2017

Annotation of Genes Involved with Biosynthetic Production of Peptidoglycan within *Meiothermus ruber* involving supposed Orthologous Genes: Mrub_0981 and b1069, Mrub_1162 and b063, Mrub_1999 and b0084.

Marckus Simmons
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Annotation of Genes Involved with Biosynthetic Production of Peptidoglycan within Meiothermus ruber involving supposed Orthologous Genes: Mrub_0981 and b1069, Mrub_1162 and b063, Mrub_1999 and b0084.

By: Marckus Simmons and Dr. Lori Scott

Introduction

Peptidoglycan Biosynthesis

The envelope that surrounds of gram-negative bacteria such as *Escherichia coli* is composed of two membranes, the inner cytoplasmic membrane and the outer membrane (Freer et al., 1971) The cell wall is composed of the outer membrane and the underlying peptidoglycan layer. The periplasm is the space between the cell wall and the cytoplasmic membrane, which contains periplasmic enzymes (Neu and Heppel, 1965) and binding proteins (Neu and Heppel, 1965). The functions of the bacterial cell envelope are providing structural integrity and shape to the cell, serving as a selective permeability barrier, and mediating interactions with the extracellular environment. (Ruiz, 2013) One of the best studied bacterial envelopes is that of *Escherichia coli*, which is considered the archetype of the Gram-negative envelope.

Peptidoglycan (PDG) is a substance found within the cell wall that plays a role with molecule transportation and it also plays a protective role by preventing osmotic lysis (Sham, 2014). Microbiologists utilize PDG as a means of bacterial identification in a technique called a Gram stain, for which there are two categories: Gram positive (Gram +) and Gram negative (Gram-). The difference between the two is related to the amount of peptidoglycan present within the cell wall. Gram + cell walls a contain 30-70% of peptidoglycan, while Gram – cell walls comprise about 10% (Schleifer, 1972). The crystal violet stain used in the first step of a Gram stain is easily rinsed away with ethanol in Gram – bacteria. Since Gram + cell walls have more peptidoglycan, they are able to hold the stain better than a Gram – cell wall.
Escherichia coli, the model organism for this project, is a Gram – species and the genes that code for PDG synthesis in the cell wall have been identified (Kanehisa M et al., 2016)
For this experiment, I looked at three specific genes of *E. coli* within the pathway of peptidoglycan biosynthesis pathway. The first gene is identified as *murJ* (locus tag b1069),

which is known to code for lipid II flippase protein (Ruiz et. al., 2008). This protein has been tested and has been known to bring lipid II molecules that are essential for PDG production from the inner side of the membrane end expose it to the outer side (citation if different than the one above). This protein helps transport lipid-anchored precursors that will soon become polymerized and become fully matured peptidoglycan molecules. *MurJ* (b1069), is not a part of an operon (citation if not Ruiz). The next *E. coli* gene is *mrdA* (b1999), which codes for peptidoglycan D, D-transpeptidase. This protein is synonymous with penicillin-binding protein II, which is identified by another synonymous gene name *pbpA*. This protein is essential for maintaining the rod shape of the *E. coli* cell and plays a role with peptidoglycan biosynthesis (Keseler I.M. et. al., 2013). This gene was found to not be a part of an operon. The last gene I chose to annotate was *pbpB* (b0084*)* which codes for penicillin-binding protein III. This gene's function is to allow cell division of the *E. coli* cell (Spratt 1975). This gene is a part of an operon whose overall function has been linked to division of the bacterial cell (Spratt 1975). I chose this system especially because of its importance towards the integrity of the cell. Without the formation of the peptidoglycan matrix, the cell's wall would be much more vulnerable to outside, harmful stimuli. This system can also prove to be of importance because of the research that I conducted with Dr. Scott about how Proline production within *Meiothermus ruber* is a link to the survival ability of the cell when expose to very high temperatures.

The peptidoglycan biosynthesis pathway contains many routes, but the specific route I chose begins with metabolized amino-sugars, that then work their way down 8 more steps until reaching the step that utilizes the b1069 gene and finally the b0635 and b0084 genes (Kanehisa M et al., 2016). Figure 1 shows the focused part of the biosynthetic pathway pulled from the KEGG database.

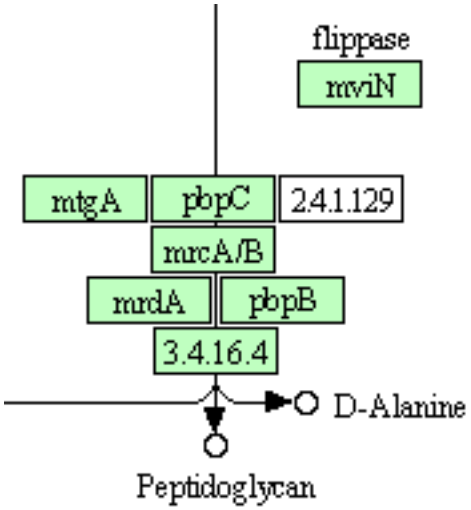


Fig. 1) Peptidoglycan biosynthesis pathway of *E. coli* representing the various genes that are expressed by the bacteria (green boxes *mviN, mrdA,* and *pbpB*). The pathway begins in the left picture and progresses to the right. Images taken from www.kegg.org.

Why study *Meiothermus ruber?*

Meiothermus ruber (*M. ruber)* is a gram negative rod shaped bacteria that is a thermophile and is known to grow in high temperatures. (Tindall et al., 2010) The bacterium was discovered and isolated in the naturally hot environments of Russia, France, China, and Iceland. But it was first isolated in the hot springs found in a city in Russia (Tindall et al., 2010). Dr. Scott, through her years of research on the strain, has formed the hypothesis that *M. ruber* is able to survive in these high tempered areas due their excessive production of proline. But this discovery is not for certain due to the very little amount of research done for this strain of bacteria. However, much research is being done to figure out the metabolic pathways within *M. ruber* are similar to those if *E. coli.*

Hypothesis

The overall goal of this study is to compare three E. coli genes and their encoded proteins known to be involved in peptidoglycan biosynthesis to putative open reading frames in M. ruber. I hypothesize that E. coli genes: b1069, b0635, and b0084 are orthologous to M. ruber genes: Mrub_0981, Mrub_1162, and Mrub_1999, respectively.

Materials and Methods

To carry out this experiment, specific instructional websites and online bioinformatics tools were utilized. The GENI-Science website (https://geni-science.org/) provided instructions on identifying a suitable study question, while the GENI-ACT site (http://www.geni-act.org/). housed the collection of bioinformatics tools used in the study. I used them to annotate the unknown genes of the understudied *Meiothermus ruber* bacterium. We anticipated that the data gathered for the *E. coli* genes, as positive controls, would be similar to the *M. ruber* genes if the our ortholog hypothesis is supported, thereby establishing their similar functional roles. First, we searched for specific biological pathways that both *M. ruber* and *E. coli* cells utilized via the KEGG website (http://www.genome.jp/kegg/pathway.html). Once we collected our three sets of genes, we then NCBI- BLAST (Madden T. et al., 2002) to compare the similarity of amino acids among the two proposed orthologous proteins in *E. coli* and *M. ruber.* Afterwards, we were clear to being annotations of the genes within the GENI-Act bioinformatics kit set up by Dr. Scott (http://www.geni-act.org/). The basic information of the genes were annotated, like the DNA coordinates, amount of nucleotides and amino acids within the sequence, and actual DNA/amino acid sequences. We then used the NCBI-BLAST website to gather a group of 10 organisms based from the blast of a specific gene. These 10 organisms' amino acid sequence were then aligned using the T Coffee analysis (Notredame et al., 2000) to help find a start codon that is conserved through all the species. Once the alignment was made, a WebLogo was created (Crooks GE et al., 2004) giving us a better understanding of whether the start codon is the correct one. Following modules contained the tools that accounted for many different features, for example; To find out where in the cell our proteins are located, we utilized various assays found on the Center for Biological Sequences website, which included Signal P 4.0 (Petersen, 2011), TMHMM 2.0 (Krogh A et al., 2016), and Lipo P (A. S. Juncker et al., 2003). The P-Sort analysis was used to predict protein localizations (Yu et al., 2010), and a Phobius chart was created as an extra amount of evidence for figuring out protein localization (Kall L et al., 2004). To further confirm if the correct start codon was being used , we utilized the Integrated Microbial Genomes & Microbial Samples database (Markowitz, 2012). To determine if there were structural similarities between the putative ortholog proteins (e.g., proteins belonged to the same protein families or possessed the same protein domains), we used like TIGRfam (Haft DH

et al., 2001), Pfam (Finn et al R.D. et al., 2016), and the PDB database (Berman H.M. et al., 2000). We used the KEGG maps (Kanehisa M et al., 2016), and EcoCyc (Keseler I.M. et al., 2013) databases to solidify our understanding of the Enzyme Function. We also used the KEGG pathway to find out if the genes we chose had paralogs, or a separate gene that resembles in function within the same organism. The last step taken, was to determine if horizontal gene transfer was a likely explanation for the evolutionary history of each gene. We created a phylogeny tree (www.phylogeny.fr) that used various species taken from an earlier T Coffee multi-species alignment (Notredame et al., 2000), and we used Integrated Microbial Genomes & Microbial Samples database (Markowitz, 2012) to view orthologous genes of species related to *M.* ruber and *E. coli*. The websites are listed below in correspondence to what module each bioinformatics step was used.

Table 1. Bioinformatics Tools used for each Geni-Act module

Basic Information	Geni-act.org
Sequence-based Similarity Data	www.ncbi.nlm.nih.gov/blast www.ebi.ac.uk/Tools/msa/tcoffee weblogo.berkeley.edu
Cellular Localization Data	www.cbs.dtu.dk/services/TMHMM www.cbs.dtu.dk/services/SignalP www.cbs.dtu.dk/services/LipoP/ www.psort.org/psortb phobius.sbc.su.se
Alternative Reading Frame	img.jgi.doe.gov/cgi-bin/edu/main.cgi
Structure-based Evidence	tigrblast.tigr.org/web-hmm pfam.sanger.ac.uk/search www.rcsb.org/pdb/home/home.do
Enzymatic Function	www.genome.jp/kegg/pathway.html https://ecocyc.org
Duplication and Degradation	www.genome.jp/kegg/pathway.html

Horizontal Gene Transfer	www.phylogeny.fr img.jgi.doe.gov/cgi-bin/edu/main.cgi

Results

Table 2 summarizes the results comparing Mrub_0981 and *E. coli* b1069. A protein BLAST alignment produced an E-value close to 0.0, which indicates strong amino acid sequence similarity that likely reflects a common function. Searching the CDD with the two query sequences identified strong amino acid sequence similarity (*i.e.,* E-values less than the 0.001 cutoff) to the same protein domain (COG numberCOG0728, and COG name Lipid II flippase protein); The bioinformatics tools used for the cellular localization module (LipoP, SignalP, TMH, PSORT-B, and Phobius) predict that Mrub0981 is localized to the cytoplasmic membrane, which is the known location for *E. coli* b1069 (Johnson, 2009). Phobius was especially informative; Mrub-0981 is predicted to have 12 transmembrane helices and b1069 has 13. Both query sequences pulled the identical TIGRfam hit TIGR01695 *mviN/murJ* From the database. A search of thePfam database showed the two genes belong to the same superfamily of MviN-like protein (PF03023). When pulling data from the Protein Database, both genes coded for 5T77, but had no alignments available. The Pfam database, and others such as KEGG, tagged both proteins as having the same E.C. number of 3.6.3.1. Mrub_0981 is predicted to catalyze the same reaction as b1069 on the KEGG map 00550. All of this data supports my hypothesis that Mrub_0981 and E. coli b1069 are orthologs. There was no contradictory data collected.

Table 2. Mrub_0981 and b1069 are orthologs

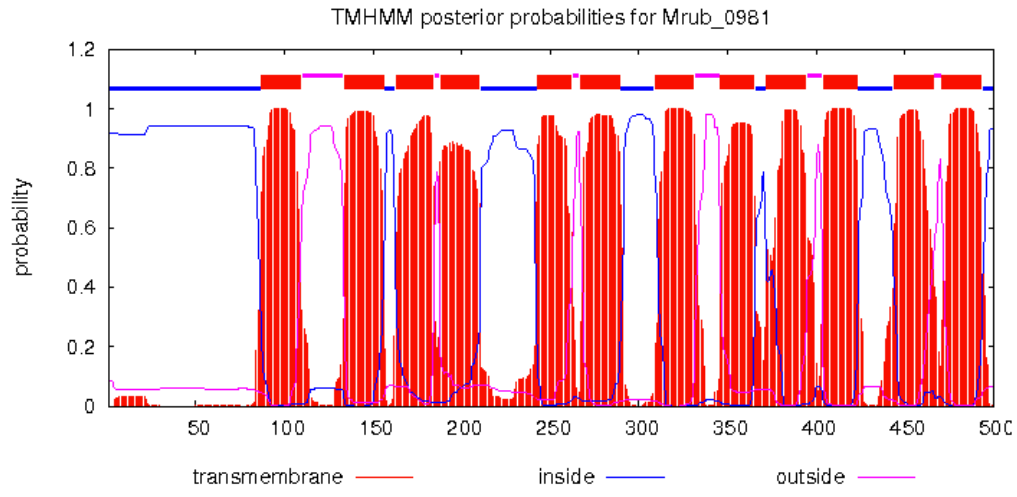
Bioinformatics Tool	**Mrub_0981**	**b1069**
Protein BLAST *E. coli* against *M. ruber*	E-value: 4e-29 Score: 107 bits	
CDD Data (COG category)	COG number: COG0728 Lipid II Flippase protein	

	E-value: 3.86e-55	E-value: 4.34e-180
Cellular Localization	Cytoplasmic Membrane	
TIGRfam- Protein Family	TIGR01695 Lipid II Flippase protein	
	E-value: 1.7e-59	E-value: 1.1e-255
Pfam Protein Family	PF03023 (MviN-like protein)	
	E-value:2.8e-66	E-value:1e-162
Protein Database	5T77: MOP flippase *murJ*	
Enzymatic Commission Number	E.C.3.6.3.1-Flippase	
KEGG Pathway map	Peptidoglycan biosynthesis pathway	

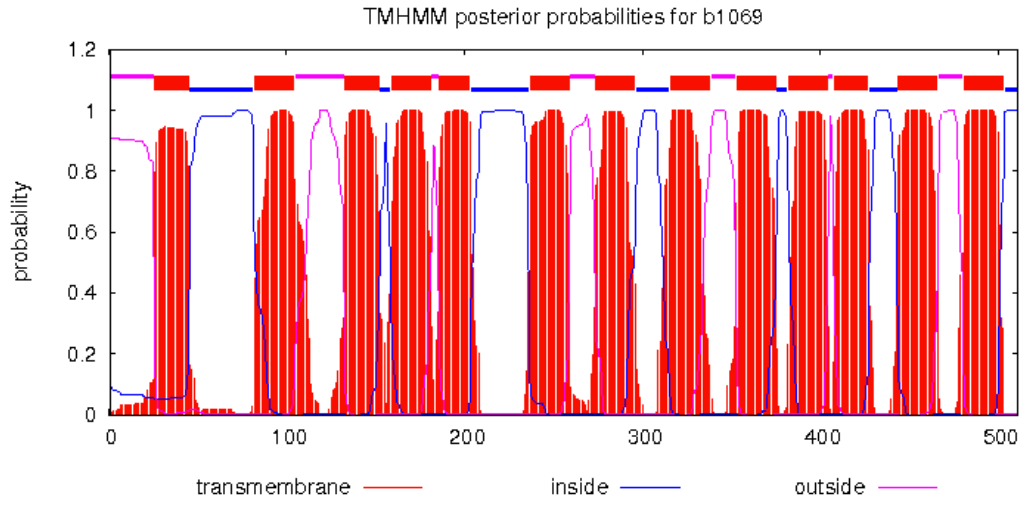
Figure 2 is the the protein BLAST alignment between *Mrub_0981* and *bo1069*, which was originally described in Table 2. There were 26% identical amino acids and 46% of amino acids of similar nature. The E-value of 4e-29 is well below the 0.001 cutoff and indicates a high level of primary sequence similarity and, consequently, functional similarity.

Fig 2) Comparison protein-BLAST search of E. coli b1069 gene against Mrub_0981 gene. Query Sequence: b1069; Subject Sequence: Mrub_0981. Analysis was done via NCBI BLAST website http://www.ncbi.nlm.nih.gov/blast.

Figure 3 represents the TMH hydropathy charts for Mrub_0981 and b1069. The red peaks signify the presence of multiple transmembrane helices for both proteins because they are matched with the predicted number of 12 TMH. The Mrub_0981 is predicted to contain 12 TMH while b1069 has 13 TMH. The presence of this many TMH likely indicates a cellular location within the cytoplasmic membrane.



Panel A

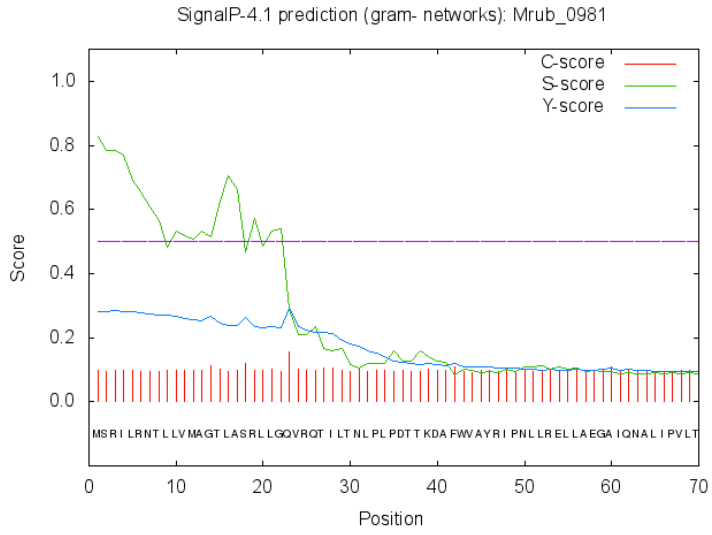


Panel B

Fig. 3) Panel shows the Transmembrane Helices (TMH) found in Mrub0981 (panel A) and b1069 (panel B). In both proteins, there are signs of TMH domains. These hydropathy charts were pulled up from TMHMM Server v 2.0 www.cbs.dtu.dk/services/TMHMM.

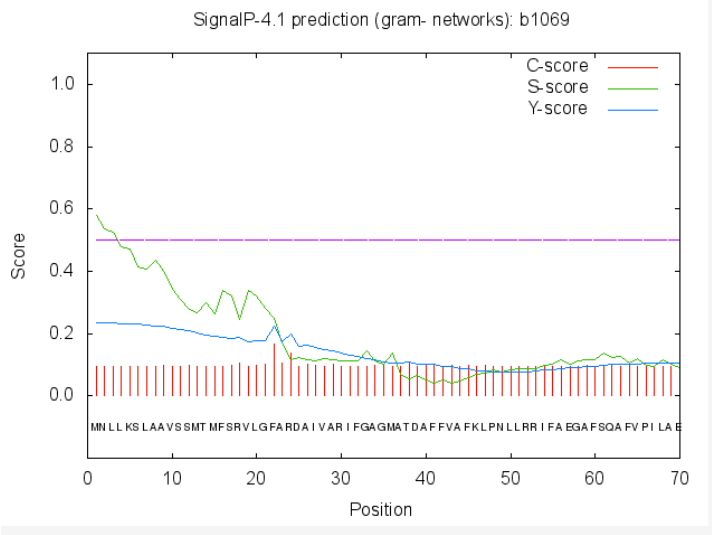
The plots in Fig.4 represent the SignalP results for Mrub_0981 and b1069. The SignalP bioinformatics tool is necessary for finding cleavage sites within a protein; and it does that giving a D value for each protein. That D value is calculated using an S Value, Y value, and a Cutoff value that is represented by the purple line on the chart. The cutoff value given was .570 for the Mrub_0981 chart with a D value of .440, not meeting the needed standard. And the cutoff value for b1069 being .510 with a D value of .279, again not meeting the standard. This final result in the data suggests that there are no cleavages sites found within the proteins of these two organisms.

Panel A




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    23        0.154
  max. Y    23        0.292
  max. S     1        0.827
  mean S     1-22     0.607
       D     1-22     0.440   0.570    NO
Name=Mrub_0981  SP='NO' D=0.440 D-cutoff=0.570 Networks=SignalP-noTM
```

Panel B




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    22        0.165
  max. Y    22        0.223
  max. S     1        0.580
  mean S     1-21     0.374
       D     1-21     0.279   0.510    NO
Name=b1069      SP='NO' D=0.279 D-cutoff=0.510 Networks=SignalP-TM
```

Fig. 4) Mrub_0981 and b1069 do not contain cleavage sites; the D values for both organisms did not meet the cutoff values of .510. These charts were pulled from the SignalP server v 4.1 database. www.cbs.dtu.dk/services/SignalP

Figure 5 shows the metabolic pathway of peptidoglycan biosynthesis and it also shows the two genes (Mrub_0981, b1069*)* being expressed at identical spots of the pathway. Both of the genes are found to code for the same lipid II flippase protein. This gives a good indication that the genes area evolutionarily related and orthologous based on function.

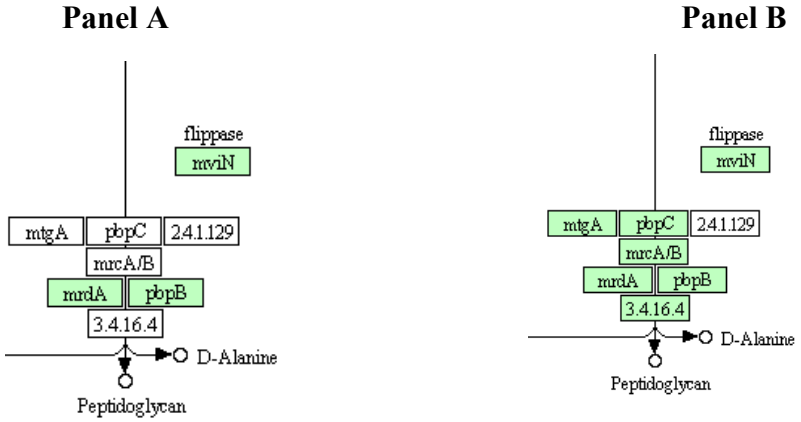
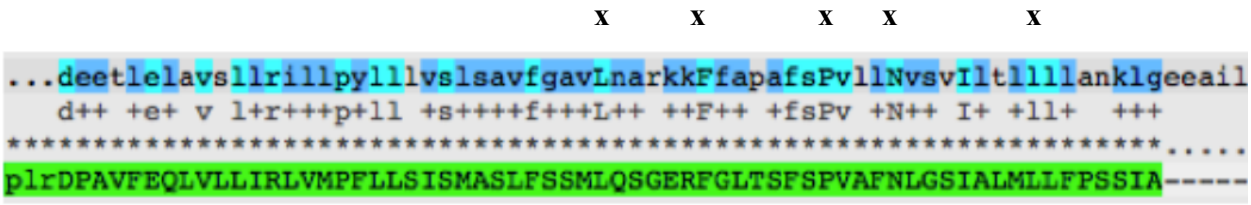


Fig. 5) Mrub_0981 and b1069 are both expressed in the same metabolic pathway. Panel A is the KEGG pathway when looking at *Meiothermus ruber (mviN).* Panel B is the KEGG pathway when looking at *Escherichia coli (mviN).* This chart has pulled from the KEGG website www.genome.jp/kegg/pathway.html.

Figure 6 is the HMM alignment that shows the conservation of amino acids within the proteins of Mrub_0981 and b1069. These alignments contain many conserved amino acids that are identical between the two species. The fact that these two genes pulled up so many consensus amino acids, reinforces the notion that the genes are similar in structure and function.

Panel A



Panel B

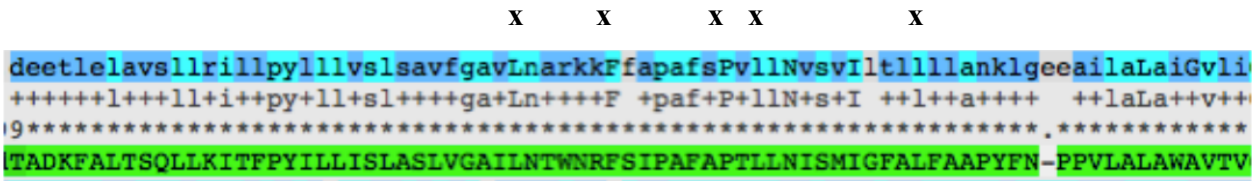
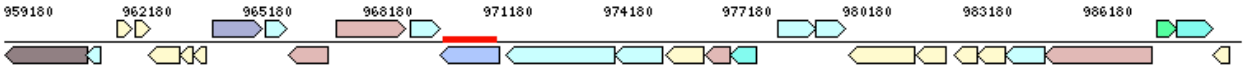


Fig. 6) Mrub_0981 and b1069 have identical highly conserved amino acids (marked by x) within their proteins in the same domain of MviN-like protein. Panel A represents Mrub_0981 and panel B is b1069. These charts were made using the Pfam website pfam.sanger.ac.uk/search.

Figure 7 shows the gene neighborhood that flank Mrub_0981 and b1069. The red line identifies the query sequence. The color of a gene is unique to its predicted KEGG pathway; this illustration is called "Color by KEGG" and is derived from the IMG Gene Details pages (Markowitz, 2012). If the query was part of an operon, the flanking genes would have the same KEGG color. Our query genes appear to be isolated genes and not part of an operon. When compared to each other, there is a slight difference in color. This does not indicate a different KEGG map, however. On further investigation, both genes were identified coding *murJ, mviN* putative peptidoglycan lipid II flippases

Panel A



Panel B

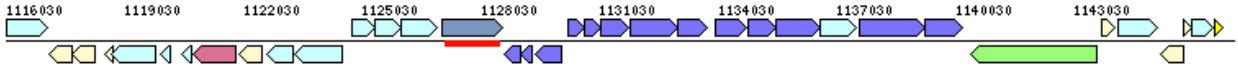


Fig. 7) Mrub_0981 and b1069 genes are not a part of an operon. Chromosome Viewer was colored via their KEGG database. Panel A: Mrub_0981 Chromosome Viewer. Panel B: b1069 Chromosome Viewer. Images were obtained from https://img.jgi.doe.gov.

Figure 8 shows a phylogeny tree that is composed of species that chosen from the previous T Coffee alignment. The species *X. sacchari* and *X. tanslucens* both belong to the phylum, Protobacteria. While the rest of the species all belong to Deinococcus Thermus, showing that there is a small chance of horizontal gene transfer.

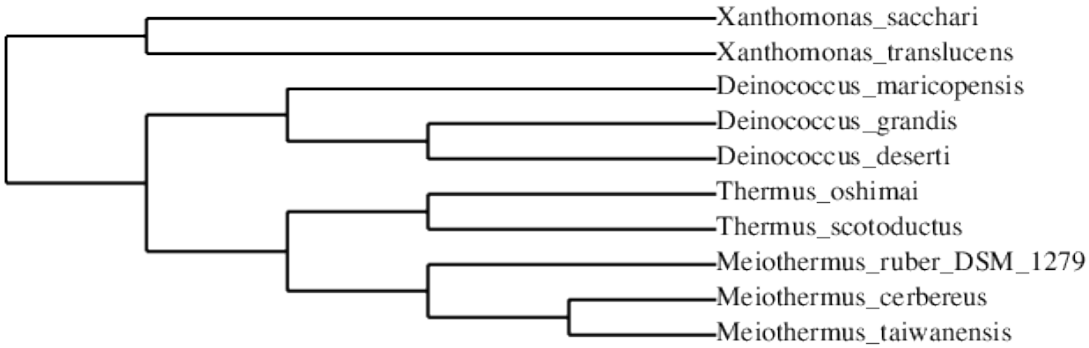


Fig. 8) Phylogeny tree of species that express Mrub_0981 gene. Chart was created under the "A La Carte" setting of Phylogeny.fr at: http://www.phylogeny.fr

Table 3 summarizes the results comparing Mrub_1162 and *E. coli* b0635. A protein BLAST alignment produced an E-value close to 0.0, which indicates strong amino acid sequence similarity that likely reflects a common function. Searching the CDD with the two query

sequences identified strong amino acid sequence similarity (*i.e.,* E-values less than the 0.001 cutoff) to the same protein domain (COG numberCOG0768, and COG name FtsI/penicillin-binding protein 2); The bioinformatics tools used for the cellular localization module (LipoP, SignalP, TMH, PSORT-B, and Phobius) predict that Mrub1162 is localized to the cytoplasmic membrane, which is the known location for *E. coli* b0635 (Johnson, 2009). Phobius was especially informative; Mrub_1162 is predicted to have 1 transmembrane helix and b0635 has 1. Both query sequences pulled the identical TIGRfam hit TIGR03423 *ftsI* from the database. A search of the Pfam database showed the two genes belong to the same two superfamies of PF03717: Penicillin-binding protein dimerization domain and PF00905: Penicillin-binding protein transpeptidase domain. When pulling data from the Protein Database, the Mrub_1162 gene coded for 4MNR: Transpeptidase domain of peptidoglycan glycosyltransferase; while b0635 coded for 5VDY: Dimeric form of Penicillin binding protein 2. The Pfam database, and others such as KEGG, tagged both proteins as having the same E.C. number of 2.4.1.129. Mrub_1162 is predicted to catalyze the same reaction as b0635 on the KEGG map 00550. All of this data supports my hypothesis that Mrub_1162 and E. coli b0635 are orthologs. There was no contradictory data collected.

Table 3. Mrub_1162 and b0635 are orthologs

<table>
<tr><th>Bioinformatics Tool Used</th><th>Mrub_1162</th><th>b0635</th></tr>
<tr><td>Protein BLAST E. coli
against M. ruber</td><td colspan="2">Score: 197 bits
E-value: 1e-59</td></tr>
<tr><td>CDD Data (COG Category)</td><td colspan="2">COG Number: COG0768
ftsI/penicillin-binding protein 2
<table><tr><td>E-value: 6.44e-112</td><td>E-value: 0</td></tr></table></td></tr>
<tr><td>Cellular Localization</td><td colspan="2">Cytoplasmic Membrane</td></tr>
<tr><td>TIGRfam-
protein family</td><td colspan="2">TIGR03423
Penicillin-binding protein II
<table><tr><td>E-value: 1e-132</td><td>E-value: 0</td></tr></table></td></tr>
<tr><td></td><td colspan="2">PF03717: Penicillin-binding protein dimerization domain</td></tr>
</table>

<table>
<tr><td>Pfam-
Protein family</td><td>PF00905: Penicillin-binding protein transpeptidase domain
<table><tr><td>E-value:
1.7e-17
2.8e-47</td><td>E-value:
6.6e-46
1.9e-90</td></tr></table></td></tr>
<tr><td>Protein Database</td><td><table><tr><td>4MNR-Transpeptidase domain of peptidoglycan glycosyltransferase</td><td>5VDY- Dimeric form of Penicillin binding protein 2</td></tr></table>
<table><tr><td>5.2e-43</td><td>3.2e-19</td></tr></table></td></tr>
<tr><td>Enzyme Commission Number</td><td>E.C.- 2.4.1.129
Peptidoglycan glycosyltransferase</td></tr>
<tr><td>KEGG Pathway Map</td><td>Peptidoglycan Biosynthesis</td></tr>
</table>

Figure 2 is the the protein BLAST alignment between Mrub_1162 and b0635, which was originally described in Table 3. There were 29% identical amino acids and 43% of amino acids of similar nature. The E-value of 1e-59 is well below the 0.001 cutoff and indicates a high level of primary sequence similarity and, consequently, functional similarity.

.

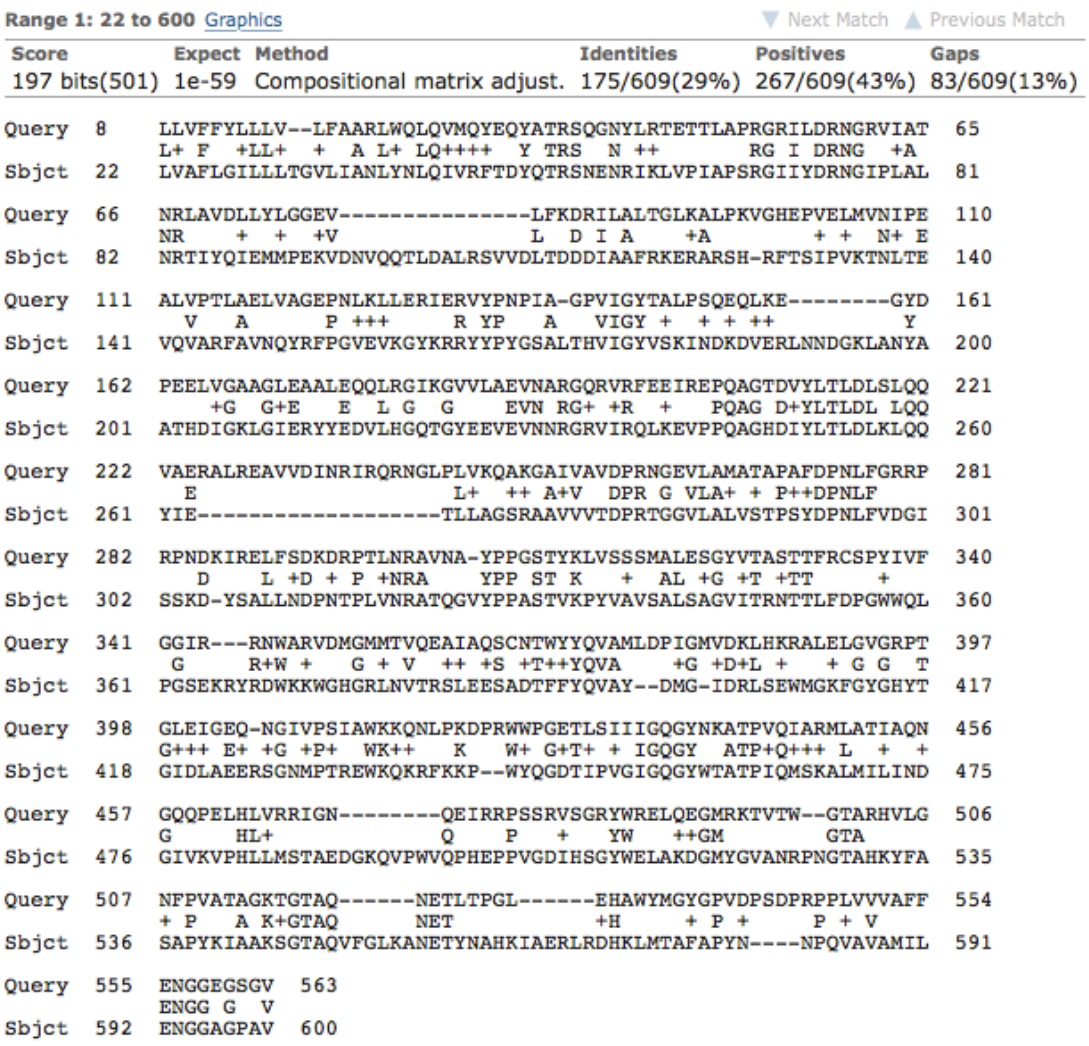
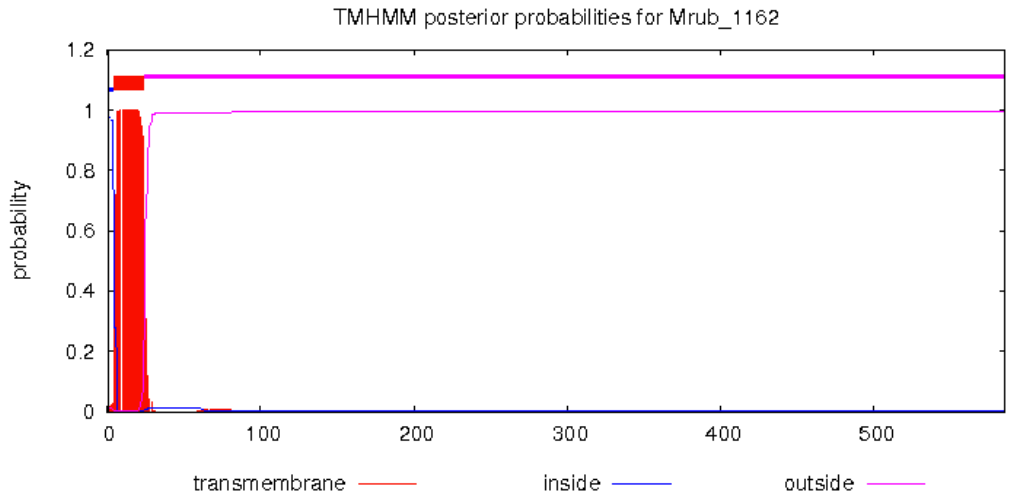
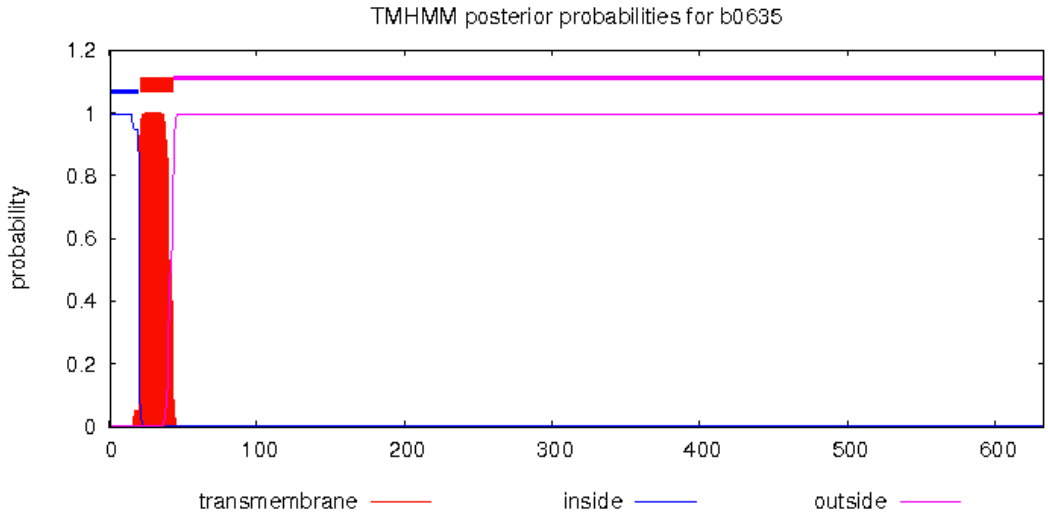


Fig. 2) Protein BLAST of b0635 gene against Mrub_1162 gene showing similar protein sequence. Query sequence: b0635. Subject sequence: Mrub_1162. Alignment was formed using NCBI Protein BLAST bioinformatics tool at http://www.ncbi.nlm.nih.gov.

Figure 3 represents the TMH hydropathy charts for Mrub_1162 and b0635. The red peaks signify the presence of multiple transmembrane helices for both proteins because they are well above the probability cutoff of 1 TMH. The Mrub_1162 is predicted to contain 1 TMH while b0084 has 1 TMH. The presence of this many TMH likely indicates a cellular location within the cytoplasmic membrane.



Panel A

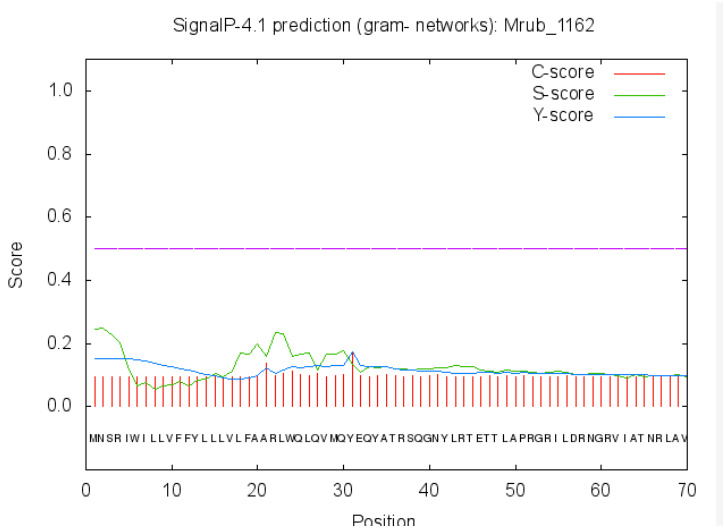


Panel B

Fig. 3) Mrub_1162 and b0635 both contain a single TMH domain within their membranes. Panel A shows the TMHMM for Mrub_1162. Panel B shows the TMHMM of b0635. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was used to obtain these charts.

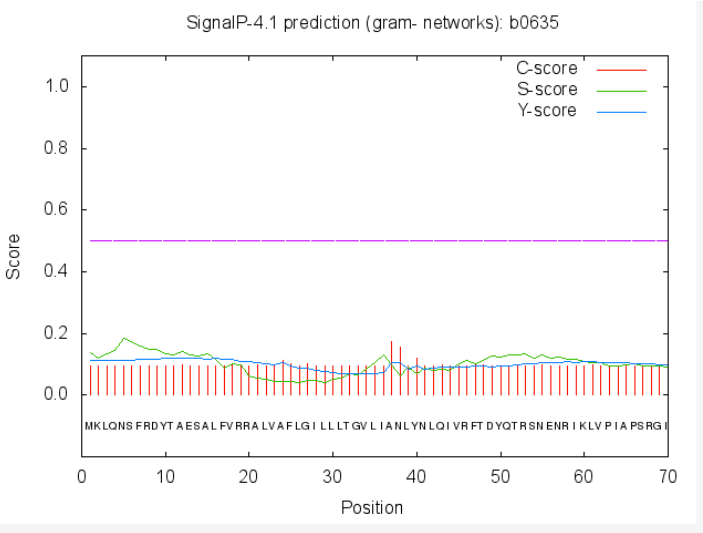
The figure 4 results were from a Signal P assay which is used to find cleavage sites within both Mrub_1162 and b0635. To find cleavage sites, a D-value is assigned to the specimen using the calculated value from corresponding S-scores, Y-scores, and a cutoff value which is represented by the purple line on the charts. The cutoff value for Mrub_1162 (Panel A) has a cutoff value .510 and a D-value of .162, in the end not making the cut. While b0635 (Panel B) has a cutoff

value of .510, but a D-value of .129, meeting the same fate as it's alleged orthologs. This data overall shows that there are no cleavage sites found in either of the gene.



Panel A

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    31        0.170
  max. Y    31        0.173
  max. S     2        0.248
  mean S     1-30     0.143
       D     1-30     0.162   0.510   NO
Name=Mrub_1162  SP='NO' D=0.162 D-cutoff=0.510 Networks=SignalP-TM
```



Panel B

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    37        0.174
  max. Y    12        0.119
  max. S     5        0.183
  mean S     1-11     0.146
       D     1-11     0.129   0.510   NO
Name=b0635      SP='NO' D=0.129 D-cutoff=0.510 Networks=SignalP-TM
```

Fig. 4) Mrub_1162 and b0635 do not show signs of containing cleavage sites, for both of their D-values were below the necessary cutoff value. Panel A shows the plot for Mrub_1162. Panel B shows the plot for b0635. These results were creating using SignalP Server v 4.1 http://www.cbs.dtu.dk/services/SignalP.

Figure 5 shows the metabolic pathway of peptidoglycan biosynthesis and it also shows the two genes (Mrub_1162, b0635*)* being expressed at identical spots of the pathway. Both of the genes are found to code for the same Penicillin-binding protein II. This gives a good indication that the genes area evolutionarily related and orthologous based on function.

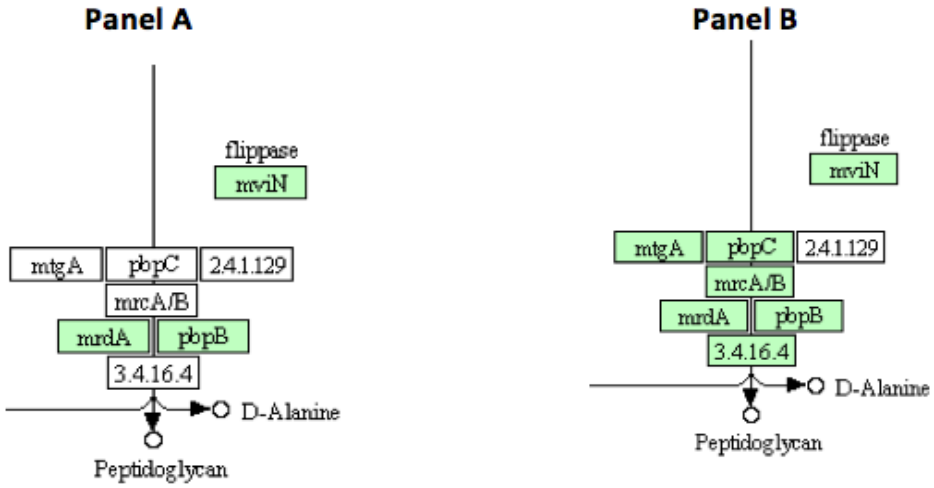
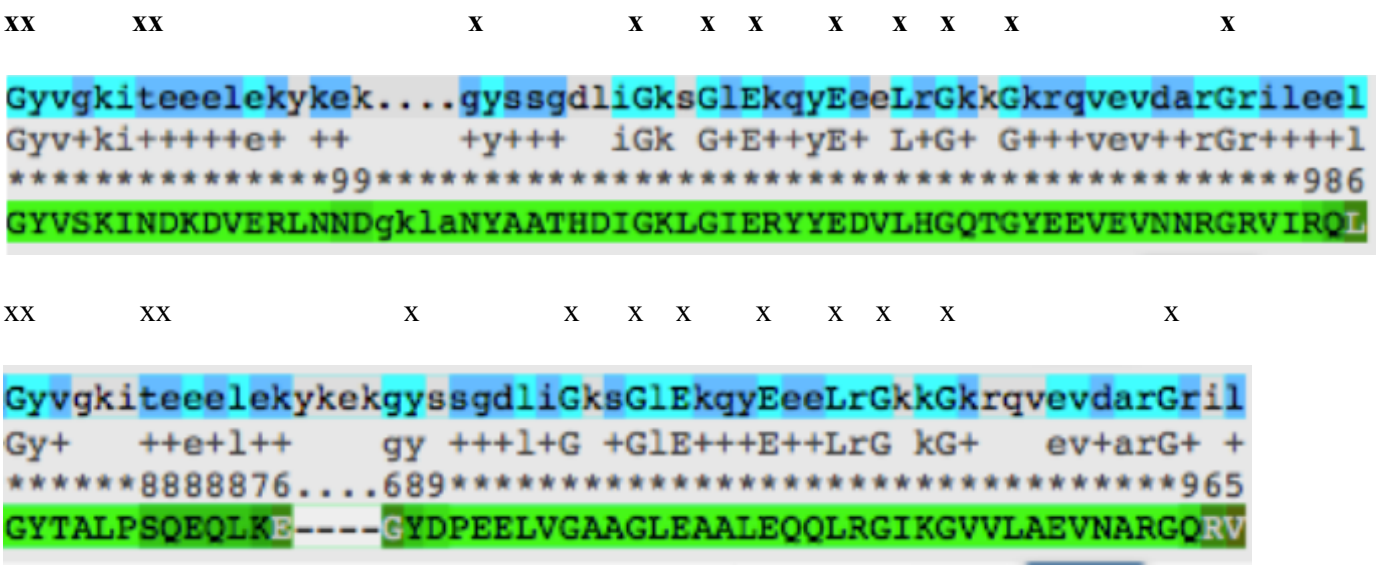


Fig. 5) Mrub_1162 and b0635 are both expressed in the same metabolic pathway. Panel A is the KEGG pathway when looking at *Meiothermus ruber (mrdA)*. Panel B is the KEGG pathway when looking at *Escherichia coli (mrdA)*. This chart has pulled from the KEGG website www.genome.jp/kegg/pathway.html.

In this figure, we are looking at an HMM alignment that shows the conservation of amino acids between Mrub_1162 and b0635 within their protein families. These alignments contain many conserved amino acids that are identical between the two species. The fact that these two genes pulled up the same consensus amino acids, reinforces the notion that the genes are similar in structure and function.

Panel A



Panel B

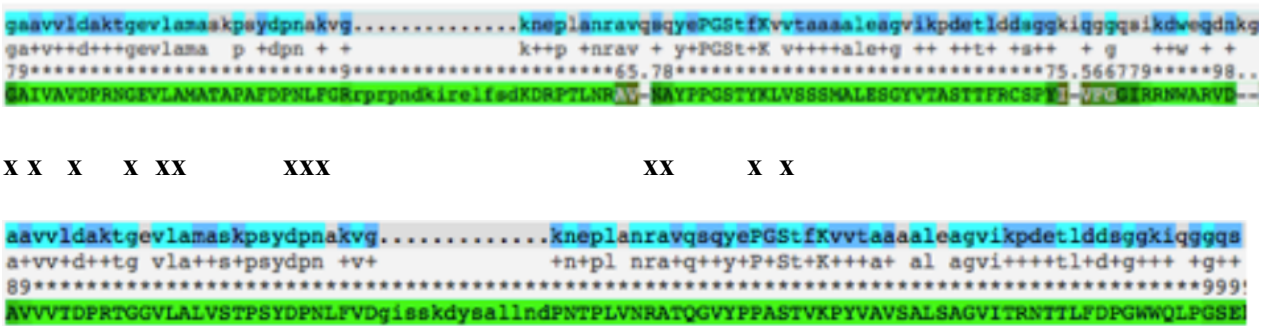
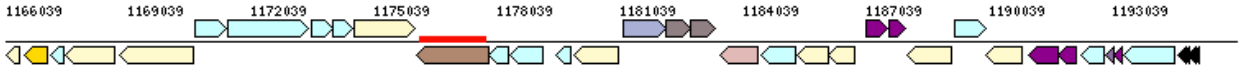


Fig. 6) Mrub_1162 and b0635 have identical highly conserved amino acids (marked with x) within their proteins in the same domain of Penicillin-binding protein II. Panel A represents PF03717: Penicillin-binding protein dimerization domain and panel B is PF00905: Penicillin-binding protein transpeptidase domain. These charts were made using the Pfam website pfam.sanger.ac.uk/search.

Figure 7 shows the gene neighborhood that flank Mrub_1162 and b0635. The red line identifies the query sequence. The color of a gene is unique to its predicted KEGG pathway; this illustration is called "Color by KEGG" and is derived from the IMG Gene Details pages (Markowitz, 2012). If the query was part of an operon, the flanking genes would have the same KEGG color. Our query genes appear to be isolated genes and not part of an operon. When compared to each other, there is a difference in color. This does not indicate a different KEGG map, however. On further investigation, both genes were identified coding *ftsI,* putative penicillin binding protein II.

Panel A



Panel B

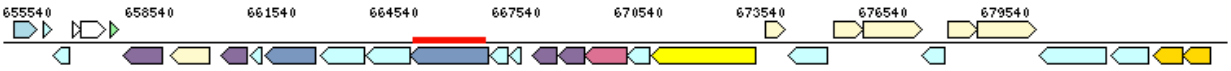


Fig. 7) Mrub_1162 and b0635 genes are not a part of an operon. Chromosome Viewer was colored via their KEGG database. Panel A: Mrub_1162 Chromosome Viewer. Panel B: b0635 Chromosome Viewer. Images were obtained from https://img.jgi.doe.gov.

Figure 8 shows a phylogeny tree that is composed of species that chosen from the previous T Coffee alignment. All of the species all belong to Deinococcus Thermus, showing that there is no chance of horizontal gene transfer.

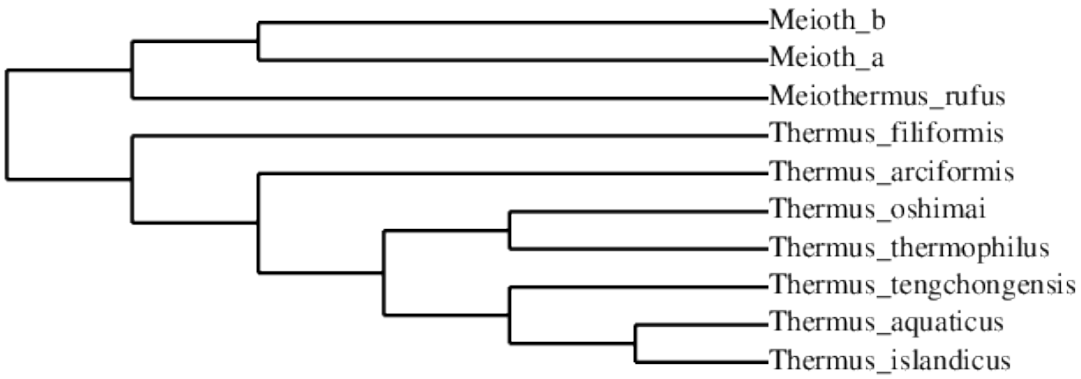


Fig. 8)) Phylogeny tree of species that express Mrub_1162 gene. Chart was created under the "A La Carte" setting of Phylogeny.fr at: http://www.phylogeny.fr

Table 4 summarizes the results comparing Mrub_1999 and *E. coli* b0084. A protein BLAST alignment produced an E-value close to 0.0, which indicates strong amino acid sequence similarity that likely reflects a common function. Searching the CDD with the two query sequences identified strong amino acid sequence similarity (*i.e.,* E-values less than the 0.001 cutoff) to the same protein domain (COG numberCOG0768, and COG name Cell division protein/ftsI); The bioinformatics tools used for the cellular localization module (LipoP, SignalP, TMH, PSORT-B, and Phobius) predict that Mrub1999 is localized to the cytoplasmic membrane, which is the known location for *E. coli* b0084 (Johnson, 2009). Phobius was especially informative; Mrub_1999 is predicted to have 1 transmembrane helix and b0084 has 1. Both query sequences pulled the identical TIGRfam hit TIGR03423 *ftsI* from the database. A search of the Pfam database showed the two genes belong to the same two protein families of PF03717: Penicillin-binding protein dimerization domain and PF00905: Penicillin-binding protein transpeptidase domain. When pulling data from the Protein Database, the Mrub_1999 gene coded for 3EQU: Penicillin binding protein III; while b0084 coded for 4BJP: Penicillin binding protein III. The Pfam database, and others such as KEGG, tagged both proteins as having the same E.C. number of 2.4.1.129. Mrub_1999 is predicted to catalyze the same reaction as b0084 on the KEGG map 00550. All of this data supports my hypothesis that Mrub_1999 and E. coli b0084 are orthologs. There was no contradictory data collected.

Table 4. Mrub_1999 and b0084 are orthologs

Bioinformatics Tools Used	**Mrub_1999**	**B0084**
BLAST *E. coli* against *M.ruber*	Score: 154 bits E-value: 4e-45	
CDD Data (COG category)	COG Number: COG0768 Cell Division Protein ftsI	
	E-Value: 6.23e-82	E-Value: 0
Cellular Localization	Cytoplasmic Membrane	
TIGRfam- Protein Family	TIGR03423 Penicillin binding protein II	
	E-Value: 2.4e-14	E-Value: 3.7e-38
Pfam- Protein Family	PF00905: Penicillin binding protein transpeptidase domain PF03717: Penicillin binding protein dimerization domain	
	E-Value: 4.3e-50	E-Values: 4.3e-84 2.4e-21
Protein Database	3EQU: Penicillin binding protein III	4BJP: Penicillin binding protein III
	E-Value: 1.7e-41	E-Value: 0
Enzyme Commission Number	EC: 2.4.1.129. Peptidoglycan Glycotransferase	
KEGG Pathway Map	Peptidoglycan Biosynthesis	

Figure 2 is the the protein BLAST alignment between Mrub_1999 and b0084, which was originally described in Table 4. There were 29% identical amino acids and 45% of amino acids of similar nature. The E-value of 4e-45 is well below the 0.001 cutoff and indicates a high level of primary sequence similarity and, consequently, functional similarity.

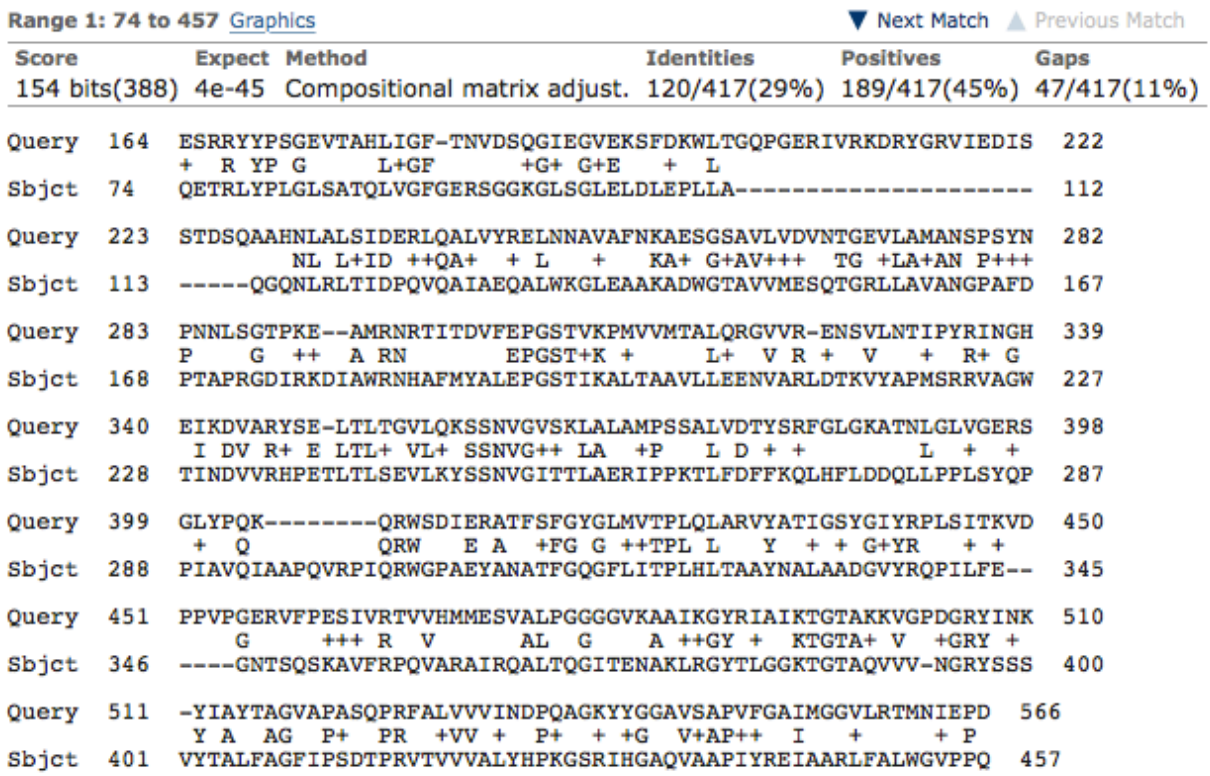
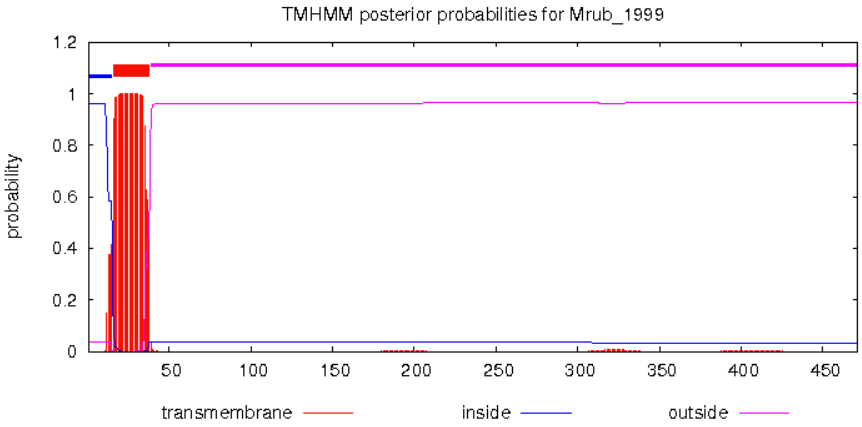


Fig. 2) Mrub_1999 and b0084 shows similarities in protein sequences when analyzed under BLAST. Query sequence: *M. ruber*. Subject sequence: *E. coli*. This analysis was performed using NCBI BLAST http://www.ncbi.nlm.nih.gov.

Figure 3 represents the TMH hydropathy charts for Mrub_1999 and b0084. The red peaks signify the presence of multiple transmembrane helices for both proteins because they are well above the probability cutoff of 1 TMH. The *Mrub_0981* is predicted to contain 1 TMH while *b1069* has 11 TMH. The presence of this many THM likely indicates a cellular location within the cytoplasmic membrane.

Panel A



Panel B

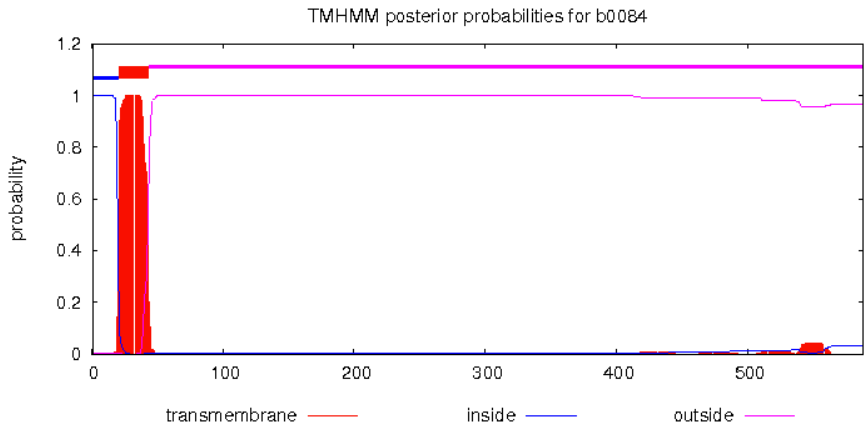
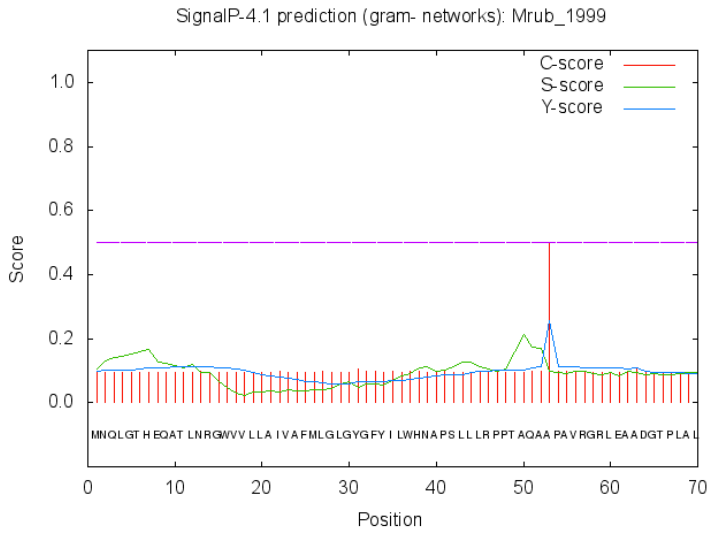


Fig. 3) Mrub_1999 and b0084 both contain a single TMH domain within their membranes. Panel A shows the TMHMM for Mrub_1999. Panel B shows the TMHMM of b0084. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was used to obtain these charts.

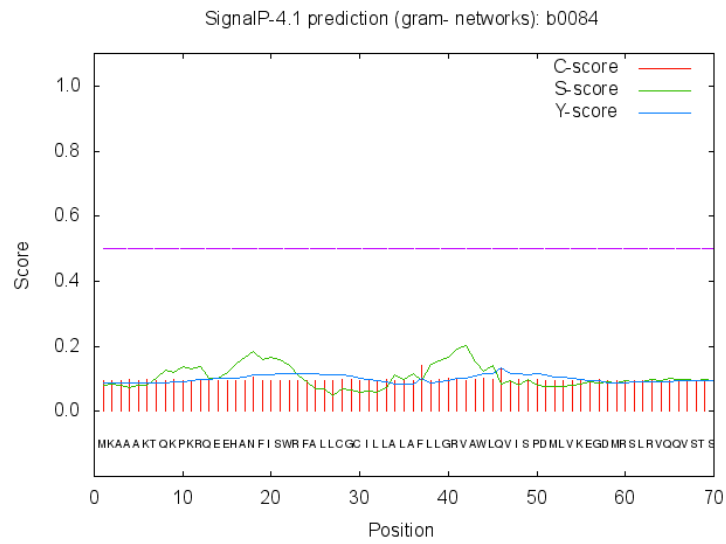
The figure 4 results were from a Signal P assay which is used to find cleavage sites within both Mrub_1999 and b0084. To find cleavage sites, a D-value is assigned to the specimen using the calculated value from corresponding S-scores, Y-scores, and a cutoff value which is represented by the purple line on the charts. The cutoff value for Mrub_1999 (Panel A) has a cutoff value .510 and a D-value of .196, in the end not making the cut. While b0084 (Panel B) has a cutoff

value of .510, but a D-value of .128, meeting the same fate as it's alleged orthologs. This data overall shows that there are no cleavage sites found in either of the gene.



Panel A

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    53        0.496
  max. Y    53        0.257
  max. S    50        0.213
  mean S     1-52     0.093
       D     1-52     0.196   0.510   NO
Name=Mrub_1999  SP='NO' D=0.196 D-cutoff=0.510 Networks=SignalP-TM
```



Panel B

```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    37        0.141
  max. Y    46        0.135
  max. S    42        0.201
  mean S     1-45     0.115
       D     1-45     0.128   0.510   NO
Name=b0084      SP='NO' D=0.128 D-cutoff=0.510 Networks=SignalP-TM
```

Fig. 4) Mrub_1999 and b0084 do not show signs of containing cleavage sites, for both of their D-values were below the necessary cutoff value. Panel A shows the plot for Mrub_1999. Panel B shows the plot for b0084. These results were creating using SignalP Server v 4.1 http://www.cbs.dtu.dk/services/SignalP.

Figure 5 shows the metabolic pathway of peptidoglycan biosynthesis and it also shows the two genes (Mrub_1999, b0084) being expressed at identical spots of the pathway. Both of the genes are found to code for the same Penicillin-binding protein III. This gives a good indication that the genes area evolutionarily related and orthologous based on function.

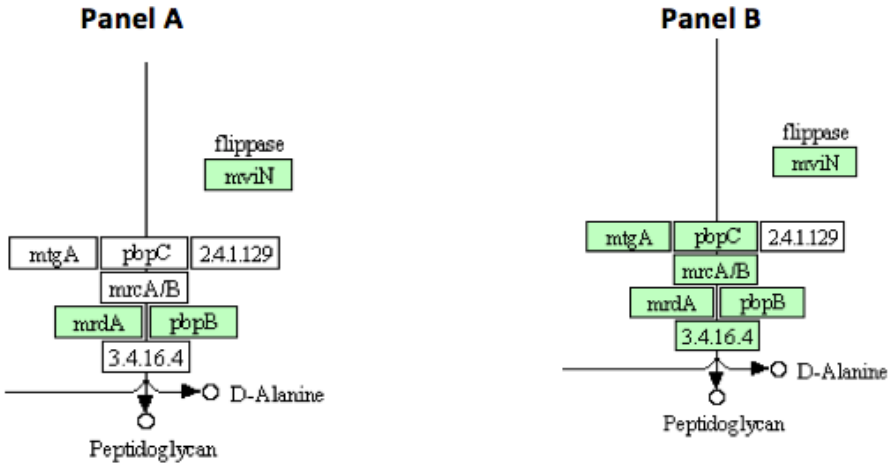
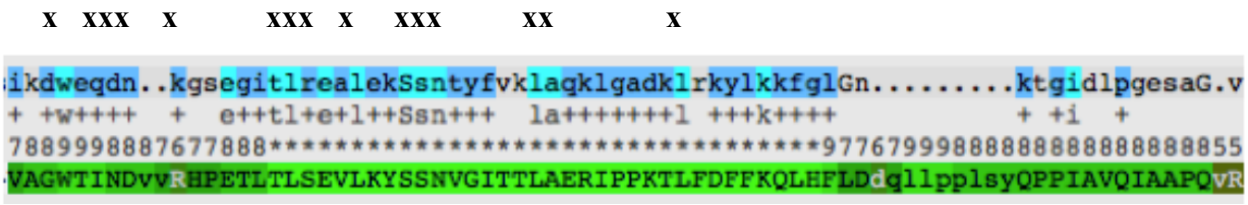


Fig. 5) Mrub_1999 and b0084 are both expressed in the same metabolic pathway. Panel A is the KEGG pathway when looking at *Meiothermus ruber (pbpB*). Panel B is the KEGG pathway when looking at *Escherichia coli (pbpB*). This chart has pulled from the KEGG website www.genome.jp/kegg/pathway.html.

In this figure, we are looking at an HMM logo that shows the conservation of amino acids within the proteins of Mrub_1999 and b0084. Contrarily to the previous genes that were discussed earlier, the HMM logo here does not show signs of having any highly conserved amino acids.

Or at least none that are a conserved as the other genes. The similarities between the two logos still gives rise to belief that the two genes are orthologous.

Panel A



Panel B

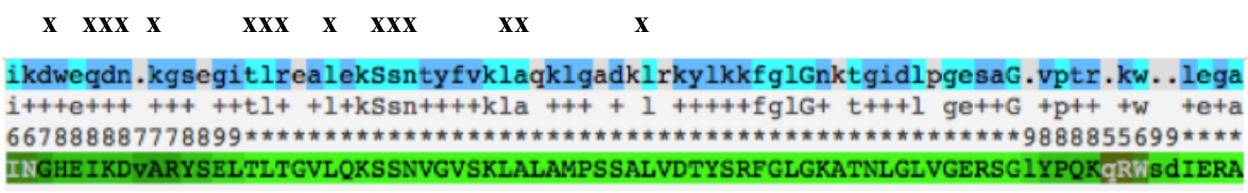
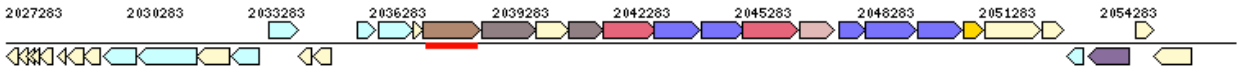


Fig. 6) Mrub_1999 and b0084 have identical highly conserved amino acids within their proteins in the same domain of Penicillin-binding protein III. Panel A represents Mrub_1999 and panel B is b0084. These charts were made using the Pfam website pfam.sanger.ac.uk/search.

Figure 7 shows the gene neighborhood that flank Mrub_1999 and b0084. The red line identifies the query sequence. The color of a gene is unique to its predicted KEGG pathway; this illustration is called "Color by KEGG" and is derived from the IMG Gene Details pages (Markowitz, 2012). Since the panel B query was part of an operon, the flanking genes are the same KEGG color. The panel A query gene appears to be and isolated gene and not part of an operon. When compared to each other, there is a slight difference in color. This does not indicate a different KEGG map, however. On further investigation, both genes were identified coding *pbpB,* penicillin binding protein III.

Panel A



Panel B

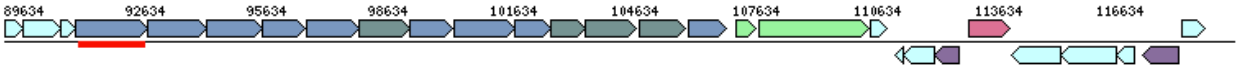


Fig. 7) Mrub_1999 gene is a part of an operon. b0084 gene is not a part of an operon. Chromosome Viewer was colored via their KEGG database. Panel A: Mrub_1999 Chromosome Viewer. Panel B: b0084 Chromosome Viewer. Images were obtained from https://img.jgi.doe.gov.

Figure 8 shows a phylogeny tree that is composed of species that chosen from the previous T Coffee alignment. All of the species all belong to Deinococcus Thermus, showing that there is no sign of horizontal gene transfer.

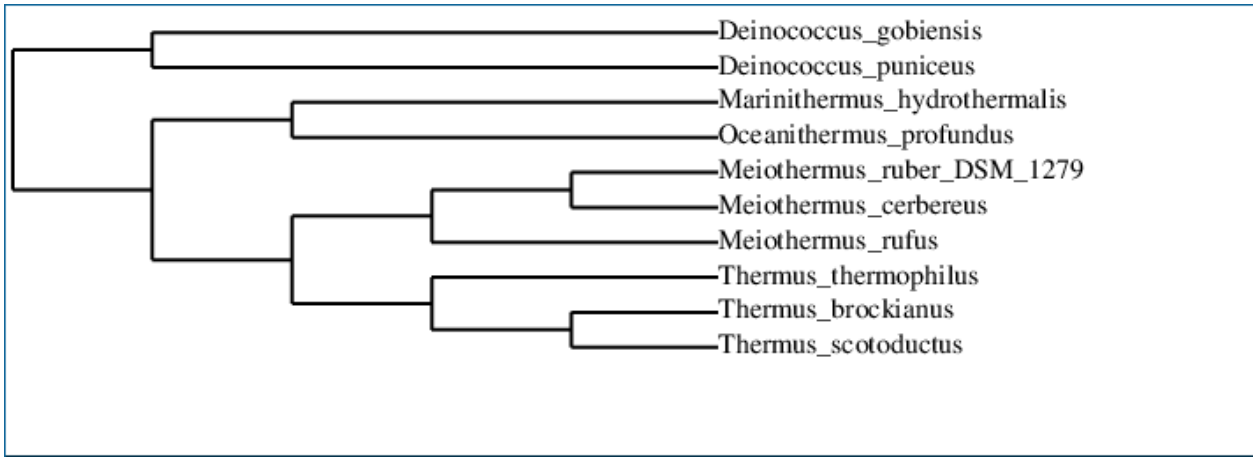


Fig. 8)) Phylogeny tree of species that express Mrub_1999 gene. Chart was created under the "A La Carte" setting of Phylogeny.fr at: http://www.phylogeny.fr

Conclusion

The results gained from this experiment has supported my hypothesis that genes: Mrub_0981 and b1069, Mrub_1162 and b0635, Mrub_1999 and b0084 are all orthologs. The first piece of evidence was a protein BLAST comparison, which resulted in E-values far below the 0.001

cutoff. Low E-values are indicative of strong sequence similarity, and, therefore, a strong likelihood of functional similarity. Next, were the TMHMM hydropathy charts whose predictions for each gene pair were the actual outcome for each chart. Having the genes contain the same amount of TMH they were predicted to have, gives strong hinting of similar cellular localization. Another analytical step that was important were the HMM alignments pulled from the Pfam database. They show conserved amino acids that belong in the protein sequence of both sets of genes. And when looking at each set, the gene orthologs were apart of the same protein family, and the amount of identical, highly conserved amino acids gave signs of similar function and structure between the genes.

Even though most of the bioinformatics data supported my hypothesis, I found one contradictory set of data (Figure 7c). When looking at the data pulled from the Chromosome Viewer "Color by KEGG Map" from the IMG database (Markowitz et al., 2012), it shows that the Mrub_1999 may not be part of an operon based on its color being different than the flanking genes. However, when I compared this same region among different species related to *M. ruber*, there

does appear to be a common set of genes in this region:

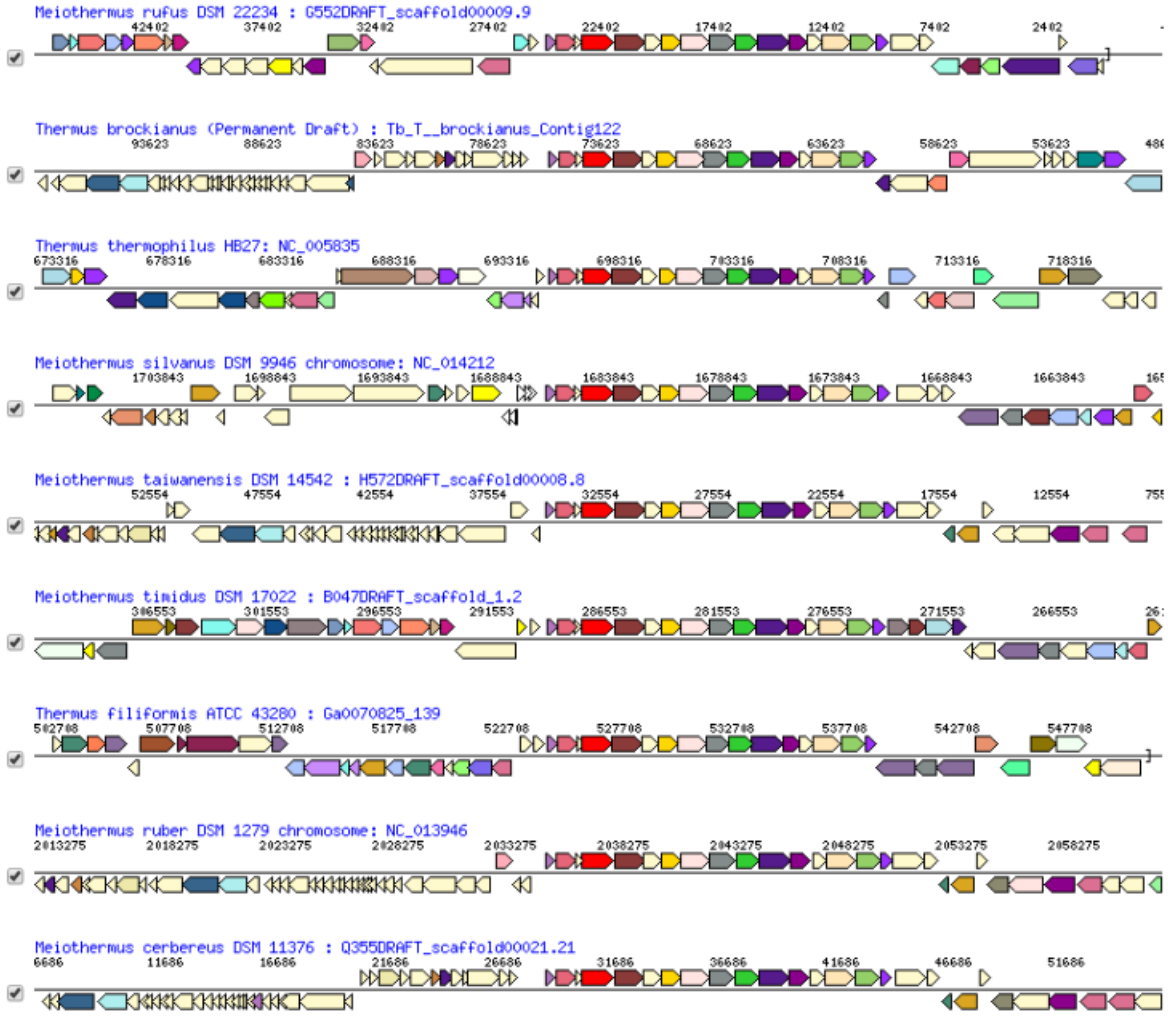


Fig. 9) Mrub_1999 gene orthologs neighborhood formatted under regions with the same top COG hit (via top homolog). Orthologs chart was pulled from IMG data base: http://pfam.xfam.org

While the b0084 gene is a part of an operon.

A portion of the modules from the GENI-Act website contained phylogenic tress that each contained various species that expressed the same gene that is present within both *E. coli* and *M. ruber*. The many species were taken from the T Coffee database (Notredame et al., 2000) and plugged into the "A La Carte" system of the phylogenic tree creator (www.phylogeny.fr). This step was used to identify if the presence of horizontal gene transfer (HGT). HGT happens when

a transfer of genes occurs between two species of the same generation. This may cause the expression of a phenotype that has the chance to undergo many mutations. And these mutations can result in a bacterium undergoing evolution at a faster, more efficient speed (www.geni-act.org). If a group of unrelated organisms contain similarities, then there is a chance that the model organism acquired a gene from a different organism that was mutated and eventually expressed, and vice versa. Fortunately, the results of the phylogenic trees obtained from my research contained very little signs of HGT due to the close relations of species that came up from the T Coffee analysis.

If I were to perform an experiment utilizing site-directed mutagenesis to disrupt the function of peptidoglycan production, I would focus on using the Mrub_0981 gene that codes for the lipid II flippase protein. There is a Tyrosine that is highly conserved at amino acid 346 within the protein sequence. I chose this amino acid because it has the property of being affiliated with phosphorylation (Betts and Russell, 2013). This helps with signal transduction across the membrane of the cell which plays a really good role with a protein that deals with peptidoglycan synthesis. The nucleotide sequence for Tyrosine is TAT, and I would make a deletion of the A so that the reading frame will shift and have the sequence read CTT, which codes for Leucine (http://nebasechanger.neb.com). Leucine has rarely been linked to protein activity making it a very good substitute (Betts and Russell, 2003). I feel that this would be a good experiment to run to disrupt the production of peptidoglycan within both *E. coli* and *M. ruber* cells. The program I used to layout this experimental approach was BioLab's NEBaseChanger.

Fig. 10) HMM Logo of Mrub_0981 gene showing a highly conserved Tyrosine. This image was found on the Pfam database at http://pfam.xfam.org.

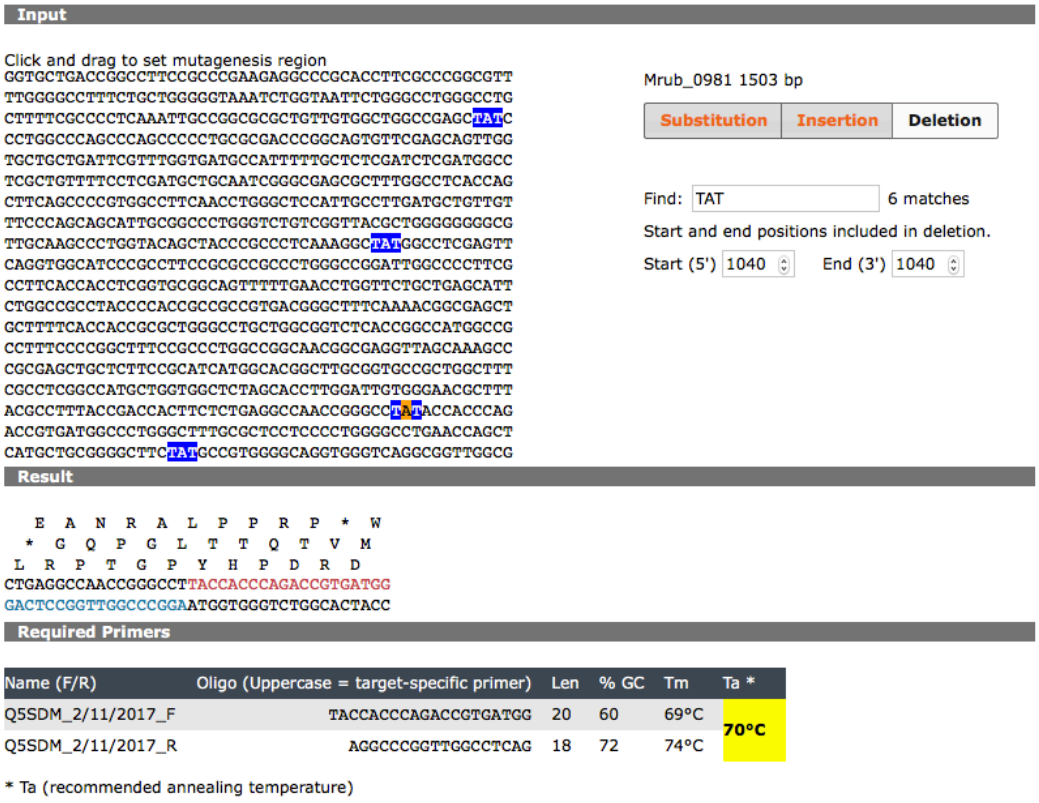


Fig. 11) Resulting chart of NEBaseChanger when looking at Mrub_0981 nucleotide sequence. Including appropriate flanking primers and annealing temperatures. The chosen TAT sequence has the highlighted "A", representing what will be deleted. This chart was pulled from the NEBaseChanger database at http://nebasechanger.neb.com.

Literature 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/
- Freer, J. H., and M. R. Salton. 1971. The anatomy and chemistry of Gram negative cell envelopes. Microbial Toxins. 4: 67-122. Available from: Academic Press Inc., New York.
- 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.
- 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/

- 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

- Matthew JB, Robert BR. 2003. Amino Acid Properties and Consequences of Substitution. Bioinformatics for Geneticists [Serial Online]. [Cited 2016 Nov 12].

- Natividad Ruiz. 2013. Filling holes in peptidoglycan biogenesis of Escherichia coli. *Current Opinion in Microbiol. 38: 1.* Available from: http://www.sciencedirect.com/science/article/pii/S1369527416300984

- NEBaseChanger. [Internet]. 2017. New England BioLabs; [2017 Feb 9]. Available from: http://nebasechanger.neb.com/

- 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

- Ruiz N (2008). "Bioinformatics identification of MurJ (MviN) as the peptidoglycan lipid II flippase in Escherichia coli." Proc Natl Acad Sci U S A 105(40);15553-7.

- Schleifer KH, Kandler O (1972). "Peptidoglycan types of bacterial cell walls and their taxonomic implications." Bacteriol Rev 36(4);407-77.

- Sham LT, Butler EK, Lebar MD, Kahne D, Bernhardt TG, Ruiz N (2014). "Bacterial cell wall. MurJ is the flippase of lipid-linked precursors for peptidoglycan biogenesis." Science 345(6193);220-2.

- Spratt BG. (1975) “Distinct penicillin binding proteins involved in the division, elongation, and shape of Escherichia coli K12.” Proc Natl Acad Sci U S A 72(8); 2999-3003.

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