124 Exp number, first 60 AAs:  0.03914
# Mrub_2124 Total prob of N-in:        0.28920
Mrub_2124       TMHMM2.0        outside      1   196
```



Panel B

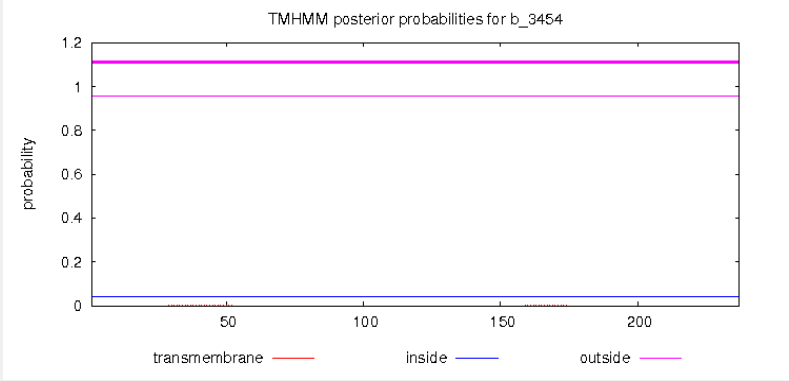
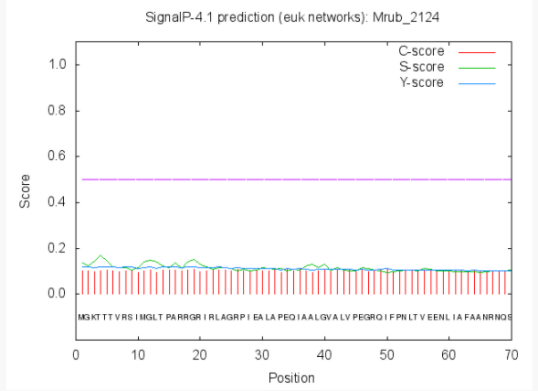


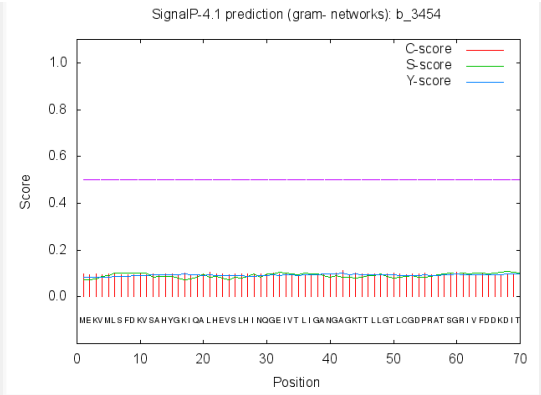
Figure 30. Mrub_2124 and *E.coli* b3454 both contain similar THM regions. Panel A is the TMHMM for Mrub_2124. Panel B is the TMHMM for *E.coli* b3454. The predicted location for the two are anchored to the cytoplasmic membrane.

The plots in figure 30 display the hydropathy plots for both Mrub_2124 and *E.coli* b3454. Red peaks shown on each plot are indicative of a transmembrane helices being present. Both Mrub_2124 and *E.coli* b3454 contain no transmembrane helices which is indicative of proteins not embedded in the membrane. Specifically, it is believed that both of these genes are anchored to the intracellular side of the inner plasma membrane. They are believed to encode the second NBD of the branched-chain amino acid ABC transporter.

Panel A



Panel B




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    42        0.111
  max. Y    42        0.103
  max. S    68        0.108
  mean S     1-41     0.090
       D     1-41     0.097   0.570   NO
Name=b_3454     SP='NO' D=0.097 D-cutoff=0.570 Networks=SignalP-noTM
```

Figure 31. Both the Mrub_2124 and *E.coli* b3454 both do not contain a cleavage site and therefore also do not contain a signal peptide. Panel A shows the plot for Mrub_2124. Panel B shows the plot for *E.coli* b3454.

SignalP graphs are displayed in figure 31 for both Mrub_2124 and *E.coli* b3454. There was no cleavage sight predicted for Mrub_2124 or *E.coli* b3454 had a cleavage. The gathered information tells us that both genes of interest do not contain a signal peptide further confirming the localization of both Mrub_2124 and *E.coli* b3454 to the cytoplasmic membrane.

Panel A

```
# Sequence CYT score=-0.200913
# Cut-off=-3
Sequence        LipoP1.0:Best   CYT     1       1       -0.200913


# NO PLOT made - less than 4 putative cleavage sites predicted
```

Panel B

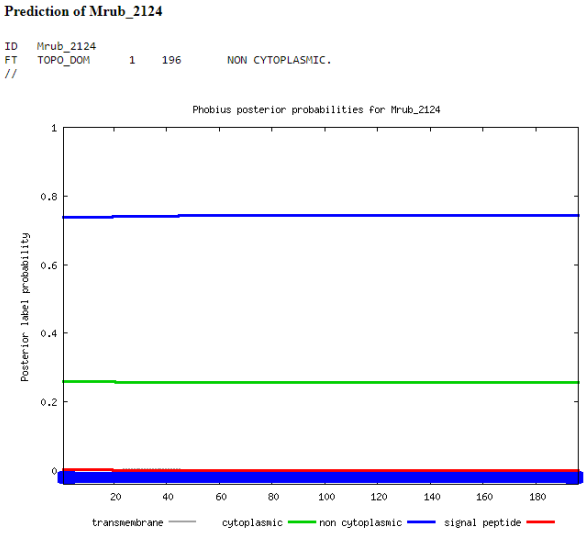
```
# b_3454 CYT score=-0.200913
# Cut-off=-3
b_3454  LipoP1.0:Best   CYT     1       1       -0.200913


# NO PLOT made - less than 4 putative cleavage sites predicted
```

Figure 32. The images produced from LipoP display that there are less than four putative cleavage sites. Panel A show the LipoP prediction for Mrub_2124. Panel B shows the LipoP prediction for *E.coli* b3454.

LipoP graphs are displayed in figure 32 for both Mrub_2124 and *E.coli* b3454. Neither Mrub_2124 nor *E.coli* b3454 were shown to have signal peptides. The lack of a signal peptide was indicated by the images produced by LipoP showing less than four putative cleavage sites for both genes of interest.

Panel A



Panel B

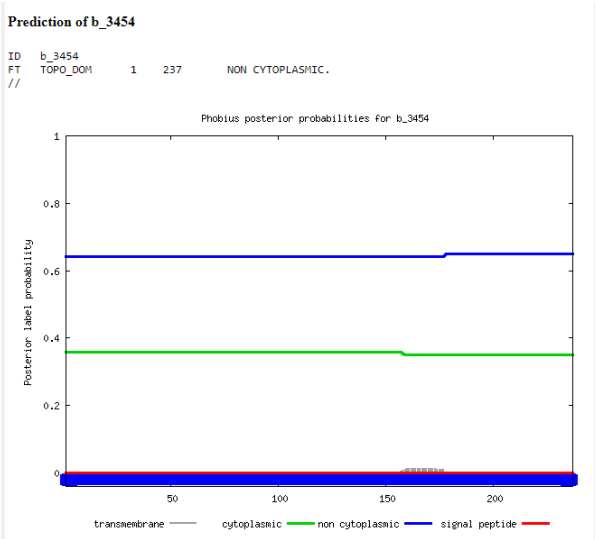


Figure 33. The given graphs display the possible locations of the given genes of interest. Panel A show the Phobius graph for Mrub_2124. Panel B shows the Phonius graph for *E.coli* b3454.

The Phobius graphs displayed in figure 33 for both Mrub_2124 and *E.coli* b3454 are very similar. Both Mrub_2124 and *E.coli* b3454 were found to fluctuate from the cytoplasm and non-cytoplasm as any transporter would, however the graphs show us both are localized in the cytoplasmic membrane. Both Phobius graphs also further confirm the lack of signal peptides in both Mrub_2124 and *E.coli* b3454.

Panel A

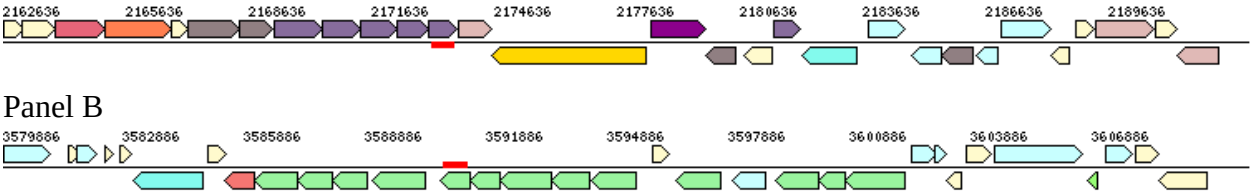


Figure 34. The Mrub_2124 and *E.coli* b3454 genes are a part of an operon. Panel A is the Mrub_2124 chromosome viewer. Panel B is the *E.coli* b3454 chromosome viewer. Chromosome viewer maps were colored by KEGG.

The panels shown in figure 34 display that both Mrub_2124 and *E.coli* b3454 are both a part of an operon. The similar colors upstream and downstream are indicative of the same function, indicating an operon is present. Mrub_2124 is in a light purple shade indicative of membrane transport while *E.coli* b3454 is a light green indicative of its involvement in amino acid metabolism. The evidence shown here supports the notion that these two genes are orthologs.

Panel A

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
ABC_tran	ABC transporter	Domain	CL0023	1	125	2	124

```
#HMM    GKStLlkllagllkpteGeilldgkdlke.qeleslrkeigvlpqepqlfpeltvren........esdeeiekalsklglkelkdtvvkssplSgGqkqrva
#MATCH  GK+t ++ ++gl+    G+i+l g++++   +++++   ++ +p+  q+fp+ltv+en        +++ e+ ++++   l    +++ ++ +++LSgG++q++a
#PP     ****************************9999*****************************9876444444444444444444444445668***********
#SEQ    GKTTTVRSIMGLTPARRGRIRLAGRPIEAlAPEQIAALGVALVPEGRQIFPNLTVEENliafaanrNQSPEPWTLARIWTLFPRLKERRNNLGYQLSGGEQQMLA
```

Panel B

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length
				Start	End	Start	End	From	To	
ABC_tran	ABC transporter	Domain	CL0023	21	166	21	165	1	**136**	137

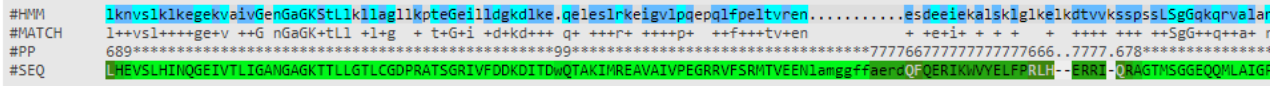


Figure 35. Mrub_2124 and b3454 contain the same highly conserved amino acids at the beginning of their alignment, both genes also code for the same domain ABC transporter. Panel A shows the pairwise alignment for Mrub_2124. Panel B shows the pairwise alignment for *E.coli* b3454.

The pairwise alignments shown above in figure 35 contain several of the same highly conserved amino acids. They both contain a highly conserved serine and glycine at the beginning of their alignments. Both Mrub_2124 and *E.coli* b3454 have the highly conserved amino acids, further validating that the two are orthologs.

Panel A

SSDB Paralog Search Result

KEGG ID : mrb:Mrub_0140 (272 a.a.)
Definition: ABC transporter related protein; K01996 branched-chain amino acid transport system ATP-binding protein
Update status: T01193 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculatic
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

Top10 Select operation Select

Search Result : 73 hits

Entry		KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
mrb:Mrub_2654	ABC transporter related protein	K01996	237	756	(-)	178	0.512	240	->
mrb:Mrub_2377	ABC transporter related protein	K01996	261	753	(-)	177	0.454	262	->
mrb:Mrub_1325	ABC transporter related protein	K01996	243	740	(-)	175	0.502	243	->
mrb:Mrub_1679	ABC transporter related protein	K01996	234	663	(-)	157	0.483	240	->
mrb:Mrub_2686	ABC transporter related protein		484	583	(-)	139	0.463	227	->
mrb:Mrub_2124	ABC transporter related protein	K01996	236	575	(-)	137	0.444	241	->
mrb:Mrub_2540	ABC transporter related protein	K01996	229	552	(-)	132	0.432	241	->
mrb:Mrub_2376	ABC transporter related protein	K01995	251	426	(-)	103	0.325	255	->
mrb:Mrub_2580	ABC transporter related protein	K06861	256	423	(-)	102	0.332	229	->
mrb:Mrub_1678	ABC transporter related protein	K01995	238	400	(-)	97	0.367	226	->

Panel B

SSDB Paralog Search Result

KEGG ID : eco:b3454 (237 a.a.)
Definition: branched-chain amino acid ABC transporter ATPase; K01996 branched-chain amino acid transport system ATP-binding protein
Update status: T00007 (aalt,achr,actc,amyc,asw,cmos,dfn,fek,fva,kmx,kpnk,mcol,msub,mtab,noe,oor,paru,pje,png,ptd,sera,sfz,slb,snl,srub : calculatic
Show : Best-best Best Paralogs Gene clusters
Threshold: 100 Go

Top10 Select operation Select

Search Result : 73 hits

Entry	KO	len	SW-score	(margin)	bits	identity	overlap	best(all)
eco:b3201 lipopolysaccharide export ABC transporter ATP	K06861	241	440	(-)	106	0.345	229	->
eco:b0809 glutamine transporter subunit	K10038	240	423	(-)	102	0.319	226	->
eco:b0652 glutamate/aspartate ABC transporter ATPase	K10004	241	374	(-)	91	0.319	226	->
eco:b3455 branched-chain amino acid ABC transporter ATP	K01995	255	367	(-)	90	0.302	255	->
eco:b0864 arginine ABC transporter ATPase	K10000	242	357	(-)	87	0.329	237	->
eco:b3271 putative amino acid ABC transporter ATPase	K09972	252	355	(-)	87	0.279	226	->
eco:b3463 cell division ATP-binding protein	K09812	222	350	(-)	86	0.306	219	->
eco:b1917 putative ABC transporter ATPase	K10010	250	344	(-)	84	0.314	226	->
eco:b0199 DL-methionine transporter subunit	K02071	343	342	(-)	84	0.298	225	->
eco:b3725 phosphate ABC transporter ATPase	K02036	257	338	(-)	83	0.310	239	->

Figure 36. Paralogs of Mrub_2124 and b3454. Panel A shows the results of a DB paralog search through the KEGG database for Mrub_2124. Panel B is the results of the same search using b3454.

The paralog data present in figure 36 depicts the known paralogs of the genes of interest. Mrub_2124 and b3454 both had seventy three hits. Since these two genes had an identical number of paralogs, it is strong evidence that the two are orthologous to each other.

Conclusion:

Site-directed mutagenesis is a method that is used to make specific and intentional changes to the DNA sequence of a gene and any gene products. In the case of our research, we looked at the orthologs Mrub_2122 and *E.coli* b3456 and mutated their sequences using NEBaseChanger (Biolabs; Betts and Russell, 2003).

Panel A

```
        10         20         30         40         50         60
MRWIFPTVLL GLLLAFPPIA AAMGLEFYQG LLTKIMIFAL AASSLNLILG YGGMVSFGHA

        70         80         90        100        110        120
AYFGLGAYTV AILMREGVTS GNVIWAVAVG LSALLALLIG AISLRTRGIY FIMITLAFAQ

       130        140        150        160        170        180
MIYYLVVSFK QYGGEDGLRA RRPEFGLFEL NDTTLYYLTL GVLLAALYLL YRLVHSRFGR

       190        200        210        220        230        240
VLQAIRENEA RALALGYPVF HFQLVAFVLA GALAGLAGVL MAHYTQYASP NLLAWQQSGH

       250        260        270        280        290        300
LMMMVILGGV GQFWGGVLGA LVLSLVEEIL QDLTIHWQLG VGLILLFIVL FAPKGLAGLM

RRGS
```

Panel B

```
        10         20         30         40         50         60
MKPMMHIAMAL LSAAMFFVLA GVFMGVQLEL DGTKLVVDTA SDVRWQWVFI GTAVVFFFQL

        70         80         90        100        110        120
LRPAFQKGLK SVSGPKFILP AIDGSTVKQK LFLVALLVLA VAWPFMVSRG TVDIATLTMI

       130        140        150        160        170        180
YIILGLGLNV VVGLSGLLVL GYGGFYAIGA YTFALLNHYY GLGFWTCLPI AGLMAAAAGF

       190        200        210        220        230        240
LLGFPVLRLR GDYLAIVTLG FGEIVRILLL NNTEITGGPN GISQIPKPTL FGLEFSRTAR

       250        260        270        280        290        300
EGGWDTFSNF FGLKYDPSDR VIFLYLVALL LVVLSLFVIN RLLRMPLGRA WEALREDEIA

       310        320        330        340        350        360
CRSLGLSPRR IKLTAFTISA AFAGFAGTLF AARQGFVSPE SFTFAESAFV LAIVVLGGMG

       370        380        390        400        410        420
SQFAVILAAI LLVVSRELMR DFNEYSMLML GGLMVLMMIW RPQGLLPMTR PQLKLKNGAA

KGEQA
```

Panel C

NEBaseChanger Summary - Thu Feb 08 2018

Input sequence:Mrub_2122

Type of mutagenesis: deletion

Mutagenesis region: 148 to 150

Result

```
   A  A  S  T  *  F  C  T  A  A  W  *
  S  S  L  N  L  I  L  Y  G  G  M  V  S
 Q  Q  P  Q  P  D  S  V  R  R  H  G  E
CAGCAGCCTCAACCTGATTCTGTACGGCGGCATGGTGAGC
GTCGTCGGAGTTGGACTAAGACATGCCGCCGTACCACTCG
```

Required Primers

Name (F/R)	Oligo (Uppercase = target-specific primer)	Len	% GC	Tm	Ta *
Q5SDM_2/9/2018_F	TACGGCGGCATGGTGAGC	18	67	72°C	69°C
Q5SDM_2/9/2018_R	CAGAATCAGGTTGAGGCTGCTG	22	55	68°C	

* Ta (recommended annealing temperature)

Panel D

NEBaseChanger Summary - Thu Feb 08 2018

Input sequence:b_3456

Type of mutagenesis: deletion

Mutagenesis region: 151 to 153

Result

```
   W  Q  W  V  F  I  T  A  V  V  F  F
  L  A  V  G  V  Y  H  G  G  G  L  F
 V  G  S  G  C  L  S  R  R  W  S  F
GTTGGCAGTGGGTGTTTATCACGGCGGTGGTCTTTTTC
CAACCGTCACCCACAAATAGTGCCGCCACCAGAAAAAG
```

Required Primers

Name (F/R)	Oligo (Uppercase = target-specific primer)	Len	% GC	Tm	Ta *
Q5SDM_2/9/2018_F	ACGGCGGTGGTCTTTTTC	18	56	65°C	65°C
Q5SDM_2/9/2018_R	GATAAACACCCACTGCCAAC	20	50	64°C	

* Ta (recommended annealing temperature)

Figure 37. Site-directed mutagenesis of Mrub_2122 and b3456. Panels A and B show the amino acid sequences of Mrub_2122 and b3456, respectively. A 3-nucleotide deletion was made which codes for a conserved glycine at positions 50 and 51 for Mrub_2122 and b3456, respectively. Panel C shows the results of the *M. ruber* mutation. Panel D shows the results of the *E. coli* mutation.

Figure 37 represents the intended mutations against the Mrub_2122 and the orthologous b3456 in *E. coli*. Functional genomics will be performed using successfully mutated bacteria to see how the mutations mentioned will affect branched-chain amino acid transport in *M. ruber* and *E. coli*.

The *E.coli* b3458 and Mrub_2120 share all the same outputs produced from the CDD, PFAM, TIGRfam and PDB databases. Both genes had the same CDD hit of LivK (COG0683) with acceptable e-values of 2.40e-102 and 8.79e-59, respectively, indicating the significance of the hits. Both genes also had the same PFAM hit of Periplasmic Binding Protein (PF13458) with e-values of 1.7e-61 and 1.6e-71, respectively, indicating the significance of the hits. There was also an identical hit for both genes in the TIGRfam database of urea_ABC_UrtA: urea ABC transporter (TIGR03407) with e-values of 4.2e-4 and 7.8e-8, respectively, indicating the significance of the hits. The cell localization data suggests that these genes are localized to the periplasmic space and likely code for the substrate-binding protein that brings the branched-chain amino acids to the transporter, so they can be shuttled across the inner membrane to the cytoplasm. Mrub_2120 was found to have two paralogs present while b3458 only had one using the DB paralog search. This makes sense given the two are orthologous. Lastly, these two genes were the only two that had hits in the PDB database which were the 4EVQ crystal structure of ABC transporter solute binding protein in complex with 4-hydroxybenzoate with an e-value of 8.70812E-42 and the 1USG L-leucine-binding protein with an e-value of 0.00. Both of these

results were accurate as they indicate the functions of both genes and have e-values that make them significant hits.

The *E.coli* b3457 and Mrub_2121 share all the same outputs produced from the CDD, PFAM, TIGRfam and PDB databases. Both genes had the same CDD hit of LivH (COG0559) With acceptable e-values of 4.91e-87 and 1.08e-52, indicating the significance of the hits. Both genes also had the same PFAM hit of BPD_transp_2 (PF02653) with acceptable e-values of 6.7e-71 and 1e-40, indicating the significance of the hits. There was also an identical hit for both genes in the TIGRfam database of urea_t_UrtC_arc: urea ABC transporter, perm (TIGR03409) with acceptable e-values of 2.4e-15 and 3.1e-31, indicating the significance of the hits. The cell localization data suggests that these genes are localized/embedded in the inner plasma membrane and likely code for one of the TMD's of the transporter. Lastly, these two genes had no hits within the PDB database indicating that that have yet to be crystallized.

The *E.coli* b3456 and Mrub_2122 share all the same outputs produced from the CDD, PFAM, TIGRfam and PDB databases. Both genes had the same CDD hit of LivM (COG4177) with acceptable e-values of 5.12e-53 and 6.05e-78, indicating the significance of the hits. Both genes also had the same PFAM hit of BPD_transp_2 (PF02653) with acceptable e-values of 9e-60 and 9.5e-42, indicating the significance of the hits. There was also an identical hit for both genes in the TIGRfam database of urea_t_UrtC_arc: urea ABC transporter, perm (TIGR03727) with acceptable e-values of 1e-11 and 2.1e-09, indicating the significance of the hits. The cell localization data suggests that these genes are localized/embedded in the inner plasma membrane and likely code for one of the TMD's of the transporter. Lastly, these two genes had no hits within the PDB database indicating that that have yet to be crystallized.

The *E.coli* b3455 and Mrub_2123 share all the same outputs produced from the CDD, PFAM, TIGRfam and PDB databases. Both genes had the same CDD hit of LivG (COG0411) With acceptable e-values of 1.61e-139 and 1.43e-107, indicating the significance of the hits. Both genes also had the same PFAM hit of ABC Transporter (PF0005) with acceptable e-values of 3.1e-32 and 2.1e-12, indicating the significance of the hits. There was also an identical hit for both genes in the TIGRfam database of urea_trans_UrtD: urea ABC transporter (TIGR03411) with acceptable e-values of 3.7e-65 and 4.4e-71, indicating the significance of the hits. The cell localization data suggests that these genes are localized to the cytoplasm and likely code for one of the NBD's of the transporter. Mrub_2123 was found to have seventy two paralogs present while b3455 had seventy using the DB paralog search. This makes sense given the two are orthologous. Lastly, these two genes had no hits within the PDB database indicating that that have yet to be crystallized.

The *E.coli* b3454 and Mrub_2124 share all the same outputs produced from the CDD, PFAM, TIGRfam and PDB databases. Both genes had the same CDD hit of LivF (COG0410) With acceptable e-values of 5.11e-136 and 7.80e-84, indicating the significance of the hits. Both genes also had the same PFAM hit of ABC Transporter (PF0005) with acceptable e-values of 3.3e-33 and 7.4e-20, indicating the significance of the hits. There was also an identical hit for both genes in the TIGRfam database of urea_t_UrtE: urea ABC transporter (TIGR03410) with acceptable e-values of 4.7e-60 and 2.5e-32, indicating the significance of the hits. The cell localization data suggests that these genes are localized to the cytoplasm and likely code for one of the NBD's of the transporter. Both b3454 and Mrub_2124 had 73 hits using the DB paralog search. This would make sense given the two are orthologous. Lastly, these two genes had no hits within the PDB database indicating that that have yet to be crystallized.

This project is part of the *Meiothermus ruber* genome analysis project, which predicts gene function using the bioinformatics tools collected under the umbrella of the Guiding Education through Novel Investigation–Annotation Collaboration Toolkit (GENI-ACT). This study analyzed the genes Mrub_2120, Mrub_2121, Mrub_2122, Mrub_2123 and Mrub_2124 which can be translated to *livK, livH, livM, livG* and *livF*, respectively. Together, these genes form an operon encoding for an ABC transporter that selectively transports branched-chain amino acids across the intracellular plasma membrane of bacteria. Additionally, the suspected orthologs of these genes were analyzed in *E. coli* as well: b3458, b3457, b3456, b3455 and b3454, respectively. Upons analysis, it was found that Mrub_2120, Mrub_2121, Mrub_2122, Mrub_2123 and Mrub_2124 are orthologs of b3458, b3457, b3456, b3455 and b3454, respectively. Both operons make up the branched-chain amino acid transporter that shuttles leucine, isoleucine and valine from the periplasmic space to the cytoplasm in gram-negative bacteria.

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/.

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.

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.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignments. J Mol Biol. 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.

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.

Vasilis V, Vasiliou K and Nebert DW. 2009. “Human ATP-Binding Cassette (ABC) Transporter Family.” *Human Genomics*. 3(3): 281–290.

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.