Meiothermus ruber Genome Analysis Project

Biology

2017

Mrub_2294, Mrub_2293, and Mrub_1942 genes are orthologs of *E. coli* b2476, b1131, and b4006 genes, respectively

Lindsay Popp
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_ 2294, Mrub_2293, and Mrub_1942 genes are orthologs of *E.coli* b2476, b1131, and b4006 genes, respectively

Lindsay Popp
Dr. Lori R Scott Laboratory
Biology Department, Augustana College
639 38th Street, Rock Island, IL 61201

Introduction

Why Study Meiothermus Ruber?

Meiothermus ruber (M. ruber) grows in high temperature environments that can range from 35-70。C. It is a unique bacterium in that it is thermophilic, red-pigmented, gram-negative, and non-motile (Tindall *et al.*, 2010). *M. ruber* was first isolated in the Russian city of Kamchatka (Loginova *et. al.,* 1975). Unlike *E.coli*, there is little known information about *M. ruber*. *E.coli* has over 30,000 sources available on Pubmed, while *M.ruber* has approximately 28 publications online (Scott, personal communication). When there is so much information available on *E.coli*, one might ask why bother studying the *M.ruber* bacterium. The Joint Genome Institute initiated a project known as the Genomic Encyclopedia of Bacteria and Archaea (GEBA) that focuses on sequencing the genomes of lesser known bacteria like *M.ruber.* It is anticipated that scientists who study these lesser known species may identify new cellular not found in the well-studied species such as *E. coli.* Because so much is known about *E.coli*, as evidenced by the many online bioinformatics platforms dedicated to it (*e.g.,* Ecocyc (Keseler *et. al.*, 2013), it is used as a "positive" control in this study. We compare what is known in *E. coli* to fill the gaps of what we understand about *M. ruber*.

Purine Metabolism

As shown in Figure 1, the reaction involving the proteins PurC, PurB, and PurH is known as De novo purine nucleotide biosynthetic pathway. It is apart of purine metabolism (Senecoff & Meagher, 1993). Genes that encode for the enzymes in the reaction are indicated by the placement above the reaction arrows. The product and substrate for this type of reaction are catalyzed by AIR synthetase (purM, or Pur5, gene), which are identified. The intermediates used in this reaction are identified as follows: PRPP, 5-phosphoribosyl-a-1 -pyrophosphate; PRA, 5-phosphoribosylamine; CAR, 5‘-phosphoribosyC1-glycinamide; FGAR, 5’-phosphoribosyl-N-formylglycinamide; FGAM, 5’-phosphoribosyl-~-formylglycinamidine; AIR, 5’-phosphoribosyl-5-aminoimidazole; CAIR, 5’-phosphoribosyl-5-aminoimidazole carboxylic acid; SAICAR; 5’-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole; AICAR, 5’-phosphoribosyl-4-carboxamide-5-aminoimidazole; FAICAR, 5’-phosphoribosyl-4-carboxamide-5-formamidoimidazole. Above each gene is a number that is the classification for the analogous Arabidopsis Pur gene.

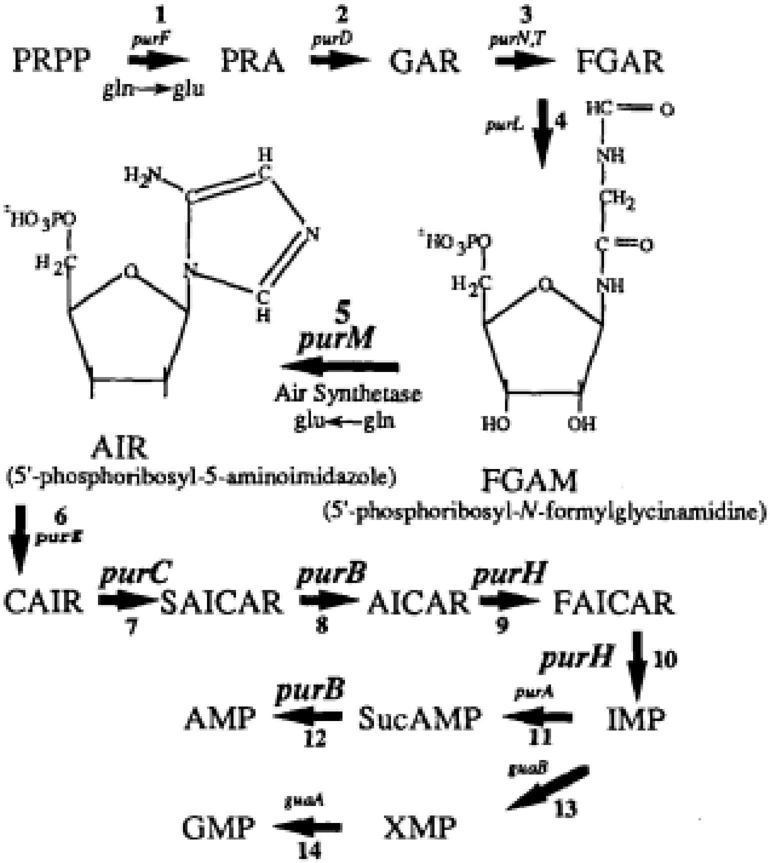


Figure 1. The de novo purine nucleotide biosynthetic pathway that is apart of purine metabolism. The proteins involved in this process are PurC, PurB, and PurH. The proteins cause the conversion of AIR to FAICAR, which utilizes the inosine 5′-monophosphate(IMP) compound. Genes/enzymes, substrates, products, and the most favorable direction for the reaction are shown in this image. Image taken from https://biocyc.org/ECOLI/NEW-IMAGE?type=REACTION&object=SAICARSYN-RXN

Purine Metabolic Pathway

Purine biosynthesis starts with the synthesis of inosine 5′-monophosphate (IMP) because of its versatility to be converted to all other purines. Figure 2 illustrates the pathway where this synthesis occurs. In *E. coli,* IMP is catalyzed by five enzymes from the molecule 5-amino-1-(5-phospho-D-ribosyl)imidazole (AIR) (Zhang, 2008). The first four steps are carried out by the enzymes phosphoribosylaminoimidazole-succinocarboxamide synthase (PurC), adenylosuccinate lyase (PurB), N^5-carboxyaminoimidazole ribonucleotide synthetase (PurK), and N^5-carboxyaminoimidazole ribonucleotide mutase (PurE), respectively. The final step involves the enzyme AICAR transformylase / IMP cyclohydrolase (PurH). In *E. coli*, the first step involves PurK converting the stable N5 to the unstable (AIR) N^5-carboxyaminoimidazole ribonucleotide (Mueller, 1994). Then, the molecule is converted to 5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate (CAIR) by PurE. The next reaction is carried out by PurC which catalyzes the amino group of aspartate to the carboxyl group of CAIR. The resulting product forms 5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole (SAICAR). The amino group in this compound has the nitrogen atom for the final purine base. The elimination of the remaining component of the aspartate molecule converts the substrate to fumarate in the next step, which coincides with the formation of aminoimidazole carboxamide ribonucleotide (AICAR) by PurB. The final step involves PurH catalyzing AICAR to phosphoribosyl-formamido-carboxamide (FAICAR) and subsequently to IMP with the elimination of a water molecule.

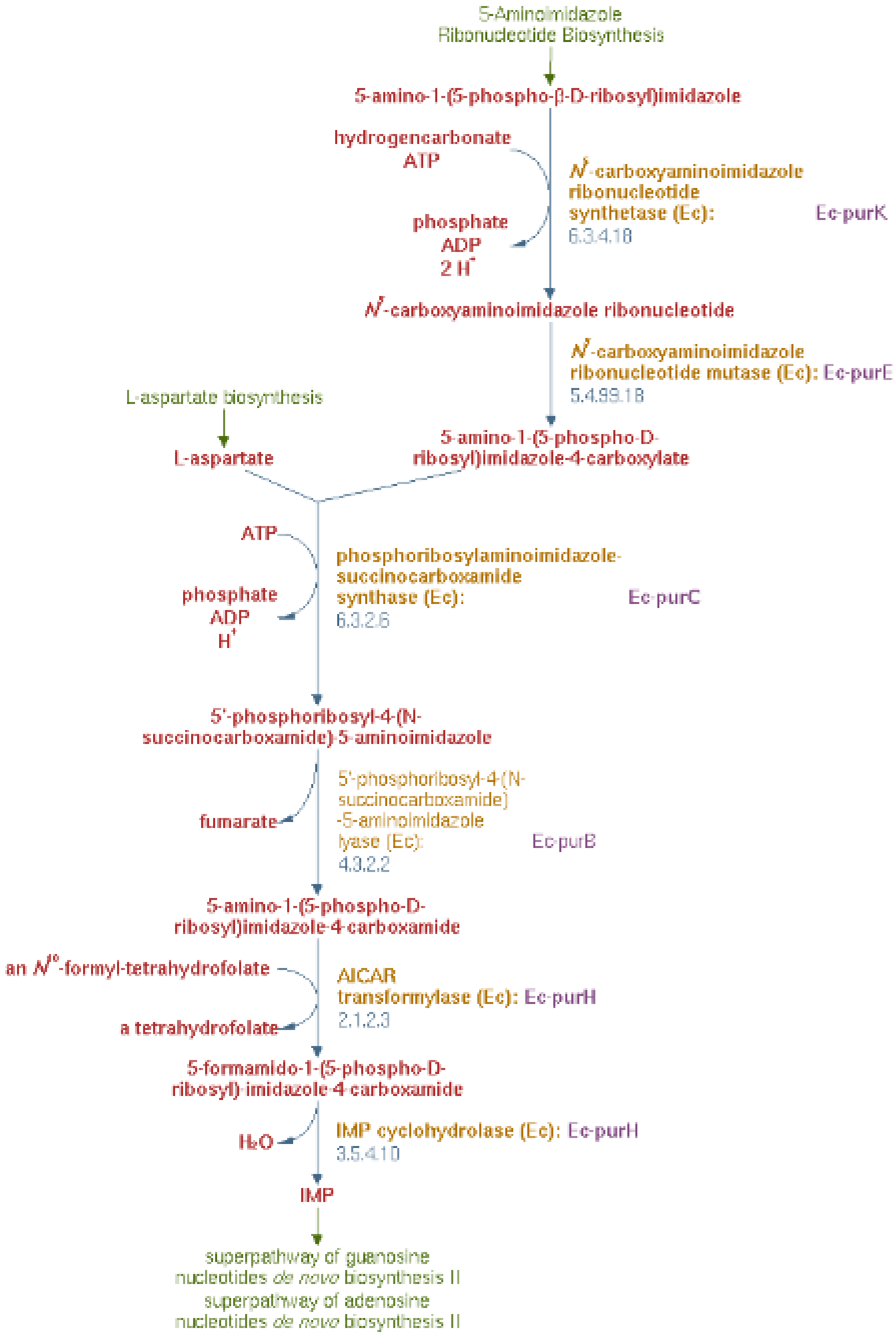

5-Aminoimidazole Ribonucleotide Biosynthesis
5-amino-1-(5-phospho-β-D-ribosyl)imidazole
hydrogencarbonate
ATP
phosphate
ADP
2 H^+
N^5-carboxyaminoimidazole ribonucleotide synthetase (Ec):
6.3.4.18
Ec-purK
N^5-carboxyaminoimidazole ribonucleotide
N^5-carboxyaminoimidazole ribonucleotide mutase (Ec): Ec-purE
5.4.99.18
L-aspartate biosynthesis
L-aspartate
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate
ATP
phosphate
ADP
H^+
phosphoribosylaminoimidazole-succinocarboxamide synthase (Ec):
6.3.2.6
Ec-purC
5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole
fumarate
5'-phosphoribosyl-4-(N-succinocarboxamide) -5-aminoimidazole lyase (Ec):
4.3.2.2
Ec-purB
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide
an N^{10}-formyl-tetrahydrofolate
a tetrahydrofolate
AICAR transformylase (Ec): Ec-purH
2.1.2.3
5-formamido-1-(5-phospho-D-ribosyl)-imidazole-4-carboxamide
H_2O
IMP cyclohydrolase (Ec): Ec-purH
3.5.4.10
IMP
superpathway of guanosine nucleotides *de novo* biosynthesis II
superpathway of adenosine nucleotides *de novo* biosynthesis II

Figure 2. Purine biosynthesis pathway showing the reactants, products, and genes/enzymes involved in the catalysis of the PurC, PurB, and PurH reactions in *E. coli* K12 MG1655. Image was taken from https://biocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=PWY-6123

Bioinformatics

All biological fields use bioinformatic tools to some extent, which make having knowledge about them very important for most careers (Persidis, 1999). These free resources can be a great time saving tool for those who know how to interpret the data outputs as well as know how to use them. The continuous advancement of technology allows for the safe storage of data when using these tools. Therefore, in order to achieve success in the biological field it is imperative that one needs to have a clear understanding of how to use and interpret the bioinformatics tool and data set. This project is designed to predict the function of previously uncharacterized open reading frames (ORF) from the *M. ruber* genome by comparing them to the bioinformatics outputs from known *E. coli* genes. We asked these questions: Is the ORF with the locus tag Mrub_2294 gene an ortholog of the *E. coli* b2476 gene (PurC); is the ORF with the locus tag Mrub_2293 gene an ortholog of the *E. coli* b1131 gene (PurB*)*; and is the Mrub_2294 gene an ortholog of the *E. coli* b4006 gene (PurH)? The bioinformatics programs available to us are housed within the GENI-ACT platform (Scott, personal communication), and are described in more detail in the Methods section. A common measure of sequence similarity between two aligned sequences, which is used by many of the bioinformatics tools in GENI-ACT, is the Expect value or E-value. The NCBI defines an Expect value as follows: "The E-value is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially with the score (S) that is assigned to a match between two sequences. Essentially, the E-value describes the random background noise that exists for matches between sequences. For example, an E-value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size, one might expect to see one match with a similar score simply by chance. This means that the lower the E-value, or the closer it is to "0", the higher is the "significance" of the match. However, it is important to note that searches with short sequences can be virtually identical and have relatively high E-value. This is because the calculation of the E-value also takes into account the length of the query sequence. This is because shorter sequences have a high probability of occurring in the database purely by chance." (Madden, 2002). For this project, an E-value greater than 0.001 indicates that two sequences have little structural similarity to each other; little structural similarity is assumed to mean little to no common functionality. A low E-value show a significant match between two sequences. An initial protein BLAST comparison was performed between *E. coli* purC, purB, and pur*H* against Mrub_2294, Mrub_2293, and Mrub_1942 genes. Low E-values were given as a result with data presented respectively: 2e-166, 0.0, and 0.0 respectively.

Hypothesis.

Based on this data output, we hypothesize that the three *M. ruber* ORFs are orthologs to the *E. coli* genes, respectively.

Methods

To study ORFs in the *M. ruber* genome, we used a model organism *E. coli* as a positive control. This organism has been studied a great deal in the laboratory and is easy to grow in a lab setting. This allows a smooth process of analyzing similar genes in our microorganism. In order to gather accurate information on the genes studied being orthologs to one another the GENI-ACT gene annotation procedure was followed to a certain extent (http://www.geni-act.org/education/main/). An initial protein BLAST (Madden, 2002) was performed to identify similarities between the *M.ruber* genes (Mrub_2294, Mrub_2993, Mrub_1942) and the presumed *E.coli* orthologs, the data for which was uploaded into the lab notebook of the GENI-ACT tool kit. The bioinformatic tools housed within the GENI-ACT site are as follows: CDD (Marchler-Bauer *et al.*, 2016), T-Coffee (Notredame *et al.*, 2000), KEGG (Kanehisa *et al.*, 2016), Weblogo (Crooks *et al.*, 2004), PDB (Berman *et al.*, 2000), LipoP (Juncker *et al.*, 2003), IMG (Markowitz *et al.*, 2012), TMHMM (Krogh *et al.*, 2016), TIGRfam (Haft *et al.*, 2001), PSORTB (Yu *et al.*, 2010), Pfam (Finn *et al.*, 2016), Ecocyc (Keseler *et al.*, 2013), Phobius (Kall *et al.*, 2004), and SignalP (Petersen *et al.*, 2011). Two significant steps were performed that were not described in the GENI-ACT instructions. First, the basic information of the gene including the amino acid sequence and nucleotide sequence was acquired. The amino acid sequence was put into the FASTA format and utilized for the rest of the modules. The protein BLAST was utilized to identify the top sequences within the database that held the most similarity(Madden, 2002). The top 250 hits with paired E-values of 0.01 or less were obtained in each protein BLAST search. The results were used to gather information on the sequence hits, as well as on the CDD section. Additionally, 15 sequences of varying genus from the protein BLAST results were selected and used to create a multiple sequence alignment through T-Coffee (Notredame *et al.*, 2000). The alignment was used to create a Weblogo about the amino acid sequence(Crooks *et al*, 2004).Second, we collected 15 sequences from a BLAST database search for the T-Coffee program instead of 10 hits described in the instructions. In addition, we excluded the *E.coli* and proteobacteria from the BLAST search to acquire a greater diversity of proteins for the multiple sequence alignment and Weblogo. We used the KEGG website for duplication and degradation of *M. ruber* instead of using the original Metacyc website. Another variation in the procedure was that we used the color by KEGG map when determining the Gene Context section of the lab notebook for the Horizontal Gene Transfer module. To search for the locus tags of interest IMG was used. After the locus tag information was found in IMG, the option to color the maps by GC was selected to view the GC heat maps for the genes. This setting allowed us to gather information regarding the characteristic GC% for the genes, which was provided at the top of the returned page. The average GC% for the genes was obtained as well, which was gathered by hovering over the gene of interest within the map. Finally, the IMG results were colored by KEGG map to identify if the gene was an ortholog or not(Markowitz *et al.*, 2012). Therefore, *E. coli* is not only easy to grow and widely studied, but also explains how *M. ruber* genes (Mrub_1942, Mrub_2294, Mrub_2293) can be orthologs to its own genes and the reason for its high interest from scientific researchers(Cooper, 2002).

RESULTS

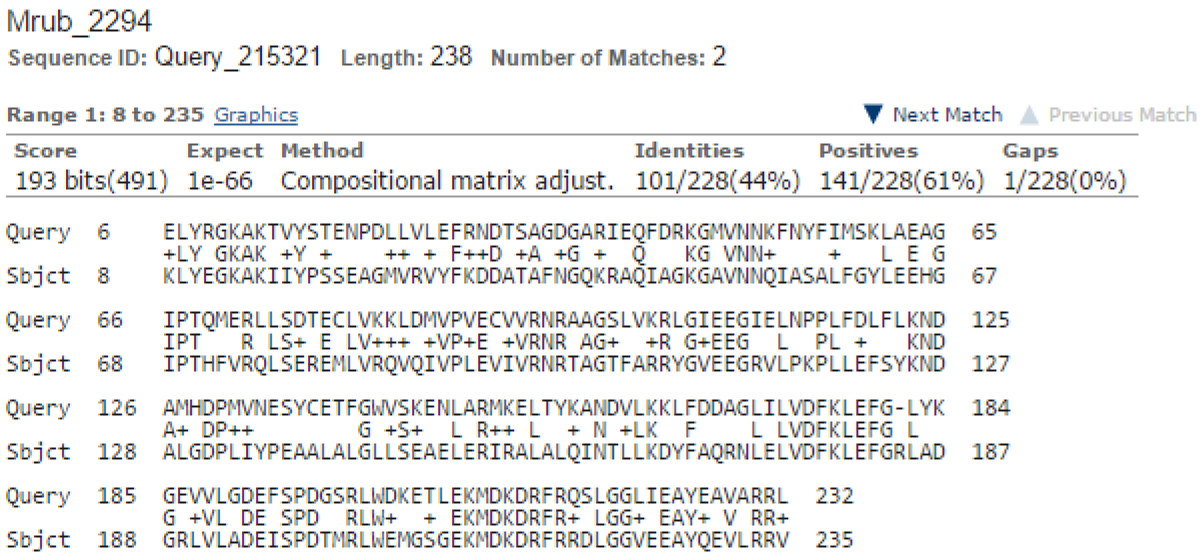
Table 1 summarizes the bioinformatics tools that we used to compare the *E. coli* b2476 gene to the Mrub_2294 gene. The first row of the table shows the initial protein BLAST. The fairly high bit score is not a problem because the sequences vary in length. The bit score of 1e-66 is close to zero, which shows that the results are significant and that it is unlikely that the proteins aligned due to chance alone. The results provide further evidence that the two genes are orthologs of one another and likely share the same ancestor. A search of the CDD with both query sequences pulled the same COG numbers (COG0152) and domain name [Phosphoribosylaminoimidazole-succinocarboxamide synthase (Nucleotide transport and metabolism)]. Both proteins were located in the cytoplasm by using the cellular localization bioinformatics tools (TMH, SignalP, LipoP and PSORT-B) and did not contain cleavage sites. A search of the TIGRfam databases pulled the same protein family (TIGR0008, PurC: phosphoribosylaminoimidazolesuccinoca. However, the PDB showed separate results with a different number and name for both *E. coli* and *M. ruber*. There was a difference in organism.The enzyme commission number and name were identical and as follows: EC 6.3.2.6 phosphoribosylaminoimidazolesuccinocarboxamide synthase. By using the KEGG website conclusions were able to be made that both genes were predicted to be apart of the same step for the Purine metabolism pathway. In conclusion, the results show that the genes are orthologs to one another and share very similar structural characteristics.

Table 1. *E. coli* b2476 and Mrub_2294 are orthologs

Bioinformatics Tools Used	***E.coli* b2476 gene (PurC)**	**Mrub_2294 gene**
BLAST *E. coli* against *M. ruber*	Score: 207 E-value: 1e-66	
CDD Data (COG category)	COG NUMBER: COG0152 Phosphoribosylaminoimidazole-succinocarboxamide synthase [Nucleotide transport and metabolism] E-value: 9.12e-116	E-value: 1e-233
Cellular Localization	Cytoplasm of the cell	

TIGRfam-protein family	TIGR00081 purC: phosphoribosylaminoimidazolesuccinoca E-value: 9e-104	E-value:4.4e-143
Pfam-protein family	PF01259 SAICAR synthetase E-value: 4e-66	E-value: 3.8e-59
Protein Database (PDB)	2GQR, SAICAR Synthetase Complexed with ADP-Mg2+ E-value: 5.24e-137	2YWV, Crystal structure of SAICAR synthetase from Geobacillus kaustophilus E-value: 1.18e-66
Enzyme Commission Number	EC 6.3.2.6 phosphoribosylaminoimidazolesuccinocarboxamide synthase	
KEGG Pathway Map	Purine Metabolism, 00230	

Figure 3 is the protein BLAST alignment of *E. coli* PurC versus Mrub_2294. From the results shown 44% of the amino acids were identical, while 141 were similar. The E-value is 1e-66, which well below the 0.001 cutoff. This shows that the amino acid sequence is not aligned due to random chance alone and shows structural similarity between the two proteins. This is the first piece of evidence that supports the hypothesis that *E. coli* PurC and Mrub_2294 are orthologs to one another.



Mrub_2294

Sequence ID: Query_215321 Length: 238 Number of Matches: 2

Range 1: 8 to 235 Graphics ▼ Next Match ▲ Previous Match

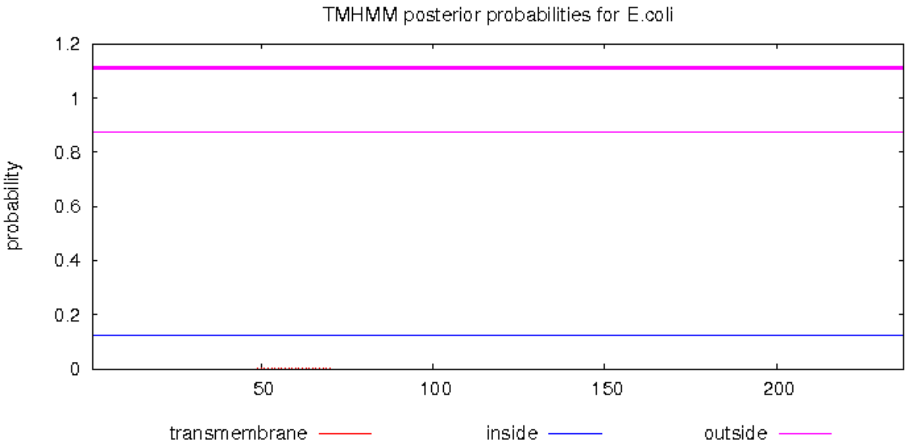
Score	Expect	Method	Identities	Positives	Gaps
193 bits(491)	1e-66	Compositional matrix adjust.	101/228(44%)	141/228(61%)	1/228(0%)

```
Query  6    ELYRGKAKTVYSTENPDLLVLEFRNDTSAGDGARIEQFDRKGMVNNKFNYFIMSKLAEAG  65
            +LY GKAK +Y +    ++ + F++D +A +G +  Q   KG VNN+    +   L E G
Sbjct  8    KLYEGKAKIIYPSSEAGMVRVYFKDDATAFNGQKRAQIAGKGAVNNQIASALFGYLEEHG  67

Query  66   IPTQMERLLSDTECLVKKLDMVPVECVVRNRAAGSLVKRLGIEEGIELNPPLFDLFLKND  125
            IPT   R LS+ E LV+++ +VP+E +VRNR AG+  +R G+EEG  L  PL +   KND
Sbjct  68   IPTHFVRQLSEREMLVRQVQIVPLEVIVRNRTAGTFARRYGVEEGRVLPKPLLEFSYKND  127

Query  126  AMHDPMVNESYCETFGWVSKENLARMKELTYKANDVLKKLFDDAGLILVDFKLEFG-LYK  184
            A+ DP++        G +S+  L R++ L  + N +LK  F    L LVDFKLEFG L
Sbjct  128  ALGDPLIYPEAALALGLLSEAELERIRALALQINTLLKDYFAQRNLELVDFKLEFGRLAD  187

Query  185  GEVVLGDEFSPDGSRLWDKETLEKMDKDRFRQSLGGLIEAYEAVARRL  232
            G +VL DE SPD  RLW+  + EKMDKDRFR+ LGG+ EAY+ V RR+
Sbjct  188  GRLVLADEISPDTMRLWEMGSGEKMDKDRFRRDLGGVEEAYQEVLRRV  235
```

Figure 3. Mrub_2294 and *E. coli* PurC have similar protein sequence. Subject sequence is Mrub_2294 and query sequence is *E. coli* PurC. Analysis was performed by using NCBI BLAST bioinformatics tool at https://blast.ncbi.nlm.nih.gov/Blast.cgi.

Figure 4 describes one of several programs used to determine the cellular location of these proteins. Each panel shows the TMH hydropathy plot for *E. coli* b2476 or Mrub_2294. The presence of red peaks on a plot indicates the presence of transmembrane helices, however, there are no peaks on either plot for our query sequences. Consequently, there are no predicted transmembrane helices in *E. coli* b2476 and Mrub_2294. This is contrary to what we anticipated because the *E. coli* protein is described as having both a cytoplasmic and membrane location (Keseler *et. al.*, 2013). However, the JGI's IMG Gene Details page for b2476 predicts no TMH (JGI, 1997). That both proteins appear to be missing a transmembrane helices could mean that another type of secondary structure attaches it to the cell membrane for both proteins.

Panel A



Panel B

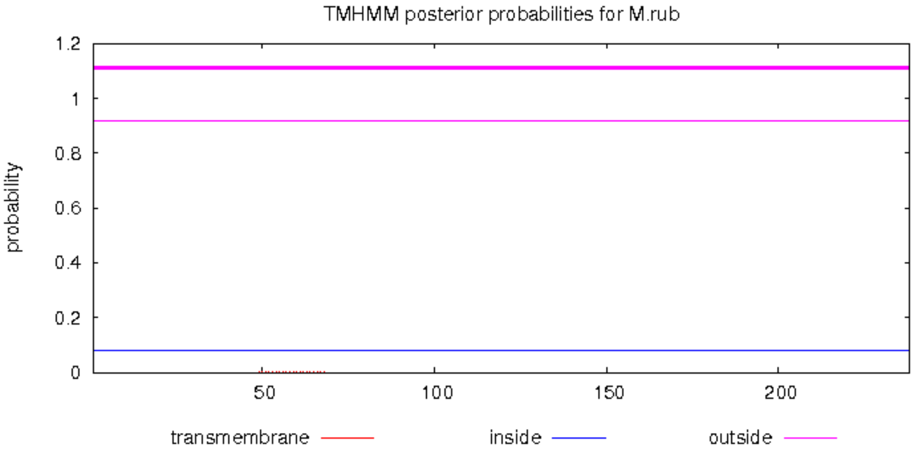
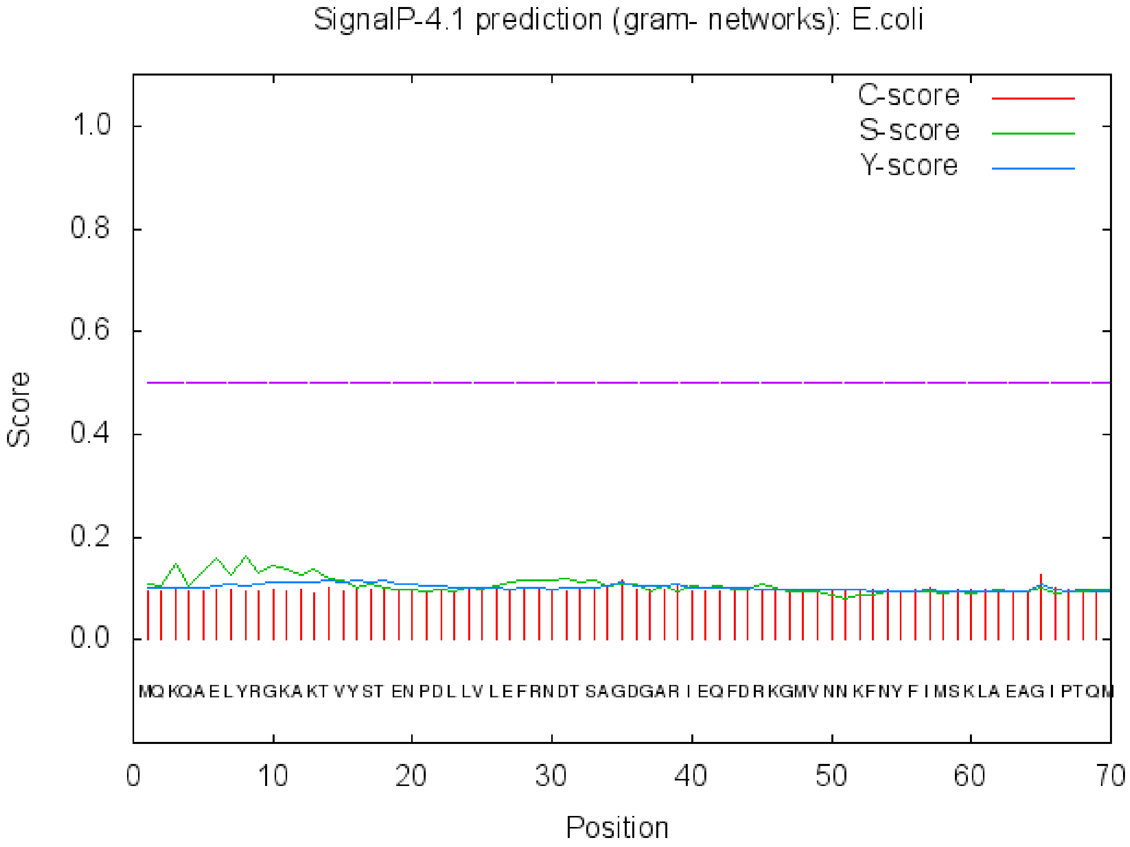


Figure 4. Both *E. coli* b2476 and Mrub_2294 do not contain TMH regions. Both are predicted to be located in the cytoplasmic region. Panel A shows the TMHMM for *E. coli* PurC/b2476; Panel B shows the TMHMM data for Mrub_2294. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was used to create these two hydropathy charts.

Figure 5 shows the SignalP graphs generated by both *E. coli* b2476 and Mrub_2294. The signalP bioinformatics tool is used to predict protein cleavage sites by assigning a D-value to each graph. This D-value is calculated from the S-value and Y-value. Alongside the D-value is a cut off value given by the purple line on the graph. For *E.coli* b2476 in Panel A the D-value of 0.122 is below the cut off value of about 0.500. In Mrub_2294 seen in Panel B, the D-value of 0.104 is below the cut off value of about 0.500. This data supports the fact that neither gene has a protein cleavage site and therefore doesn't attach or cross over the membrane. Also, both genes had start codons that were a correct distance away from a putative Shine-Dalgarno sequence as determined by the JGI IMG Sequence Viewer For Alternate ORF Search (Markowitz *et al.*, 2012). Consequently, we hypothesize that the correct reading frame has been identified for Mrub_2294.

Panel A



Panel B

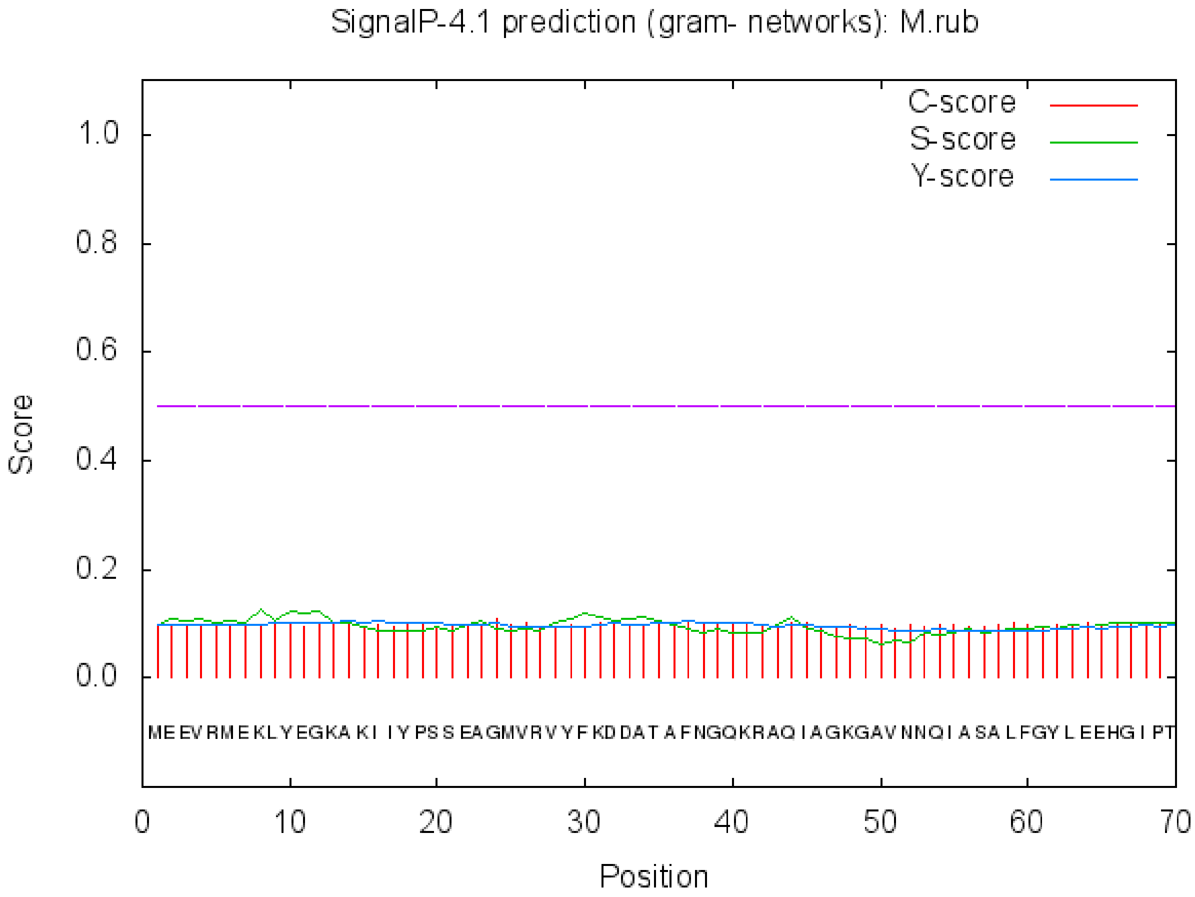
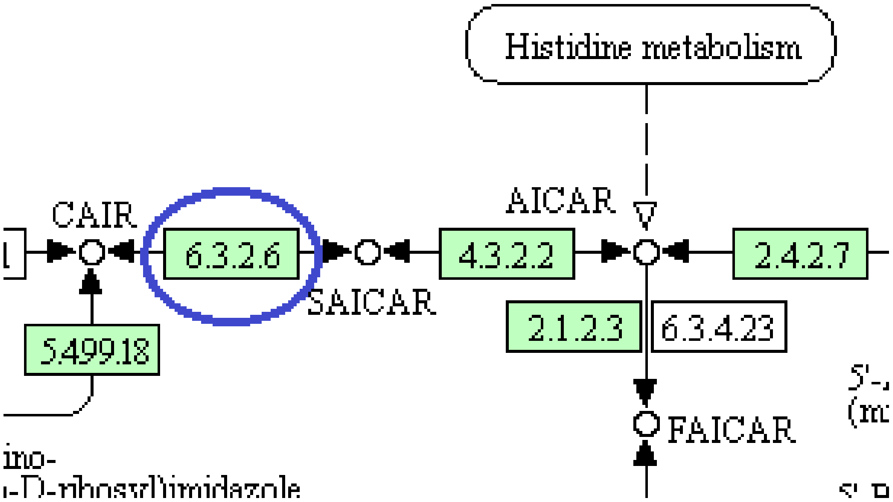


Figure 5. Cleavage sites are not present in *E. coli* b2476 and Mrub_2294. *E. coli* b2476 has a D value (D=0.104) and Mrub_2294 has a D value (D=0.122). Both values were below the cut off value. Panel A shows the plot for *E. coli* b2476; Panel B shows the plot for Mrub_2294. Signal P server v 4.1 http://www.cbs.dtu.dk/services/SignalP created these plots.

Figure 6 displays the purine metabolism biochemical pathway predicted to contain our query proteins. The green boxes represent enzymes that are present in the pathway. In the pathway it shows how both genes are involved in the conversion of CAIR to SAICAR. This provides further evidence that these two genes are orthologs of one another.

Panel A



Panel B

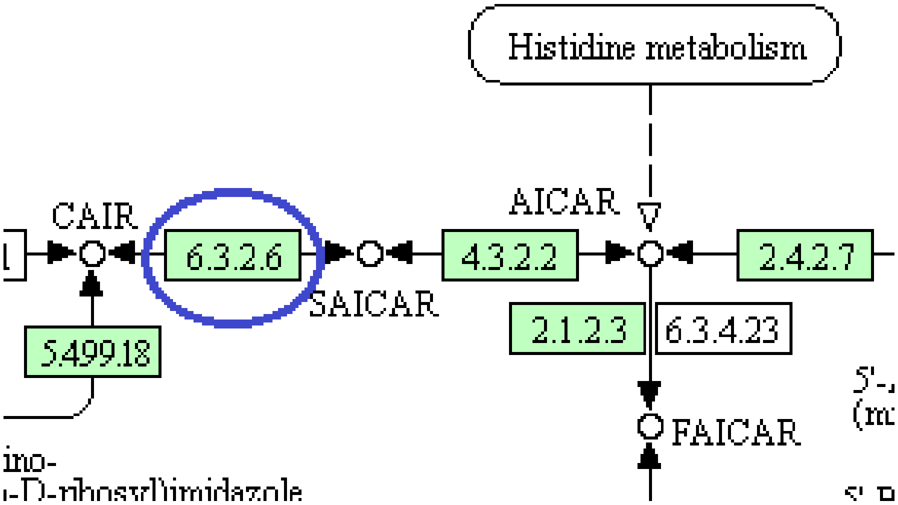
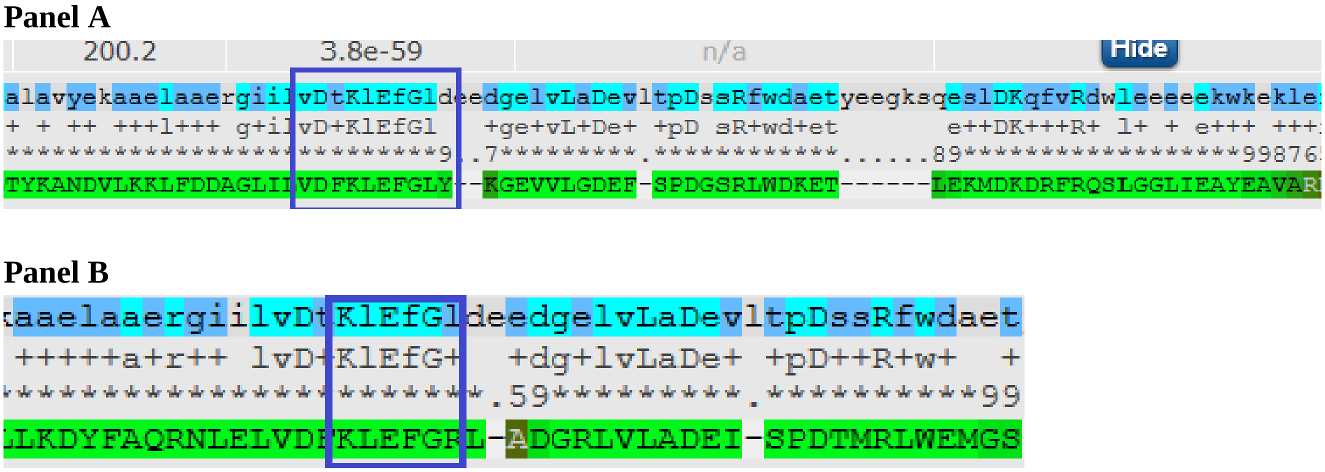


Figure 6. *E. coli* b2476 and Mrub_2294 are in the same biological pathway. Panel A shows the KEGG pathway for *E. coli* b2476, while Panel B shows the pathway for Mrub_2294. In order to find the purine metabolism pathway The Kyoto Encyclopedia of Genes and Genomes (KEGG) database at http://www.genome.jp/kegg/pathway.html was used.

The figure below shows which amino acids are highly conserved. Both pairwise alignments have aspartate, lysine, glutamate, and glycine as shared conserved amino acids. This alignment is different from the BLAST search in that it compares a query sequence to a Pfam consensus sequence, which is made up of hundreds of proteins. This further supports our hypothesis that the two genes are orthologs because they both created similar pairwise alignments to one another.



Panel A

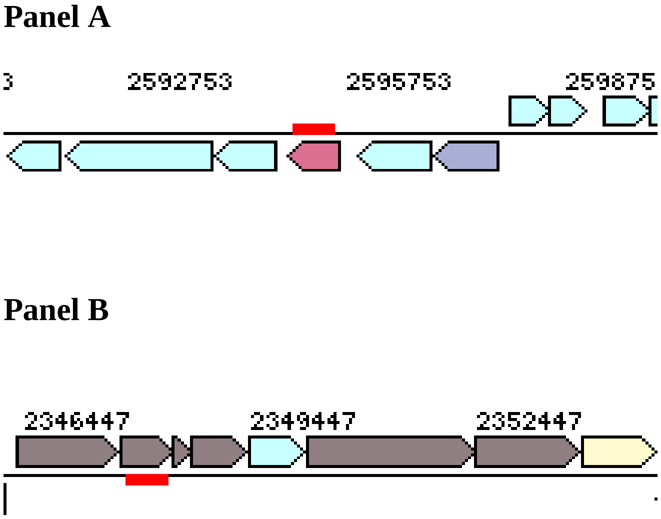
```
200.2    3.8e-59    n/a    Hide
alavyekaaelaaergiilvDtKlEfGldeedgelvLaDevltpDssRfwdaetyeegksqeslDKqfvRdwleeeeekwkekle
+ + ++ +++l+++ g+ilvD+KlEfGl   +ge+vL+De+ +pD sR+wd+et      e++DK+++R+ l+ + e+++ +++
****************************9..7*********.************......89******************99876
TYKANDVLKKLFDDAGLIIVDFKLEFGLY--KGEVVLGDEF-SPDGSRLWDKET------LEKMDKDRFRQSLGGLIEAYEAVAR
```

Panel B

```
kaaelaaergiilvDtKlEfGldeedgelvLaDevltpDssRfwdaet
 +++++a+r++ lvD+KlEfG+   +dg+lvLaDe+ +pD++R+w+  +
***********************.59*********.**********99
LLKDYFAQRNLELVDFKLEFGRL-ADGRLVLADEI-SPDTMRLWEMGS
```

Figure 7. Represents the highly conserved amino acids in the *E. coli* b2476 and Mrub_2294 alignments. Both have similar conserved amino acids to one another and are identified in the above figure. Also, they have the same Pfam name (SAICAR synthetase). The pairwise alignment was created by the Pfam website http://pfam.sanger.ac.uk/search.

Figure 8 represents the ortholog neighborhoods that are seen from color by KEGG. The genes surrounding the query gene in both Panel A and Panel B are not all the same color, which means that they are not made up of an operon. Although the two query genes are not the same exact color, they still provide evidence that they are orthologs because they are both a shade of purple. This indicates they share a similar function of nucleotide transport and metabolism (Markowitz *et. al.*, 2012).



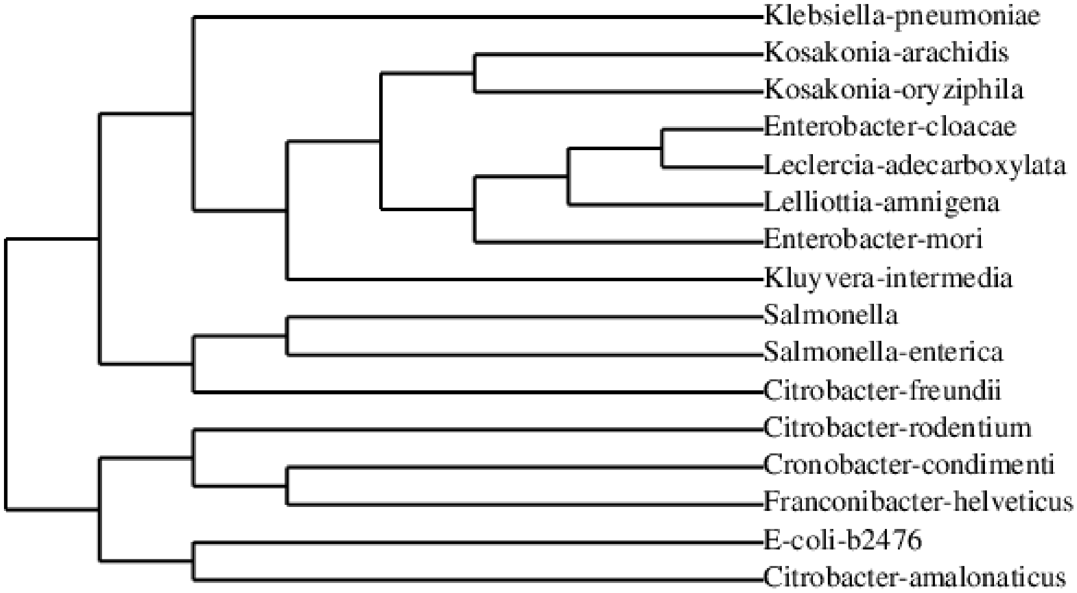
Panel A

Panel B

Figure 8. *E. coli* b2476 and Mrub_2294 are not part of an operon. The gene being viewed has a red bar above or below it. Panel A is *E. coli* b2476 and Panel B is Mrub_2294. Color by KEGG was used for this figure. Image was taken from http://img.jgi.doe.gov/.

Figure 9 shows the traditional phylogenetic trees for *E. coli* and *M. ruber*. Panel A has genus that are all part of the Proteobacteria phylum, while Panel B has genus made up of either Deinococcus-Thermus or Firmicutes. Since the species most closely related to Mrub_2294 are within the same phylum, we are predicting that no HGT has occurred in the recent evolutionary history of this gene.

Panel A



Panel B

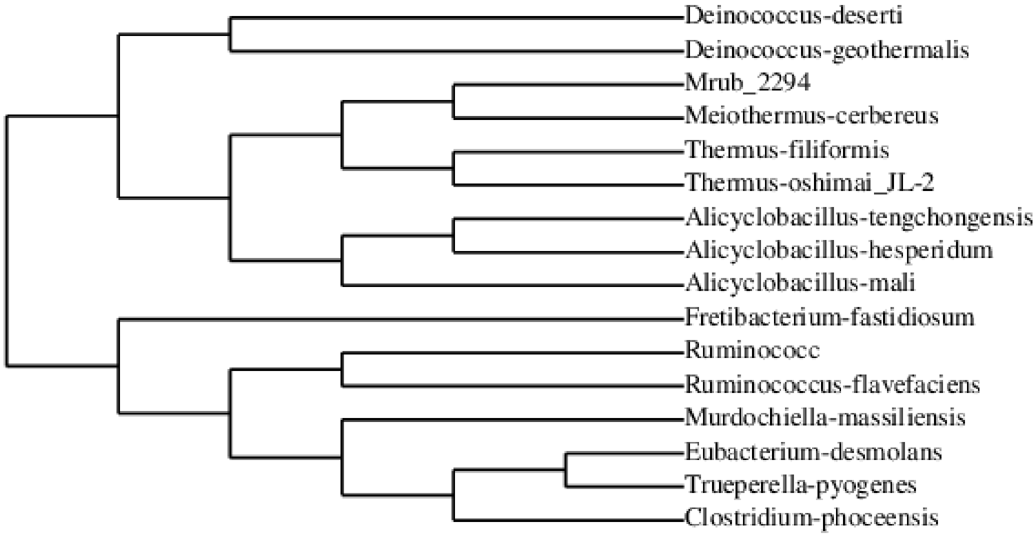


Figure 9. Both Panel A and Panel B are not likely to be HGT. Panel A represents *E. coli* b2476 and Panel B represents Mrub_2294. Both phylogenetic trees were created by http://www.phylogeny.fr.

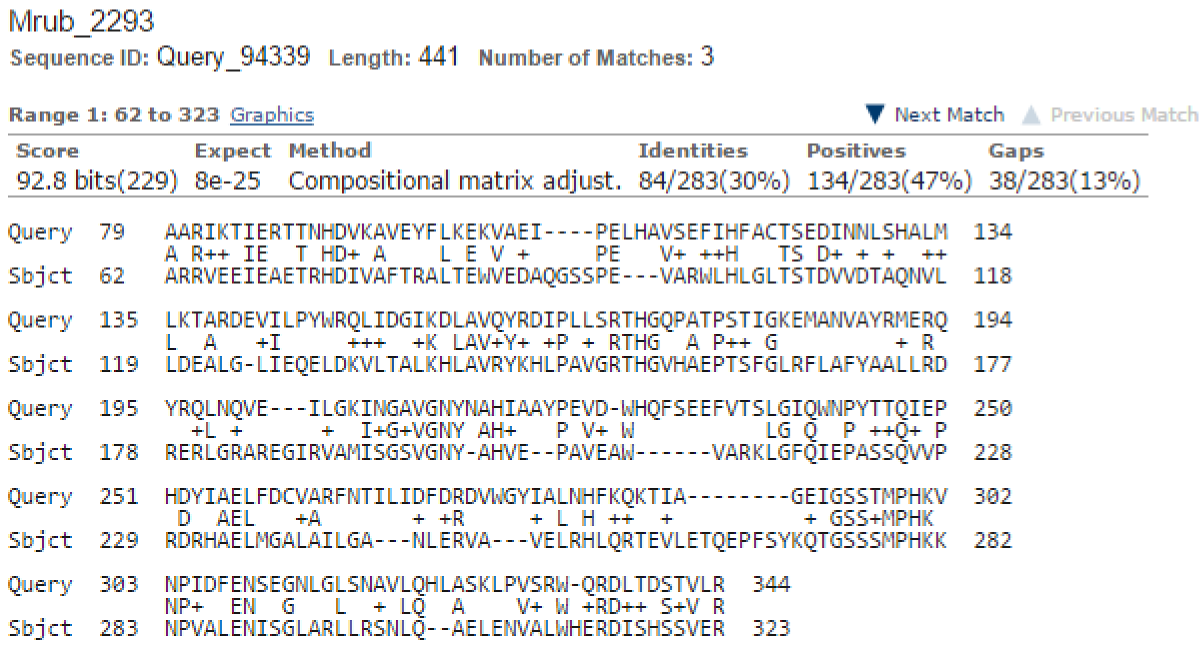
Table 2 summarizes the bioinformatics tools that we used to compare the *E. coli* b1131 gene to the Mrub_2293 gene. The first row of the table shows the initial BLAST done on the genes that was stated earlier in the introduction. The moderately high bit score is not a problem because the sequences vary in length. The E-value score of 8e-25 is close to zero, which shows that the results are significant and that it is unlikely that the proteins aligned due to chance alone. The low value shows similarity of amino acids between the genes as well. The results provide further evidence that the two genes are orthologs of one another and share the same ancestor. The CDD showed similar results as well with matching COG numbers (COG0015) and names (Adenylosuccinate lyase [Nucleotide transport and metabolism]). Both genes were located in the cytoplasm. Also, the TIGRfam-protein family showed similarity with identical numbers and names: TIGR00928, PurB: adenylosuccinate lyase. However, the PDB showed separate results with a different number and name for both *E. coli* and *M.ruber*. There was a difference in organism. The enzyme commission number and name were identical and as follows: EC 6.3.2.6, adenylosuccinate lyase. Finally, both genes were predicted to be apart of the same step for the Purine metabolism pathway. In conclusion, the results show that the genes are orthologs to one another and share very similar characteristics.

Table 2. *E. coli* b1131 and Mrub_2293 are orthologs

Bioinformatics Tools Used	***E. coli* b1131 gene (PurB)**	**Mrub_2293 gene**
BLAST *E. coli* against *M. ruber*	Score: 92.8 E-value: 8e-25	
CDD Data (COG category)	COG NUMBER: COG0015 Adenylosuccinate lyase [Nucleotide transport and metabolism] E-value: 2.48e-173	E-value: 4.75e-171
Cellular Localization	Cytoplasm of the cell	
TIGRfam-protein family	TIGR00928 PurB: adenylosuccinate lyase E-value: 2.5e-249	E-value:2.6e-175
Pfam-protein family	PF00206 Lyase E-value: 1.3e-84	E-value: 4.1e-53
Protein Database (PDB)	2PTQ , Crystal structure of Escherichia coli adenylosuccinate lyase mutant H171N with bound AMP and fumarate E-value: 0.0	1F1O , STRUCTURAL STUDIES OF ADENYLOSUCCINATE LYASES E-value: 5.31e-106
Enzyme Commission Number	EC 4.3.2.2 adenylosuccinate lyase	
KEGG Pathway Map	Purine Metabolism, 00230	

Figure 10 explains the BLAST alignment results of *E. coli* b1131 versus Mrub_2293. From the results shown 30% of the amino acids were identical, while 134 were similar. The E-value is 8e-

25, which is fairly small and closer to zero. This shows that the amino acid sequence is not aligned due to random chance alone and shows structural similarity between the two genes. This is the first piece of evidence that supports the hypothesis that *E. coli* b1131 and Mrub_2293 are orthologs to one another.



Mrub_2293

Sequence ID: Query_94339 Length: 441 Number of Matches: 3

Range 1: 62 to 323 Graphics

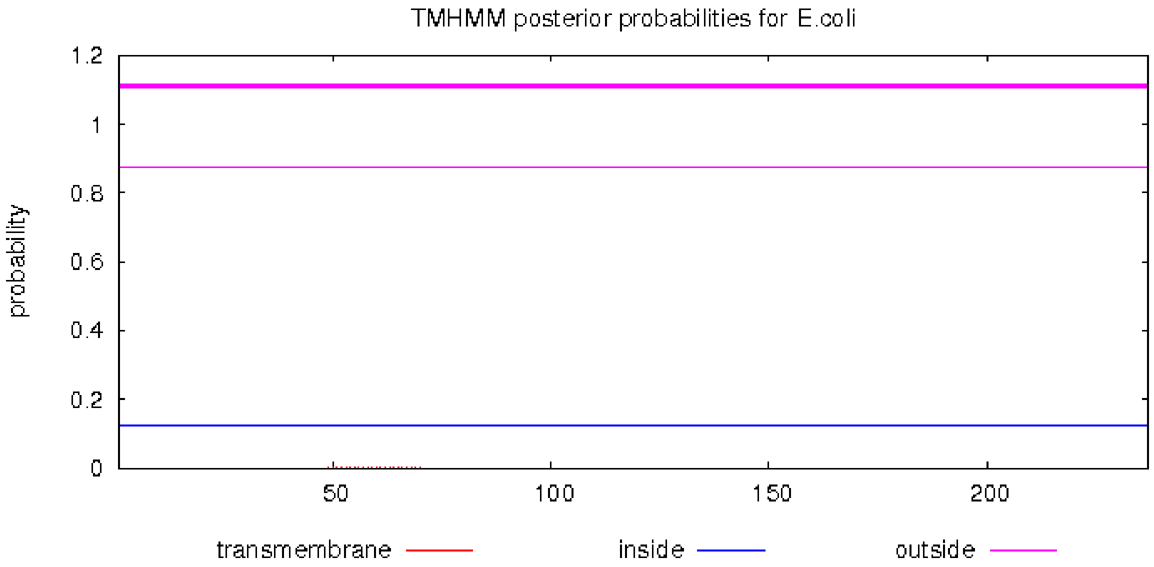
Score	Expect	Method	Identities	Positives	Gaps
92.8 bits(229)	8e-25	Compositional matrix adjust.	84/283(30%)	134/283(47%)	38/283(13%)

```
Query  79   AARIKTIERTTNHDVKAVEYFLKEKVAEI----PELHAVSEFIHFACTSEDINNLSHALM  134
            A R++ IE  T HD+ A    L E V +     PE   V+ ++H   TS D+ + +  ++
Sbjct  62   ARRVEEIEAETRHDIVAFTRALTEWVEDAQGSSPE---VARWLHLGLTSTDVVDTAQNVL  118

Query  135  LKTARDEVILPYWRQLIDGIKDLAVQYRDIPLLSRTHGQPATPSTIGKEMANVAYRMERQ  194
            L  A   +I     +++  +K LAV+Y+ +P + RTHG  A P++ G         + R
Sbjct  119  LDEALG-LIEQELDKVLTALKHLAVRYKHLPAVGRTHGVHAEPTSFGLRFLAFYAALLRD  177

Query  195  YRQLNQVE---ILGKINGAVGNYNAHIAAYPEVD-WHQFSEEFVTSLGIQWNPYTTQIEP  250
              +L +      +  I+G+VGNY AH+   P V+ W          LG Q  P ++Q+ P
Sbjct  178  RERLGRAREGIRVAMISGSVGNY-AHVE--PAVEAW------VARKLGFQIEPASSQVVP  228

Query  251  HDYIAELFDCVARFNTILIDFDRDVWGYIALNHFKQKTIA--------GEIGSSTMPHKV  302
             D  AEL   +A       + +R     + L H ++  +          + GSS+MPHK
Sbjct  229  RDRHAELMGALAILGA---NLERVA---VELRHLQRTEVLETQEPFSYKQTGSSSMPHKK  282

Query  303  NPIDFENSEGNLGLSNAVLQHLASKLPVSRW-QRDLTDSTVLR  344
            NP+  EN  G   L  + LQ  A    V+ W +RD++ S+V R
Sbjct  283  NPVALENISGLARLLRSNLQ--AELENVALWHERDISHSSVER  323
```

Figure 10. Mrub_2293 and *E. coli* b1131 have similar protein sequence. Subject sequence is Mrub_2293 and query sequence is *E. coli* b1131. Analysis was performed by using NCBI BLAST bioinformatics tool at https://blast.ncbi.nlm.nih.gov/Blast.cgi.

Figure 11 describes the cellular location for both genes. Each panel shows the TMH hydropathy plot for *E. coli* b1131 or Mrub_2293. The presence of red peaks on the plot represent transmembrane helices, however, there are no peaks on either plot meaning there are no transmembrane helices in *E. coli* b1131 and Mrub_2293 (JGI, 1997). Additionally, both plots are consistent with one another proving significance that both genes are located in the cytoplasm opposed to the membrane (Keseler *et. al.*, 2013).

Panel A



Panel B

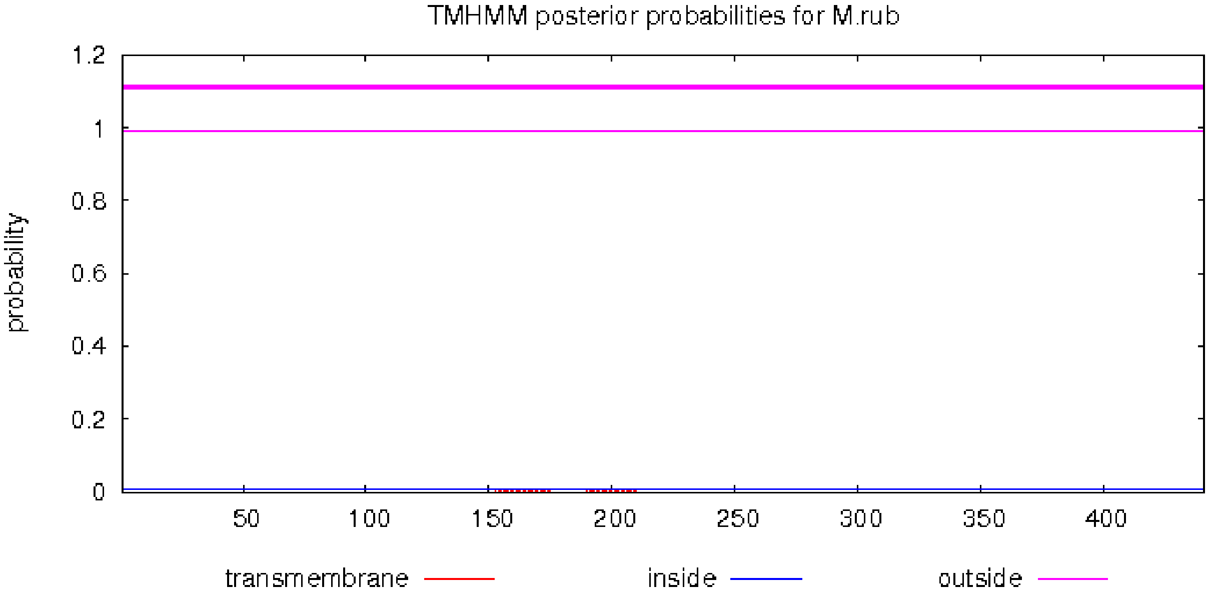
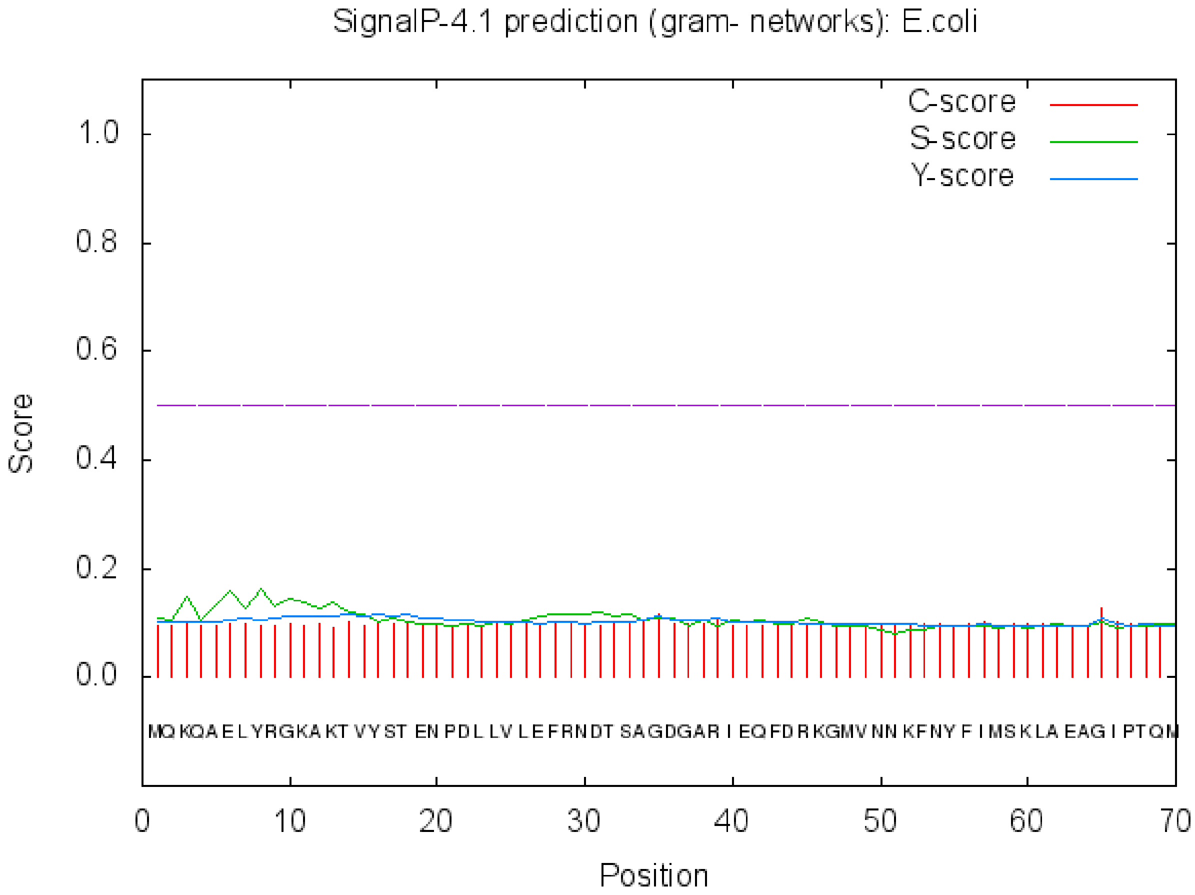


Figure 11. Both *E. coli* b1131 and Mrub_2293 do not contain TMH regions. Both are predicted

to be located in the cytoplasmic region. Panel A shows the TMHMM for *E. coli* b1131; Panel B shows the TMHMM data for Mrub_2293. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was used to create these two hydropathy charts. Figure 12 shows the SignalP graphs generated by both *E. coli* b1131 and Mrub_2293. The signalP bioinformatics tool is used to predict protein cleavage sites by assigning a D-value to each graph. This D-value is calculated from the S-value and Y-value. Alongside the D-value is a cut off value given by the purple line on the graph. For *E. coli* b1131 in Panel A the D-value of 0.131 is below the cut off value of about 0.500. In Mrub_2293 seen in Panel B, the D-value of 0.110 is below the cut off value of about 0.500. This data supports the fact that neither gene has a protein cleavage site and therefore doesn't attach or cross over the membrane. Also, both genes had start codons that were a correct distance away from the shine-dalgarno sequence and were in the correct reading frame. No new reading frames had to be proposed for the JGI/IMG module (Markowitz *et al.*, 2012).

Panel A



Panel B

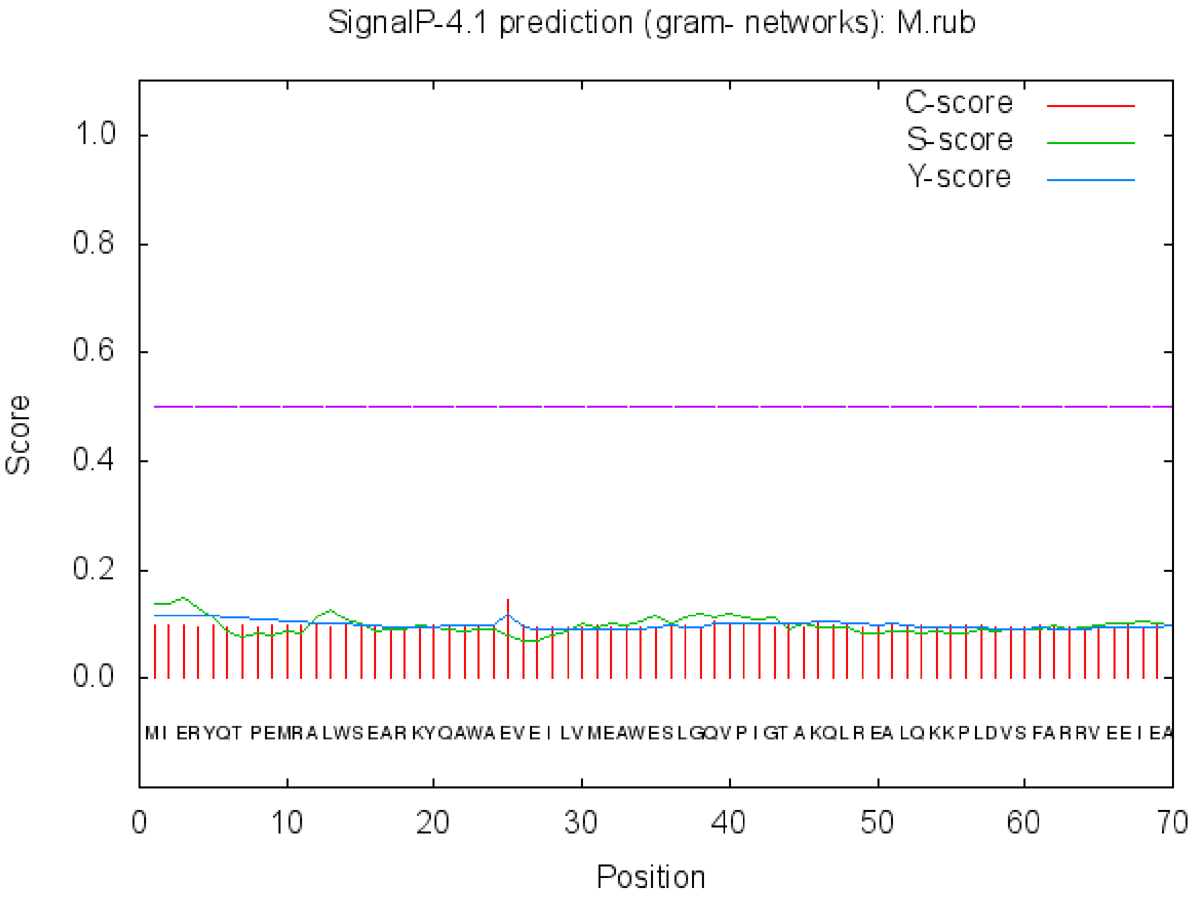
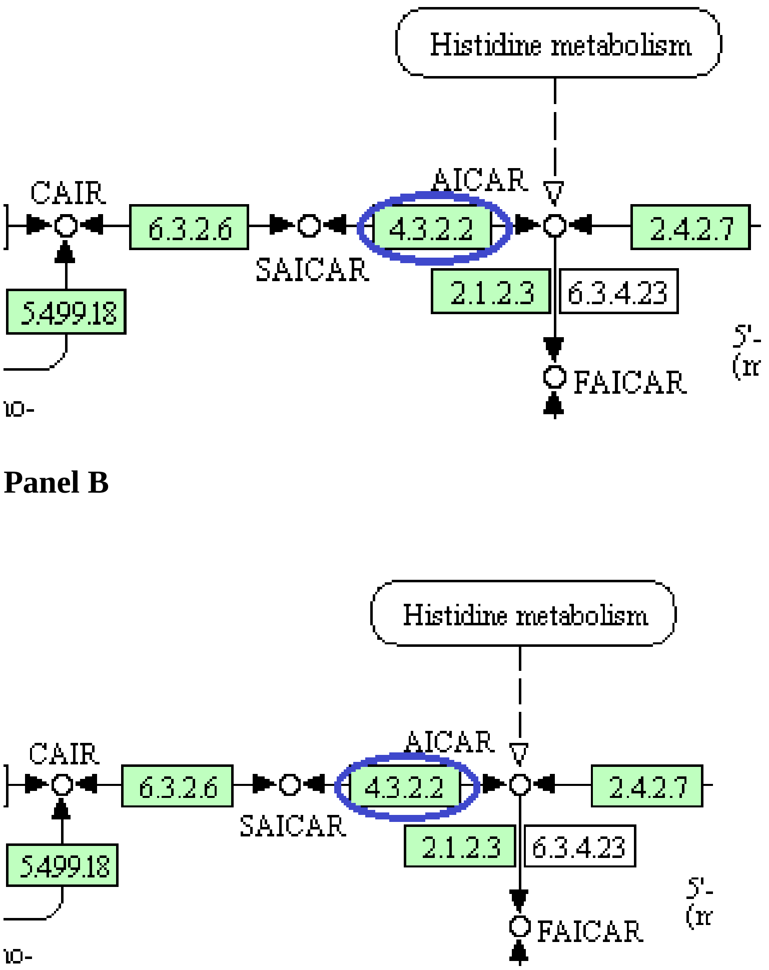


Figure 12. Cleavage sites are not present in *E. coli* b1131 and Mrub_2293. *E. coli* b1131 has a D value (D=0.131) and Mrub_2293 has a D value (D=0.110). Both values were below the cut off value. Panel A shows the plot for *E. coli* b1131; Panel B shows the plot for Mrub_2293. Signal P server v 4.1 http://www.cbs.dtu.dk/services/SignalP created these plots.

Figure 13 displays the purine metabolism biochemical pathway that both genes are apart of. The green boxes represent enzymes that are present in the pathway. In the pathway it shows how both genes are involved in the conversion of SAICAR to AICAR. This provides further evidence that these two genes are orthologs of one another.



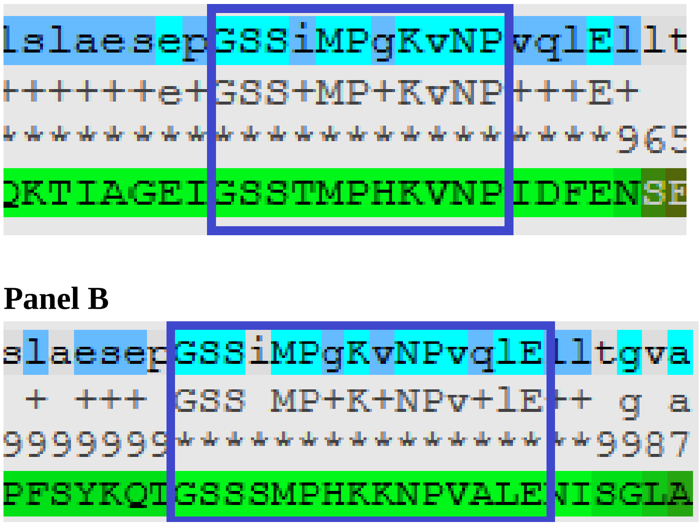
Panel A

Panel B

Figure 13. *E. coli* b1131 and Mrub_2293 are in the same biological pathway. Panel A shows the KEGG pathway for *E. coli* b1131, while Panel B shows the pathway for Mrub_2293. In order to find the purine metabolism pathway The Kyoto Encyclopedia of Genes and Genomes (KEGG) database at http://www.genome.jp/kegg/pathway.html was used.

The below figure shows which amino acids are highly conserved. Both pairwise alignments have serine, methionine, proline, lysine, asparagine, glutamate and glycine as shared conserved amino acids. This alignment is different from the BLAST search in that it compares a query sequence to a consensus sequence, which is made up of hundreds of proteins. This further supports our hypothesis that the two genes are orthologs because they both created similar pairwise alignments to one another.

Panel A



```
lslaesepGSSiMPgKvNPvqlEllt
++++++e+GSS+MP+KvNP+++E+
***********************965
QKTIAGEIGSSTMPHKVNPIDFENSE
```

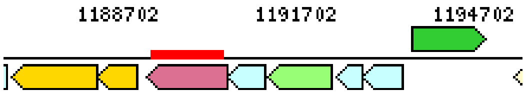
Panel B

```
slaesepGSSiMPgKvNPvqlEllTgva
 + +++ GSS MP+K+NPv+lE++ g a
9999999*****************9987
PFSYKQTGSSSMPHKKNPVALENISGLA
```

Figure 14. Represents the highly conserved amino acids in the *E. coli* b1131 and Mrub_2293 alignments. Both have similar conserved amino acids to one another and are identified in the above figure. Also, they have the same Pfam name (Lyase). The pairwise alignment was created by the Pfam website http://pfam.sanger.ac.uk/search.

Figure 14 represents the ortholog neighborhoods that are seen from color by KEGG. The genes surrounding the query gene in both Panel A and Panel B are not all the same color, which means that they are not made up of an operon. Although the two query genes are not the same exact color, they still provide evidence that they are orthologs because they are both a shade of purple. This indicates they share a similar function of nucleotide transport and metabolism (Markowitz *et. al*., 2012).

Panel A



Panel B

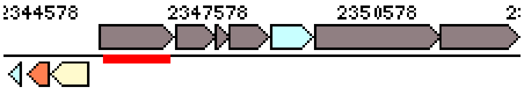
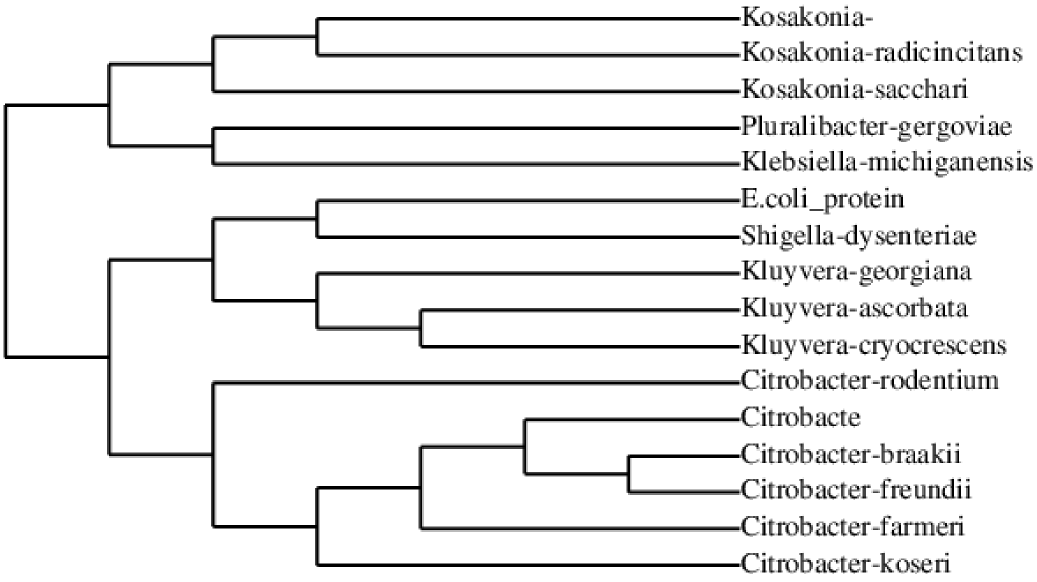


Figure 14. *E. coli* b1131 and Mrub_2293 are not apart of an operon. The gene being viewed has a red bar above or below it. Panel A is *E. coli* b1131 and Panel B is Mrub_2293. Color by KEGG was used for this figure. Image was taken from http://img.jgi.doe.gov/.

Figure 15 shows the traditional phylogenetic trees for *E. coli* and *M. ruber*. Panel A has genus that are all apart of the proteobacteria phylum, while Panel B has genus made up of either Deinococcus-thermus or Firmicutes. Since they are all of similar phylum, then HGT is not present in both genes.

Panel A



Panel B

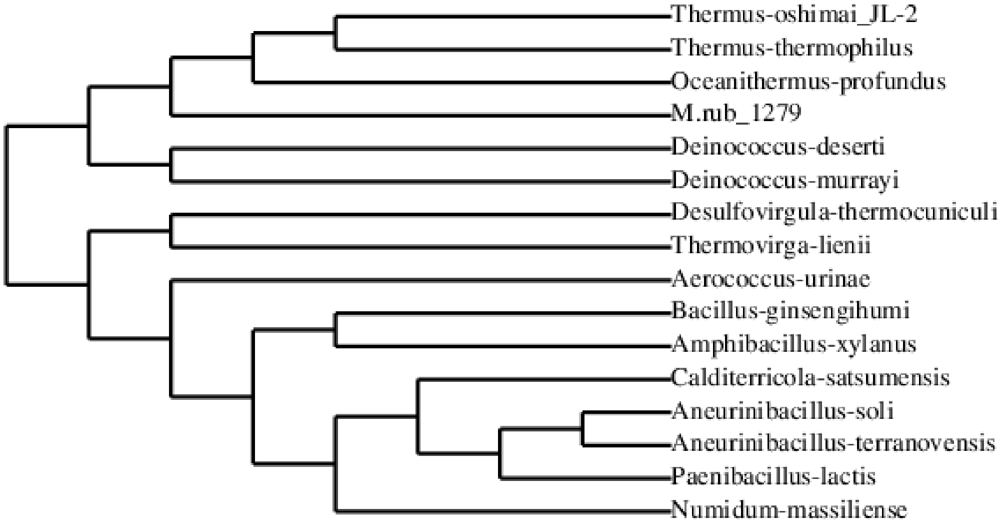


Figure 15. Both Panel A and Panel B are not likely to be HGT. Panel A represents *E. coli* b1131 and Panel B represents Mrub_2293. Both phylogenetic trees were created by http://www.phylogeny.fr.

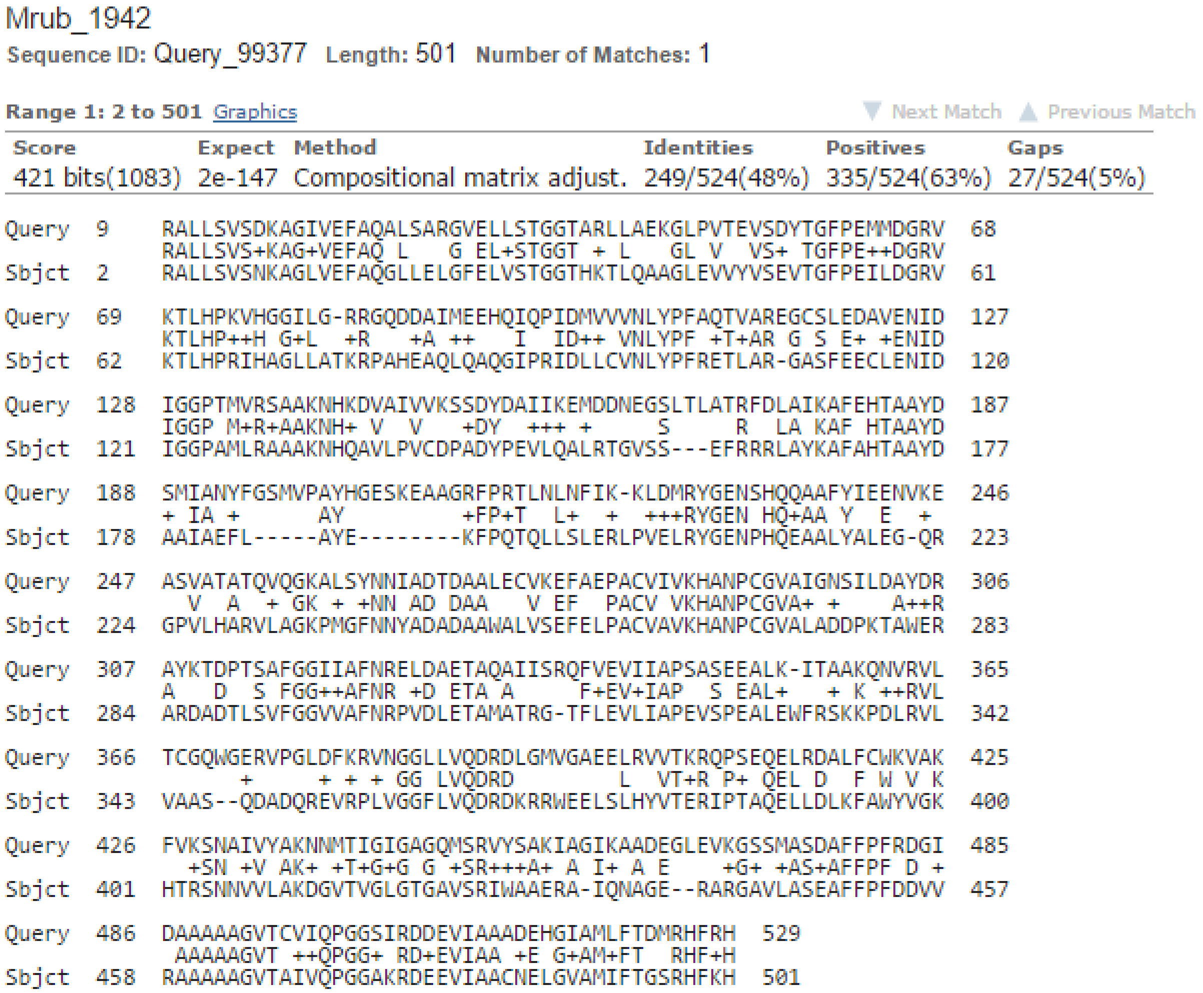
Table 3. *E. coli* b4006 and Mrub_1942 are orthologs

Bioinformatics Tools Used	***E. coli* b4006 gene (PurH)**	**Mrub_1942 gene**
BLAST *E. coli* against *M. ruber*	Score: 421 E-value: 2e-147	
CDD Data (COG category)	COG NUMBER: COG0138 AICAR transformylase/IMP cyclohydrolase PurH [Nucleotide transport and metabolism] E-value: 0.0	E-value: 0.0
Cellular Localization	Cytoplasm of the cell	
TIGRfam-protein family	TIGR00355 PurH: phosphoribosylaminoimidazolecarboxami E-value: 1.2e-283	E-value: 1.4e-172

Pfam-protein family	PF02142 MGS-like domain E-value: 6.9e-32 E-value: 1.7e-25
Protein Database (PDB)	3ZZM Crystal structure of Mycobacterium tuberculosis PurH with a novel bound nucleotide CFAIR, at 2.2 A resolution E-value: 1.012e-120 E-value: 5.36e-113
Enzyme Commission Number	EC 2.1.2.3 phosphoribosylaminoimidazolecarboxamide formyltransferase
KEGG Pathway Map	Purine Metabolism, 00230

Table 3 summarizes the bioinformatics tools that we used to compare the *E. coli* b4006 gene to the Mrub_1942 gene. The first row of the table shows the initial BLAST done on the genes that was stated earlier in the introduction. The high bit score is not a problem because the sequences vary in length. The E-value score of 2e-147 is close to zero, which shows that the results are significant and that it is unlikely that the proteins aligned due to chance alone. The low value shows similarity of amino acids between the genes as well. The results provide further evidence that the two genes are orthologs of one another and share the same ancestor. The CDD showed similar results as well with matching COG numbers (COG0138) and names (AICAR transformylase/IMP cyclohydrolase PurH [Nucleotide transport and metabolism]). Both genes were located in the cytoplasm. Also, the TIGRfam-protein family showed similarity with identical numbers and names: TIGR00355, purH: phosphoribosylaminoimidazolecarboxami. The PDB showed similar results as well: 3ZZM, Crystal structure of Mycobacterium tuberculosis PurH with a novel bound nucleotide CFAIR, at 2.2 A resolution. The enzyme commission number and name were identical and as follows: 2.1.2.3, phosphoribosylaminoimidazolecarboxamide formyltransferase. Finally, both genes were predicted to be apart of the same step for the Purine metabolism pathway. In conclusion, the results show that the genes are orthologs to one another and share very similar characteristics.

Figure 16 explains the BLAST alignment results of *E. coli* b4006 versus Mrub_1942. From the results shown 48% of the amino acids were identical, while 335 were similar. The E-value is 2e-147, which is fairly small and closer to zero. This shows that the amino acid sequence is not aligned due to random chance alone and shows structural similarity between the two genes. This is the first piece of evidence that supports the hypothesis that *E. coli* b4006 and Mrub_1942 are orthologs to one another.



Mrub_1942

Sequence ID: Query_99377 **Length:** 501 **Number of Matches:** 1

Range 1: 2 to 501 Graphics ▼ Next Match ▲ Previous Match

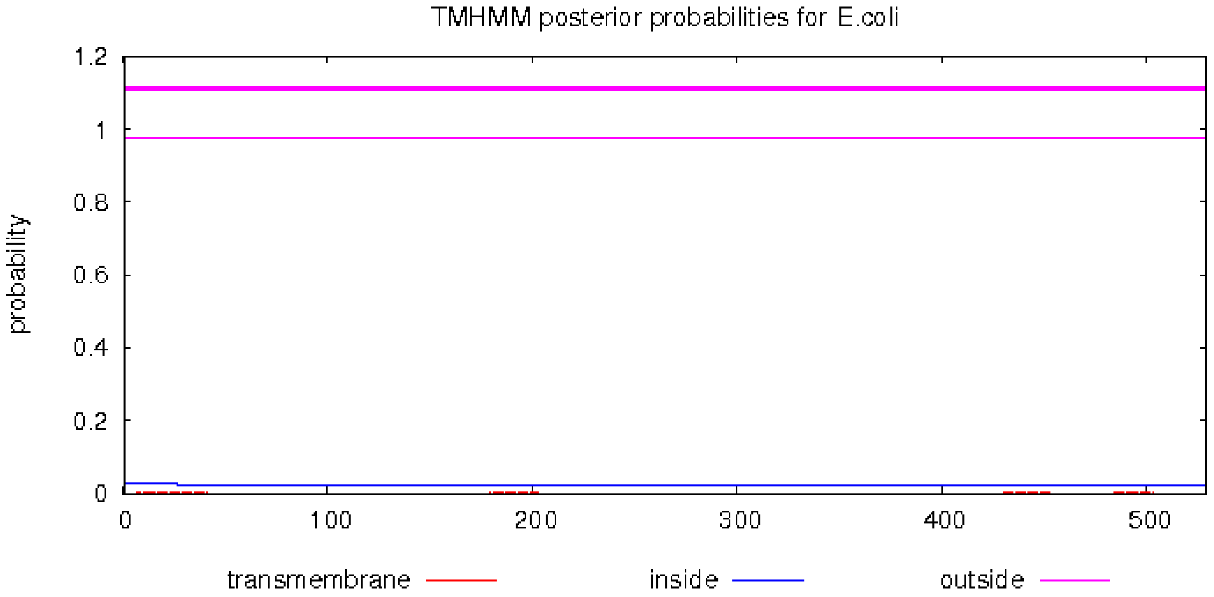
Score	Expect	Method	Identities	Positives	Gaps
421 bits(1083)	2e-147	Compositional matrix adjust.	249/524(48%)	335/524(63%)	27/524(5%)

```
Query  9     RALLSVSDKAGIVEFAQALSARGVELLSTGGTARLLAEKGLPVTEVSDYTGFPEMMDGRV  68
             RALLSVS+KAG+VEFAQ L   G EL+STGGT + L   GL V  VS+ TGFPE++DGRV
Sbjct  2     RALLSVSNKAGLVEFAQGLLELGFELVSTGGTHKTLQAAGLEVVYVSEVTGFPEILDGRV  61

Query  69    KTLHPKVHGGILG-RRGQDDAIMEEHQIQPIDMVVVNLYPFAQTVAREGCSLEDAVENID  127
             KTLHP++H G+L  +R    +A ++   I  ID++ VNLYPF +T+AR G S E+ +ENID
Sbjct  62    KTLHPRIHAGLLATKRPAHEAQLQAQGIPRIDLLCVNLYPFRETLAR-GASFEECLENID  120

Query  128   IGGPTMVRSAAKNHKDVAIVVKSSDYDAIIKEMDDNEGSLTLATRFDLAIKAFEHTAAYD  187
             IGGP M+R+AAKNH+ V  V   +DY  +++ +     S     R  LA KAF HTAAYD
Sbjct  121   IGGPAMLRAAAKNHQAVLPVCDPADYPEVLQALRTGVSS---EFRRRLAYKAFAHTAAYD  177

Query  188   SMIANYFGSMVPAYHGESKEAAGRFPRTLNLNFIK-KLDMRYGENSHQQAAFYIEENVKE  246
             + IA +      AY         +FP+T  L+  +  +++RYGEN HQ+AA Y  E  +
Sbjct  178   AAIAEFL-----AYE--------KFPQTQLLSLERLPVELRYGENPHQEAALYALEG-QR  223

Query  247   ASVATATQVQGKALSYNNIADTDAALECVKEFAEPACVIVKHANPCGVAIGNSILDAYDR  306
               V  A  + GK + +NN AD DAA   V EF  PACV VKHANPCGVA+ +    A++R
Sbjct  224   GPVLHARVLAGKPMGFNNYADADAAWALVSEFELPACVAVKHANPCGVALADDPKTAWER  283

Query  307   AYKTDPTSAFGGIIAFNRELDAETAQAIISRQFVEVIIAPSASEEALK-ITAAKQNVRVL  365
             A   D  S FGG++AFNR +D ETA A     F+EV+IAP  S EAL+   + K ++RVL
Sbjct  284   ARDADTLSVFGGVVAFNRPVDLETAMATRG-TFLEVLIAPEVSPEALEWFRSKKPDLRVL  342

Query  366   TCGQWGERVPGLDFKRVNGGLLVQDRDLGMVGAEELRVVTKRQPSEQELRDALFCWKVAK  425
                   +     + + + GG LVQDRD        L  VT+R P+ QEL D  F W V K
Sbjct  343   VAAS--QDADQREVRPLVGGFLVQDRDKRRWEELSLHYVTERIPTAQELLDLKFAWYVGK  400

Query  426   FVKSNAIVYAKNNMTIGIGAGQMSRVYSAKIAGIKAADEGLEVKGSSMASDAFFPFRDGI  485
               +SN +V AK+ +T+G+G G +SR+++A+ A I+ A E    +G+ +AS+AFFPF D +
Sbjct  401   HTRSNNVVLAKDGVTVGLGTGAVSRIWAAERA-IQNAGE--RARGAVLASEAFFPFDDVV  457

Query  486   DAAAAAGVTCVIQPGGSIRDDEVIAAADEHGIAMLFTDMRHFRH  529
              AAAAAGVT ++QPGG+ RD+EVIAA +E G+AM+FT  RHF+H
Sbjct  458   RAAAAAGVTAIVQPGGAKRDEEVIAACNELGVAMIFTGSRHFKH  501
```

Figure 16. Mrub_1942 and *E. coli* b4006 have similar protein sequence. Subject sequence is Mrub_1942 and query sequence is *E. coli* b4006. Analysis was performed by using NCBI BLAST bioinformatics tool at https://blast.ncbi.nlm.nih.gov/Blast.cgi.

Figure 17 describes the cellular location for both genes. Each panel shows the TMH hydropathy plot for *E. coli* b4006 or Mrub_1942. The presence of red peaks on the plot represent transmembrane helices, however, there are no peaks on either plot meaning there are no transmembrane helices in *E.coli* b4006 and Mrub_1942. Additionally, both plots are consistent with one another proving significance that both genes are located in the cytoplasm opposed to the membrane.

Panel A



Panel B

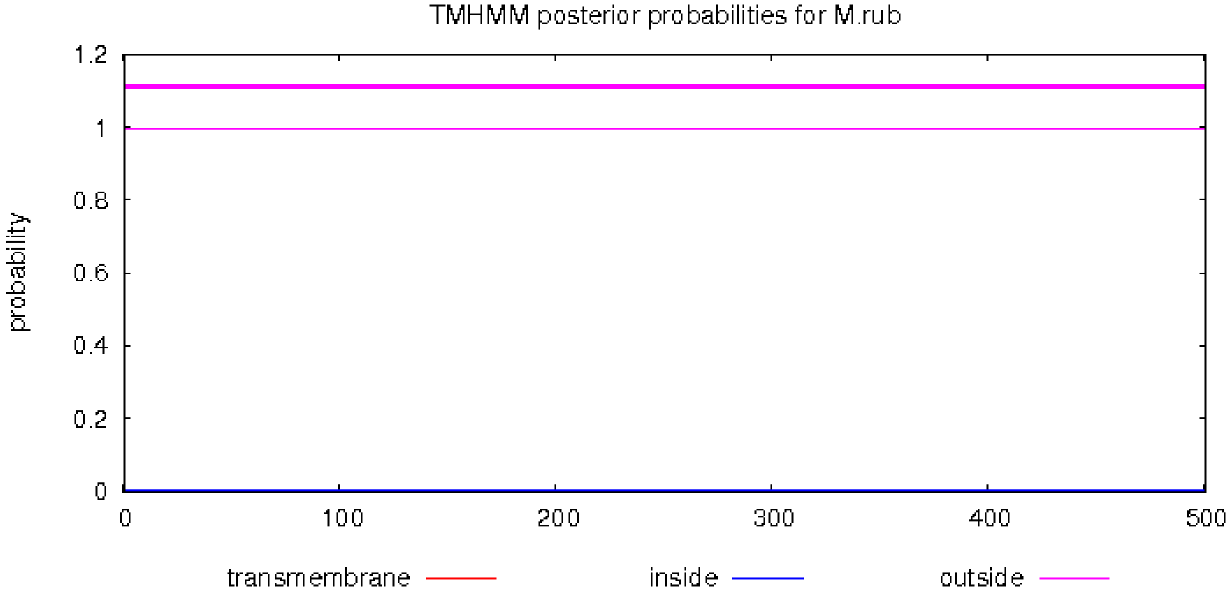
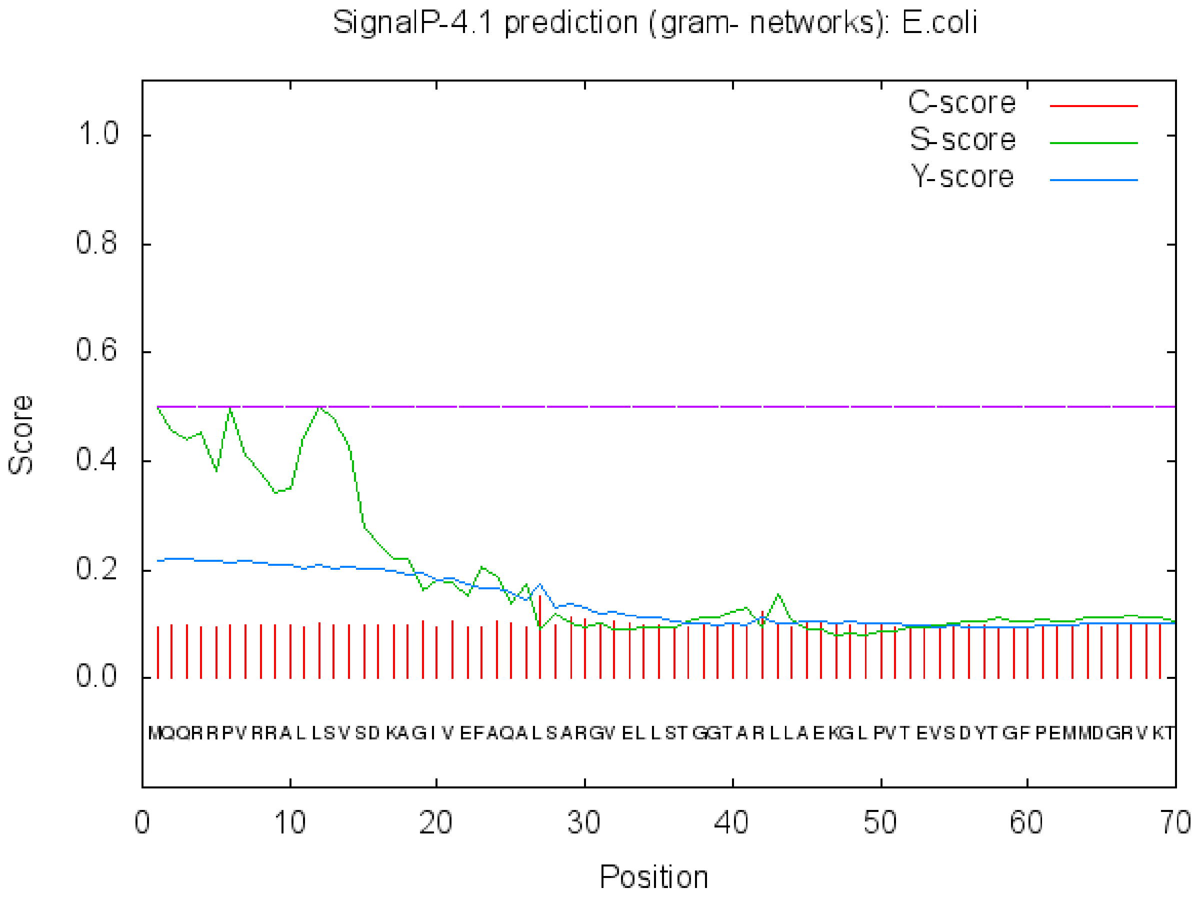


Figure 17. Both *E. coli* b4006 and Mrub_1942 do not contain TMH regions. Both are predicted to be located in the cytoplasmic region. Panel A shows the TMHMM for *E. coli* b4006; Panel B

shows the TMHMM data for Mrub_1942. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was used to create these two hydropathy charts.

Figure 18 shows the SignalP graphs generated by both *E. coli* b4006 and Mrub_1942. The signalP bioinformatics tool is used to predict protein cleavage sites by assigning a D-value to each graph. This D-value is calculated from the S-value and Y-value. Alongside the D-value is a cut off value given by the purple line on the graph. For *E. coli* b4006 in Panel A the D-value of 0.309 is below the cut off value of about 0.500. In Mrub_1942 seen in Panel B, the D-value of 0.182 is below the cut off value of about 0.500. This data supports the fact that neither gene has a protein cleavage site and therefore doesn't attach or cross over the membrane. Also, both genes had start codons that were a correct distance away from the shine-delgarno sequence and were in the correct reading frame. No new reading frames had to be proposed for the JGI/IMG module (Markowitz *et al.*, 2012).

Panel A



Panel B

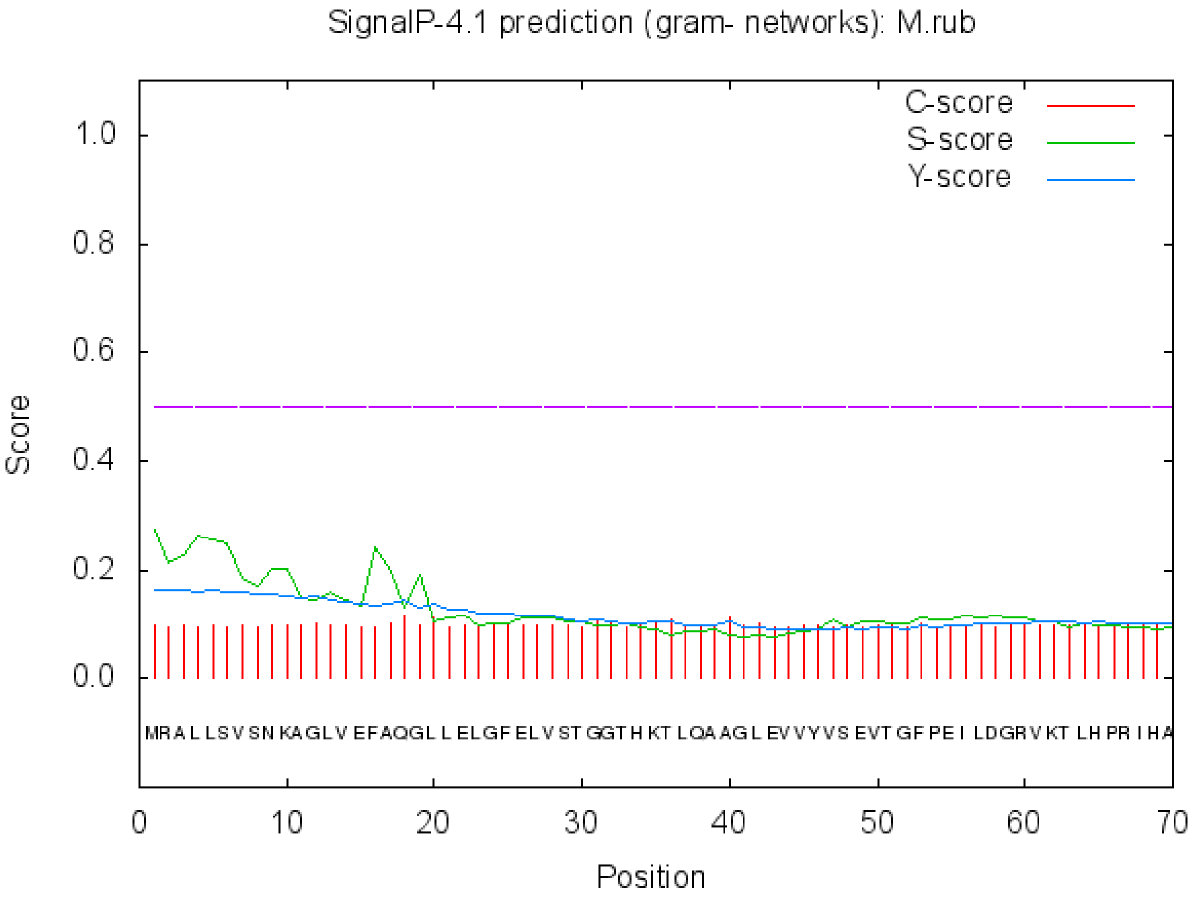
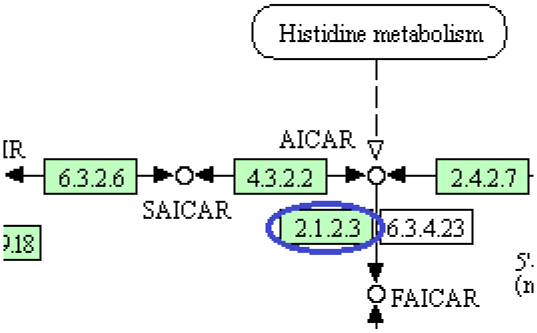


Figure 18. Cleavage sites are not present in *E. coli* b4006 and Mrub_1942. *E. coli* b4006 has a D value (D=0.309) and Mrub_1942 has a D value (D=0.182). Both values were below the cut off value. Panel A shows the plot for *E. coli* b4006; Panel B shows the plot for Mrub_1942. Signal P server v 4.1 http://www.cbs.dtu.dk/services/SignalP created these plots.

Figure 19 displays the purine metabolism biochemical pathway that both genes are apart of. The green boxes represent enzymes that are present in the pathway. In the pathway it shows how both genes are involved in the conversion of AICAR to FAICAR. This provides further evidence that these two genes are orthologs of one another.

Panel A



Panel B

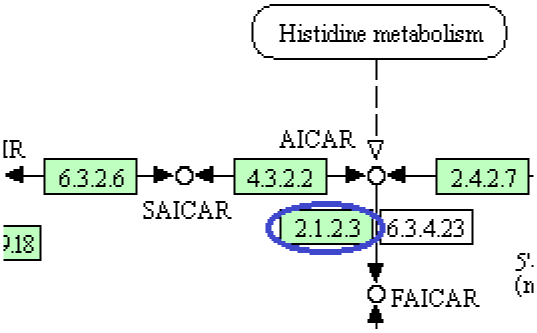
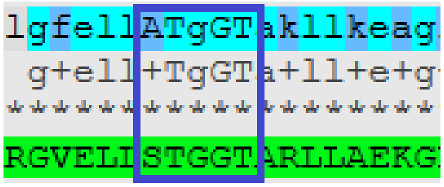


Figure 19. *E. coli* b4006 and Mrub_1942 are in the same biological pathway. Panel A shows the KEGG pathway for *E. coli* b4006, while Panel B shows the pathway for Mrub_1942. In order to find the purine metabolism pathway The Kyoto Encyclopedia of Genes and Genomes (KEGG) database at http://www.genome.jp/kegg/pathway.html was used.

The figure below shows which amino acids are highly conserved. Both pairwise alignments have alanine, threonine, and glycine as shared conserved amino acids. This alignment is different from the BLAST search in that it compares a query sequence to a consensus sequence, which is made up of hundreds of proteins. This further supports our hypothesis that the two genes are orthologs because they both created similar pairwise alignments to one another.

Panel A



```
lgfellATgGTakllkeag
 g+ell+TgGTa+ll+e+g
*******************

RGVELLSTGGTARLLAEKG
```

Panel B

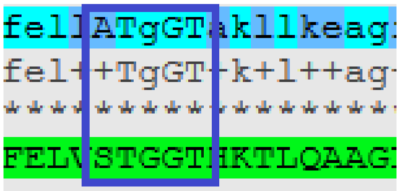
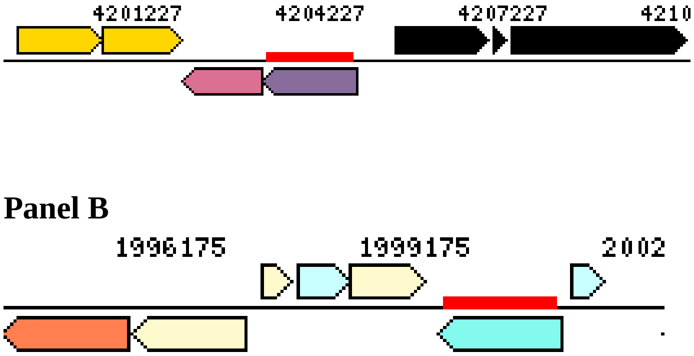


Figure 20. Represents the highly conserved amino acids in the *E. coli* b4006 and Mrub_1942 alignments. Both have similar conserved amino acids to one another and are identified in the above figure. Also, they have the same Pfam name (MGS-like domain). The pairwise alignment was created by the Pfam website http://pfam.sanger.ac.uk/search.

Figure 21 represents the ortholog neighborhoods that are seen from color by KEGG. The genes surrounding the query gene in both Panel A and Panel B are not all the same color, which means that they are not made up of an operon. However, the two query genes are of different colors, but still showed a similar function on the IMG/JGI website. This indicates they share a similar function of nucleotide transport and metabolism (Markowitz *et. al*., 2012).

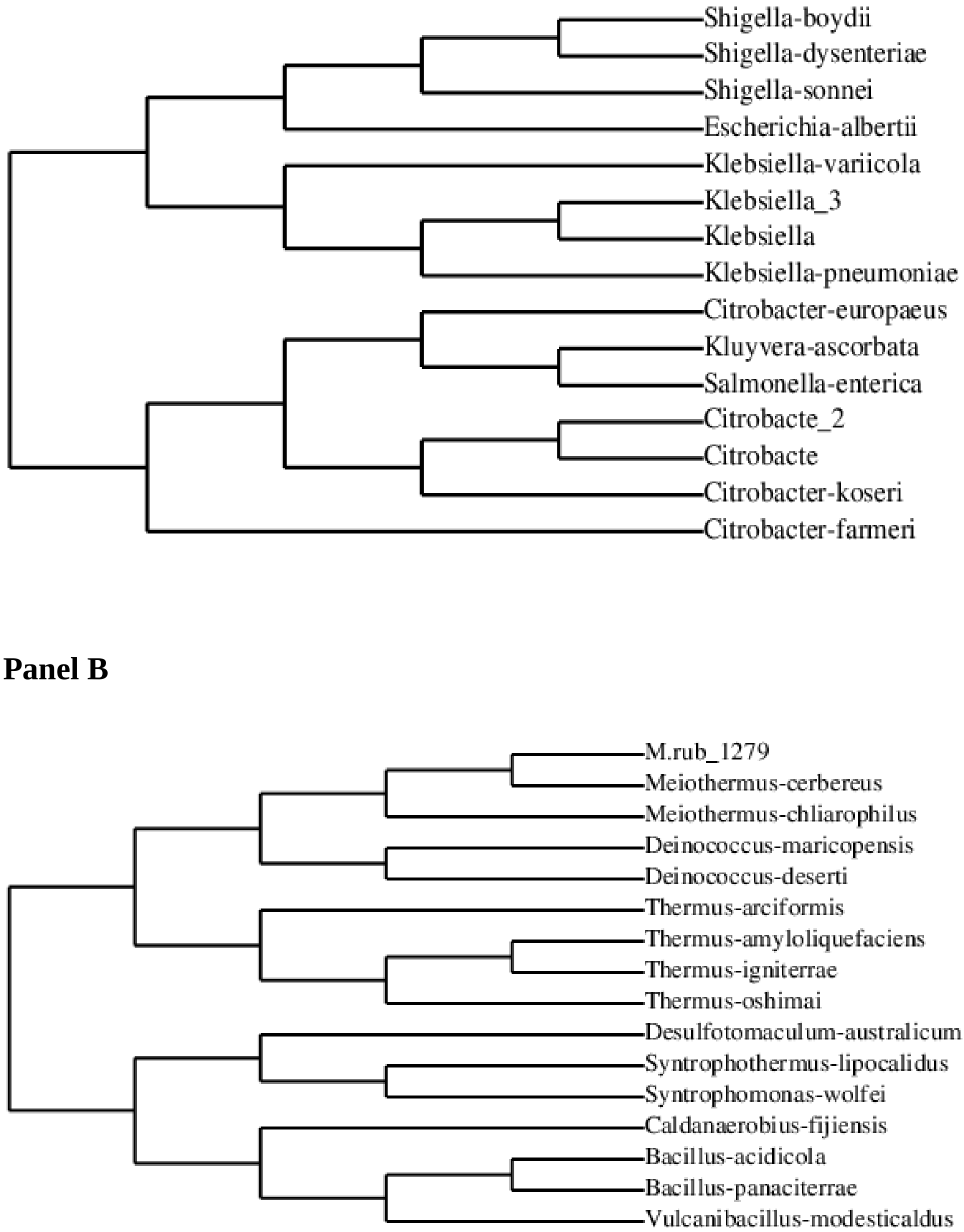
Panel A



Panel B

Figure 21. *E. coli* b4006 and Mrub_1942 are not apart of an operon. The gene being viewed has a red bar above or below it. Panel A is *E. coli* b4006 and Panel B is Mrub_1942. Color by KEGG was used for this figure. Image was taken from http://img.jgi.doe.gov/.

Figure 22 shows the traditional phylogenetic trees for *E. coli* and *M. ruber*. Panel A has genus that are all apart of the proteobacteria phylum, while Panel B has genus made up of either Deinococcus-thermus or Firmicutes. Since they are all of similar phylum, then HGT is not present in both genes.



Panel A

Panel B

Figure 22. Both Panel A and Panel B are not likely to be HGT. Panel A represents *E. coli* b4006 and Panel B represents Mrub_1942. Both phylogenetic trees were created by http://www.phylogeny.fr.

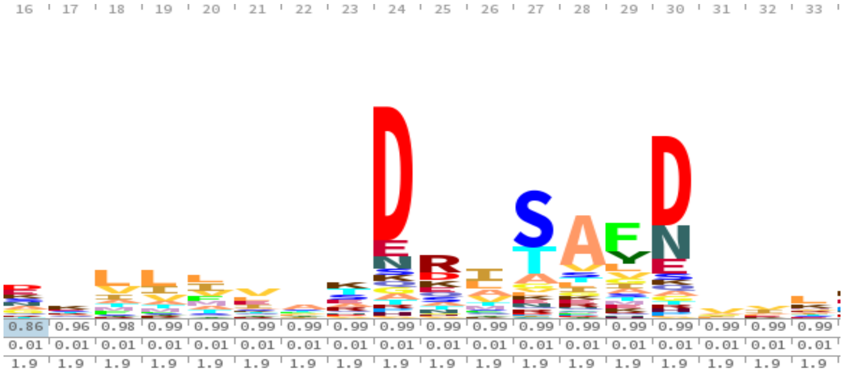
Conclusion

In conclusion, the results of the study were able to determine that the *E.coli* b2476 gene and Mrub_2294 gene are orthologous genes; *E.coli* b1131 and Mrub_2293 are orthologous genes; and *E.coli* b4006 and Mrub_1942 are orthologous genes, which means that the genes share a common ancestor. The first tool to confirm these results was the initial BLAST ran determining that the protein sequences were similar. Cellular location bioinformatics tools further confirmed the results and are as follows: TMH, SignalP, LipoP, Phobius and PSORT-B. Analysis of the TMH led to the conclusion that all six of the proteins lacked transmembrane helices, while SignalP analysis confirmed that the six proteins did not contain cleavage sites. All of the tools indicated that the genes were in the cytoplasm. TIGRfam and Pfam were able to identify similar protein sequences among each gene. The phylogenetic tree for *M. ruber* showed that the genus were either part of the Firmicutes or the Deinococcus-thermus phylum, which are closely related to one another. None of the bioinformatics tools showed significant differences between the genes and all showed similarities. *E. coli* b2476 and Mrub_2294 shared similar sequences in comparison to the Weblogo and Pfam modules. Both genes were located inside the cytoplasm and had no transmembrane helices nor cleavage sites. *E. coli* b2476 and Mrub_2294 had identical COG numbers and TIGRfam numbers and names. Both were located in the same KEGG pathway map, which was purine metabolism. Also, both shared common ancestry in the phylogenetic trees by containing a Firmicutes and Deinococcus-thermus phylum. *E. coli* b1131 and Mrub_2293 shared similar sequences in comparison to the Weblogo and Pfam modules. Both genes were located inside the cytoplasm and had no transmembrane helices nor cleavage sites. *E. coli* b1131 and Mrub_2293 had identical COG numbers and TIGRfam numbers and names. Both were located in the same KEGG pathway map, which was purine metabolism. Also, both shared common ancestry in the phylogenetic trees by containing a Firmicutes and Deinococcus-thermus phylum. *E. coli* b4006 and Mrub_1942 shared similar sequences in comparison to the Weblogo and Pfam modules. Both genes were located inside the cytoplasm and had no transmembrane helices nor cleavage sites. *E. coli* b4006 and Mrub_1942 had identical COG numbers and TIGRfam numbers and names. Both were located in the same KEGG pathway map, which was purine metabolism. Also, both shared common ancestry in the phylogenetic trees by containing a Firmicutes and Deinococcus-thermus phylum.Based upon the abundance of similarity found in the bioinformatics tools, we can conclude that *E.coli* b2476 and Mrub_2294 are orthologous genes; *E.coli* b1131 and Mrub_2293 are orthologous genes; *E.coli* b4006 and *Mrub_1942* are orthologous genes.

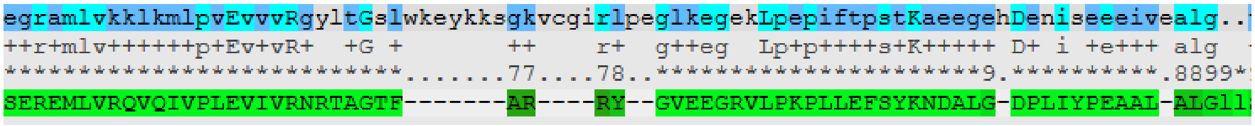
There are several different logical ways to change a DNA sequence (missense, nonsense, and substitution). Site-directed mutagenesis (SDM) is a method that is used to make DNA alterations. One may change the DNA sequence in order to study the protein activity or screen for specific mutations (http://nebasechanger.neb.com/).

The image below in Figure 23 portrayed an example of site directed mutagenesis involving proline applied to Mrub_2294. The missense mutation occurs at position 25 and 27 by chancing a CTC nucleotide sequence to a CCC. Panel A shows the HMM logo for Mrub_2294, with the tallest letters being the most conserved. Leucine is located at position 33 and is moderately conserved compared to the amino acids surrounding it. Panel B portrays the amino acid sequence and has a highly conserved leucine around position 33. This shows that leucine is a highly conserved amino acid for *E. coli* and *M. ruber*. Panel C shows the primers that would be needed to make this mutation in the DNA in a lab and shows the missense mutation. The website http://nebasechanger.neb.com/ was used to create this SDM mutation.

Panel A

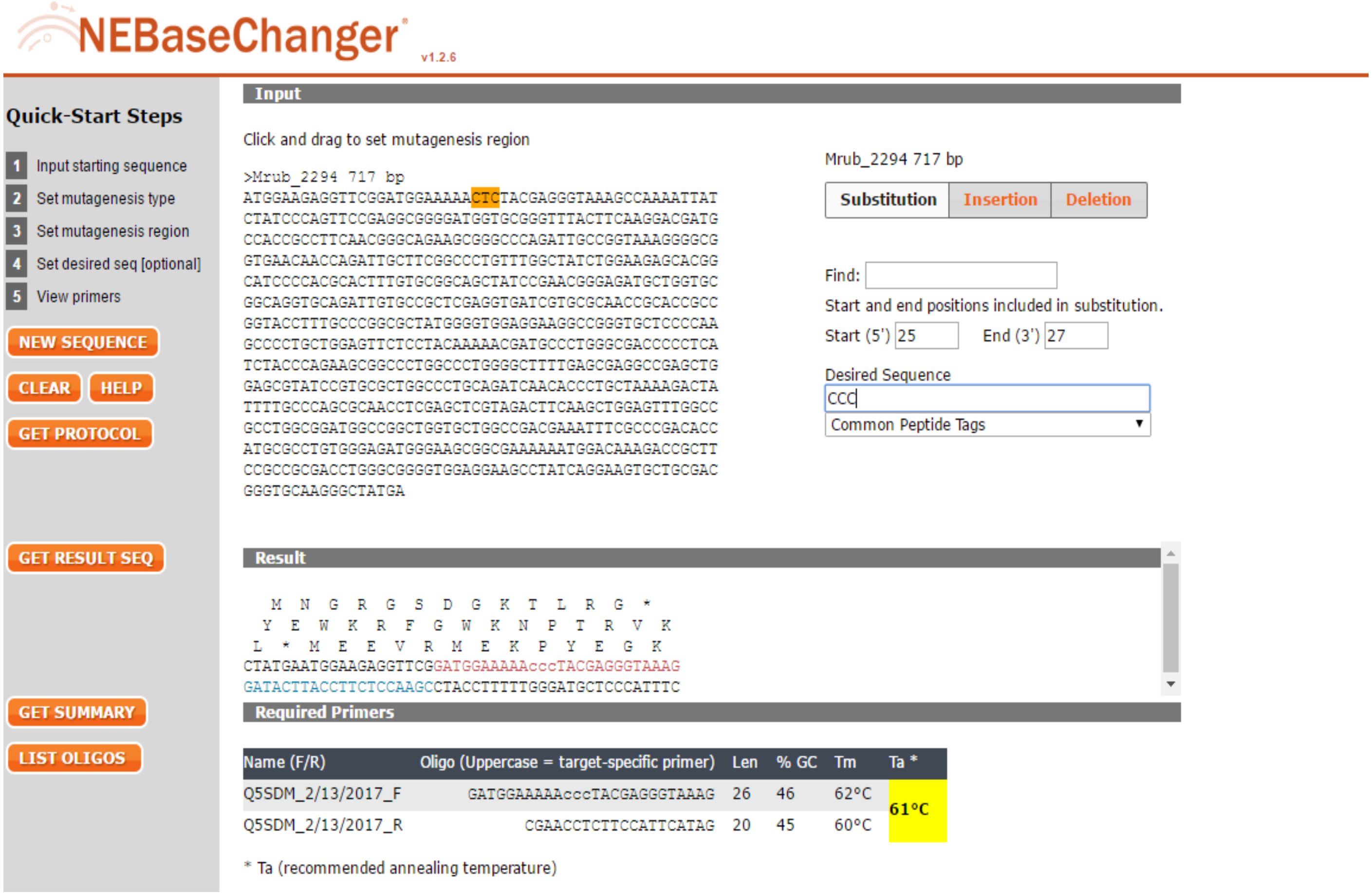


Panel B



```
egramlvkklkmlpvEvvvRgyltGslwkeykksgkvcgirlpeglkegekLpepiftpstKaeegehDeniseeeivealg..
++r+mlv++++++p+Ev+vR+  +G +       ++    r+  g++eg  Lp+p++++s+K+++++ D+ i +e+++ alg
***************************.......77....78..**********************9.**********.8899*
SEREMLVRQVQIVPLEVIVRNRTAGTF-------AR----RY--GVEEGRVLPKPLLEFSYKNDALG-DPLIYPEAAL-ALGll
```

Panel C



Quick-Start Steps

1 Input starting sequence
2 Set mutagenesis type
3 Set mutagenesis region
4 Set desired seq [optional]
5 View primers

Input

Click and drag to set mutagenesis region

```
>Mrub_2294 717 bp
ATGGAAGAGGTTCGGATGGAAAAACTCTACGAGGGTAAAGCCAAAATTAT
CTATCCCAGTTCCGAGGCGGGGATGGTGCGGGTTTACTTCAAGGACGATG
CCACCGCCTTCAACGGGCAGAAGCGGGCCCAGATTGCCGGTAAAGGGGCG
GTGAACAACCAGATTGCTTCGGCCCTGTTTGGCTATCTGGAAGAGCACGG
CATCCCCACGCACTTTGTGCGGCAGCTATCCGAACGGGAGATGCTGGTGC
GGCAGGTGCAGATTGTGCCGCTCGAGGTGATCGTGCGCAACCGCACCGCC
GGTACCTTTGCCCGGCGCTATGGGGTGGAGGAAGGCCGGGTGCTCCCCAA
GCCCCTGCTGGAGTTCTCCTACAAAAACGATGCCCTGGGCGACCCCCTCA
TCTACCCAGAAGCGGCCCTGGCCCTGGGGCTTTTGAGCGAGGCCGAGCTG
GAGCGTATCCGTGCGCTGGCCCTGCAGATCAACACCCTGCTAAAAGACTA
TTTTGCCCAGCGCAACCTCGAGCTCGTAGACTTCAAGCTGGAGTTTGGCC
GCCTGGCGGATGGCCGGCTGGTGCTGGCCGACGAAATTTCGCCCGACACC
ATGCGCCTGTGGGAGATGGGAAGCGGCGAAAAAATGGACAAAGACCGCTT
CCGCCGCGACCTGGGCGGGGTGGAGGAAGCCTATCAGGAAGTGCTGCGAC
GGGTGCAAGGGCTATGA
```

Mrub_2294 717 bp

Find:

Start and end positions included in substitution.

Start (5') 25 End (3') 27

Desired Sequence

ccc

Result

```
   M   N   G   R   G   S   D   G   K   T   L   R   G   *
  Y   E   W   K   R   F   G   W   K   N   P   T   R   V   K
 L   *   M   E   E   V   R   M   E   K   P   Y   E   G   K
CTATGAATGGAAGAGGTTCGGATGGAAAAAcccTACGAGGGTAAAG
GATACTTACCTTCTCCAAGCCTACCTTTTTGGGATGCTCCCATTTC
```

Required Primers

Name (F/R)	Oligo (Uppercase = target-specific primer)	Len	% GC	Tm	Ta *
Q5SDM_2/13/2017_F	GATGGAAAAAcccTACGAGGGTAAAG	26	46	62°C	61°C
Q5SDM_2/13/2017_R	CGAACCTCTTCCATTCATAG	20	45	60°C	

* Ta (recommended annealing temperature)

Figure 23. Panel C displays creating a missense mutation at locations 25 and 27 to change nucleotide sequence CTC to CCC. This changes Leucine to Proline. The website http://nebasechanger.neb.com/ was used to create this SDM mutation.

Works Cited

A. S. Juncker, H. Willenbrock, G. von Heijne, H. Nielsen, S. Brunak and A. Krogh. Prediction of lipoprotein signal peptides in Gram-negative bacteria.Protein Sci. 12(8):1652-62, 2003; [2016 Dec 6]. Available at: http://www.cbs.dtu.dk/services/LipoP/

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E.. [Internet]. 2000. The Protein Data Bank. [2016 Dec 6]. Available from: http://www.rcsb.org/.

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, 2004; [2016 Dec 6]. Available at: http://weblogo.berkeley.edu/

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A. 2016. The Pfam protein families database: towards a more sustainable future: Nucleic Acids Res., 44:D279-D285; [2016, Dec. 6]. Available from: http://pfam.xfam.org/

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29(1):41-3.

Kall L, Krough A, Sonnhammer E. 2004. A combined transmembrane topology and signal peptide prediction method. Journal of Molecular Biology 338(5):1027-36.

JGI - Phylogenetic Diversity. U.S. Department of Energy Joint Genome Institute. 1997. [2017, Feb 2]. Available from: http://jgi.doe.gov/our-science/science- programs/microbial-genomics/phylogenetic-diversity/

Kanehisa M, Sato Y, Kawashima M, Furumichi M. and Tanabe M. (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res., 44, D457–D462; [2016 Dec 6]. Available from: http://www.genome.jp/kegg/

Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martinez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Schroder, I., Shearer, A., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R.P., Paulsen, I., and Karp, P.D. 2013. EcoCyc: fusing model organism databases with systems biology Nucleic Acids Research 41:D605-612.

Krogh A, Rapacki K. TMHMM Server, v. 2.0. Cbs.dtu.dk. 2016 [accessed 2016 Dec 6]. http://www.cbs.dtu.dk/services/TMHMM/

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

Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097/

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. CDD: NCBI's conserved domain database. Nucleic Acids Res.28(43): D222-2: [2016 Dec 6]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25414356?dopt=AbstractPlus

Markowitz VM, Chen IA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, et al. 2012. IMG: The integrated microbial genomes database and comparative analysis system. Nucleic Acids Research 40(D1):D115-22. Available from: http://nar.oxfordjournals.org/content/40/D1/D115.full

Mueller EJ, Meyer E, Rudolph J, Davisson VJ, Stubbe J. 1994. N5-carboxyaminoimidazole ribonucleotide: evidence for a new intermediate and two new enzymatic activities in the de novo purine biosynthetic pathway of *Escherichia coli*. NCBI's conserved domain database. 1;33(8):2269-78: [2017 Feb 10]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8117684

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302 (1):205-17 Available from: http://www.ebi.ac.uk/Tools/msa/tcoffee/

N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics **26(13):1608-1615**

Senecoff J.F., Meagher, R.B. 1993. Isolating the Arabidopsis thaliana Genes for de Novo Purine Synthesis by Suppression of *Escherichia coli* Mutants (I. 5[prime]-Phosphoribosyl-5-

Aminoimidazole Synthetase). Plant Physiology 102 (2): 387-399.

Scott LR. *Meiothermus ruber* Genome Analysis Project. [Internet]. GENI-ACT; [2015 Dec 16]. Available from:http://geni-science.org/secure/projects/vi

Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne & Henrik Nielsen Discriminating signal peptides from transmembrane regions. *Nature Methods*, 8:785-786, 2011Available from: http://www.cbs.dtu.dk/services/SignalP

Tindall et al. 2010. Complete genome sequence of *Meiothermus ruber* type strain. Stand Genomic Sci 3(1): 26-36.

Zhang Y, Morar M, Ealick SE. Structural Biology of the Purine Biosynthetic Pathway. *Cellular and molecular life sciences : CMLS*. 2008;65(23):3699-3724. doi:10.1007/s00018-008-8295-8.