Mrub_3029, Mrub_2052, are predicted orthologs of b_0688, b_0394, while Mrub_0759 and Mrub_2365 are not predicted orthologs of b_1309, in Escherichia coli, which code for enzymes involved in starch and sucrose metabolism.

Max A. Benstine
Augustana College, Rock Island Illinois

Dr. Lori R. Scott
Augustana College, Rock Island Illinois

Follow this and additional works at: http://digitalcommons.augustana.edu/biolmruber

Part of the Bioinformatics Commons, Biology Commons, and the Molecular Genetics Commons

Augustana Digital Commons Citation
Benstine, Max A. and Scott, Dr. Lori R.. "Mrub_3029, Mrub_2052, are predicted orthologs of b_0688, b_0394, while Mrub_0759 and Mrub_2365 are not predicted orthologs of b_1309, in Escherichia coli, which code for enzymes involved in starch and sucrose metabolism" (2017). Meiothermus ruber Genome Analysis Project. http://digitalcommons.augustana.edu/biolmruber/27

This Student Paper is brought to you for free and open access by the Biology at Augustana Digital Commons. It has been accepted for inclusion in Meiothermus ruber Genome Analysis Project by an authorized administrator of Augustana Digital Commons. For more information, please contact digitalcommons@augustana.edu.
Mrub_3029, Mrub_2052, are predicted orthologs of b0688, b0394, while Mrub_0759 and Mrub_2365 are found not to be predicted orthologs of b1309, in Escherichia coli, which code for enzymes involved in starch and sucrose metabolism

Max A. Benstine  
Dr. Lori R. Scott Laboratory  
Biology Department, Augustana College  
639 38th Street, Rock Island, IL 61201

Introduction
Why Study Meiothermus Ruber/What is known
Originally, Meiothermus ruber (M. ruber), belonged to the genus Thermus until 1996 when it was added to the newly formed genus Meiothermus because it was determined to live in temperatures of lesser heat in comparison to species found in the Thermus genus (Tindall et al., 2010). Characteristics of M. ruber are as follows; red pigmentation, aerobic, non-motile, gram-stain negative, rod shaped cells, and has a temperature range of 35°C-70°C, optimally functioning at 60°C. M. ruber was first recorded at being collected in 1973 in the hot springs of Kamchatka Peninsula of Russia because this environment is ideal for the bacteria’s growth. Since its first collection, it has also been found in similar environments across northern Asia and Europe. The M. ruber genome analysis project is a collaborate of the Joint Genome Institute (JGI) which has been attempting to find and sort thousands of pieces of data to complete the Genomic Encyclopedia of Bacteria and Archaea (GEBA) (Hornick, 1997). Although there are nearly 15,000 drafted genome sequences available in GEBA, unfortunately, there appears to be an unequal amount of research done on some species in comparison to others. The goal of the Meiothermus ruber genome analysis project is to gather information on M. ruber and to lessen the gap in information between it and other microbes such as E. coli (Scott, 2016). Closing this information gap is necessary because it could possibly uncover genes that are essential in functions that are not present in the more analyzed bacteria. Fortunately, having a substantial amount of data for E. coli makes it a reliable control microbe that can be used to help determine missing information in the M. ruber genome (Scott, 2016).

E. coli, the Model Organism
To help gain genetic information about starch and sucrose metabolism in M. ruber, E. coli is being used as a model organism. Since E. coli has a complete genome sequenced (GENI-Science, 2011) proteins found in the same location as the M. ruber genes being studied along the KEGG pathway map of starch and sucrose metabolism were selected. The M. ruber and E. coli protein pairs were determined using matching placement and E.C. numbers in the KEGG pathway map site of the starch and sucrose metabolism of both E. coli and M. ruber. KEGG pathway confirmed the following similarities. Mrub_0759 and Mrub_2365 match with E. coli b0394, Mrub_3029 matches with b1309, and Mrub_2052 matches with b0688. However, there appears to be some discrepancy in the enzyme identification found on the KEGG pathway map of “starch and sucrose metabolism” of Escherichia coli because all three genes are found under the same E.C. number of (2.7.1.4), but when b0394 is selected, it is labeled in its page as E.C. number (2.7.1.2) which is the label for glucokinase. It can be noticed (and will be addressed)
further along in this study that when the amino acid sequences of \textit{b0394} undergo NCBI’s BLAST with either \textit{Mrub}_0759 or \textit{Mrub}_2365, very little similarity will be found. Fortunately, Ecocyc.org has identified \textit{b0394} as responsible for coding for glucokinase which helps explain why there appears to be such great differences between \textit{b0394/Mrub}_0759 and \textit{b0394/Mrub}_2365 because both \textit{M. ruber} genes are found to belong to the fructokinase protein family, rather than the glucokinase. Using bioinformatics tools such as TIGRfam (Haft \textit{et al.}, 2001) the proteins these \textit{E. coli} genes code for were able to be determined. \textit{E. coli} \textit{b1309} appears to be part of an operon and codes for sucrose phosphorylase, \textit{b0688} is part of an operon and codes for phosphoglucomutase, \textit{b0394} does not appear to belong to an operon but codes for fructokinase (GENI-ACT 2014). Another reason for using \textit{E. coli} as the control for this research is because the bacteria’s starch and sucrose metabolism is heavily used in modern day industry. For instance, \textit{E. coli} is used in the production of ethanol because it can efficiently use a cheap carbon source such as molasses in creating the necessary byproducts. Currently, there is research being funded to discover even more efficient aerobic pathways to aide in the production of ethanol (Förster \textit{et al.}, 2014). Therefore, it is possible that finding other organisms with similar pathways may be even better suited for this task. Which is why collecting more information on the genome of \textit{M. ruber} is beneficial, because this new information may possess more comparative data in sucrose metabolism that can be used by the biofuel industry.

\textbf{Biosynthesis Pathway of Starch and Sucrose Metabolism}
Figure 1. KEGG image of starch and sucrose metabolic pathway, displays locus tags of genes that are found to code for important protein structures within the pathway. Highlighted in green are the ExPASy enzyme numbers (E.C.) which are used to identify specific enzymes within the desired organism’s metabolic pathway. Circled in red are the enzymes that contain the genes of both M. ruber and E. coli which are being annotated in this project. E.C. (2.7.1.4) contains the locus tags; Mrub_0759, Mrub_2365, b0394(mak). E.C. (5.4.2.2) contains the locus tags; Mrub_2052, b0688(pgm). E.C. (2.4.1.7) contains the locus tags; Mrub_3029, b1309(ycjM). The image was acquired from the KEGG genome mapping website (Kanehisa et al. 2016)

Phosphorylation of fructose
Fructokinase (mak) appears to be responsible in E. coli for fructose metabolism when the primary phosphoenolpyruvate/glycose phosphotransferase system (PTS) is not functioning (Kornberg et al., 2000). When PTS is not available to diffuse fructose, fructokinase is the enzyme found responsible in the secondary pathway. When fructose is brought into the cell, fructo(manno)kinase (mak) utilizes ATP to phosphorylate the fructose molecules into fructose-6-phosphate (as shown below in figure 2) (MetaCyc.org). In this study the gene b0394 which has been determined to code for fructokinase in E. coli was used as the model to compare the lesser-known M. ruber genes; Mrub_0759 and Mrub_2365. This was to help determine if these two paralogs also play a role in fructose phosphorylation. It is important to note that b0394 in table 1 and 2 later in this paper will refer to glucokinase due to a discrepancy in the KEGG pathway map site, which was discussed in more detail in the “E. coli, model organism” section.

![Fructose phosphorylation reaction](https://biocyc.org/META/NEW-IMAGE?type=REACTION&object=FRUCTOKINASE-RXN)

Glycogen Degradation II
Phosphoglucomutase (Pgm) is an enzyme responsible for the last step of the glycogen degradation II pathway (metacyc) as shown by the pathway displayed in Panel A below. In the presence of the substrate, alpha-D-glucose 1,6-bisphosphate (http://enzyme.expasy.org),
phosphoglucomutase will act as the catalyst for the rearrangement of the intermediate, alpha-D-glucopyranose-1-phosphate into D-glucopyranose-6-phosphate (displayed in Panel B, below) to be further used as a substrate in the glycolysis I pathway. The proposed reason D-glucopyranose-6-phosphate is more readily used to initiate glycolysis and form glycogen for energy storage is because the addition of the phosphate group on the 6th carbon atom inhibits the glucopyranose from diffusing out of the cell membrane (Berg, 2002). In this study we used the gene \textit{b0688} which helps code for phosphoglucomutase, as the model gene to determine if \textit{Meiothermus ruber}'s gene \textit{Mrub_2052} also codes for the same enzyme which helped determine if there may be any common ancestry between the two.

Panel A

Panel B

Figure 3. Panel A displays the MetaCyc image of the last two steps found in the Glycogen Degradation II pathway. Boxed off in red is the final step where Phosphoglucomutase catalyzes the transformation of alpha-D-glucopyranose-1-phosphate into D-glucopyranose-6-phosphate. Panel B shows the MetaCyc image for the balanced chemical reaction of the transformation of alpha-D-glucopyranose-1-phosphate into D-glucopyranose-6-phosphate. Images were taken from https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5941 and https://biocyc.org/META/NEW-IMAGE?type=REACTION&object=PHOSPHOGLUCMUT-RXN, respectively.

Sucrose Degradation IV

Not much is known about Sucrose Phosphorylase (\textit{SucP}) in \textit{E. coli} aside from that it is the putative enzyme responsible for catalyzing the reaction between sucrose (consisting of glucose and fructose) with a phosphate group in order to cleave the sucrose and create D-fructose and alpha-D-glucose-1-phosphate (Reid, 2005). These two products can then be used by the two afore mentioned enzyme catalysts in this study (fructokinase will phosphorylate beta-d-fructofuranose and Phosphoglucomutase will catalyze the phosphorylation of alpha-D-glucopyranose-1-phosphate). This information is useful in this study because it helps determine if the gene \textit{Mrub_2052} has any similarity in function. After this has been determined it can further be explored if there is potential common ancestry between the genes \textit{b1309} and \textit{Mrub_2052}. 
Figure 4. Displays the cleavage of sucrose into fructose and glucose because sucrose phosphorylase was introduced to the substrate in order to catalyze the phosphorylation of sucrose into fructose and glucose. Image retrieved from: https://biocyc.org/META/NEW-IMAGE?type=REACTION&object=SUCROSE-PHOSPHORYLASE-RXN

**Purpose/Hypothesis**

The purpose of this research project is to predict if the genes Mrub_0759, Mrub_2365, Mrub_3029, and Mrub_2052 are orthologs of b0394(mak), b1309(ycjM), and b0688(pgm) genes in *Escherichia coli*, respectively. GENI-ACT was used to effectively organize the retrieved data and provide links to various bioinformatics tools discussed in the *Methods* section of this report. To determine if these genes are orthologous, bioinformatics tools provide “Expect-values” (E-values) which are important for recognizing the significance of each gene comparison (Madden et al. 2002). According to the National Center for Biotechnology Information (NCBI), website we hope to find low E-values because this indicates that the gene sequence alignments are significant, rather than likely being due to chance which is what high E-values indicate (Madden, 2002). For each gene pair an initial BLAST was run to determine E-values, Mrub_0759/b0394 have an E-value of (0.37), therefore, we hypothesize the two genes are not orthologs. Mrub_2365/b0394 also have an E-value of (0.83) which is why it is hypothesized the two genes are also not orthologs. Mrub_3029/b1309 have an E-value of (3e-168), therefore, it is also hypothesized the two genes are orthologs. Mrub_2052/b0688 have an E-value of (0.0), which is why it is hypothesized that these two genes are also orthologous to each other.

**Materials/Methods**

At the start of this project, we used the GENI-Science site (http://www.geni-science.org/) to find a suitable project question. Once we had our genes of interests, the GENI-ACT site (http://www.geni-act.org/), which is an online lab notebook, was used to organize the bioinformatics data collected on the *E.coli* and *M. ruber* genes. GENI-ACT also provides links to the various bioinformatics programs used in this project, as well as, basic information for the gene being researched, such as; Locus Tag, gene coordinates, nucleotide sequence, amino acid sequence, and sequence lengths. Once the putative orthologous *M. ruber* and *E.coli*’s proteins/genes were identified, the similarities among the species were compared using NCBI’s BLAST (Madden, 2002). BLAST’ING these protein sequences also helped determine if the proteins had common domains by using the Conserved Domain Database search. This database helps determine, if the protein belongs to a highly specific Cluster Orthologous Group (COG) (Marchler-Bauer et al., 2016). The next bioinformatics tool used was TCOFFEE, 15 BLAST hits were collected for each *E.coli* and *M. ruber* gene to create a multi-species alignment in the TCOFFEE program (Notredame et al. 2000). The TCOFFEE alignment helps create a multiple sequence alignment which helps compare the closely related species and determine if our selected organism shares a similar start codon or has an alternate start codon (http://img.jgi.doe.gov/). It is important to know that or sequence is aligned properly as we go about using other bioinformatics tools because if it is not, it could skew our data and result in
high E-values (Scott, 2016). WEBlogo is a program that uses the TCOFFEE sequence alignments to create an image of the most conserved amino acids across the different species. Conservation rate and presence of the amino acid at the specific location along a sequence is portrayed by the height of the amino acid abbreviation letter and how large it is, respectively (Crooks et al. 2004). Next, Gram stains were determined to see what type of cell wall each organism had. E. coli and M. ruber both are Gram-Negative (Koury, 2010). The TMHMM program (http://www.cbs.dtu.dk/services/TMHMM) predicts the number of transmembrane helices in a protein (Krough et al., 2016). SignalP (http://www.cbs.dtu.dk/services/SignalP) was another program used to help determine cellular localization for the proteins. If there was a significant D-value calculated then this was interpreted as signal-peptides being present. This D-value is calculated by using the; C-score (signal peptide cleavage site), S-score (signal peptide position), Y-score (combined C and S scores), and cutoff value, portrayed by the horizontal purple line in the graph (Peterson et al. 2011). LipoP (http://www.cbs.dtu.dk/services/LipoP/) program was used to predict a second type of N-terminus signal peptide specific to a lipoprotein (Koury, 2010). PSortB (http://www.psort.org/psortb) was the program used in predicting the number of transmembrane helices (Yu et al., 2010). Lastly, Phobius (http://phobius.sbc.su.se) is a program used as a secondary conformation for predictions of signal peptides and transmembrane-helices within the cells (Koury, 2010). Next, structural similarities between the orthologous protein pairs were determined. Common protein families were identified using TIGRfam (Haft et al., 2001). The Pfam program is helpful for identifying protein domains in a specific query sequence, it also helps create HMM logos which help visualize highly conserved amino acids found in similar proteins in different species (Sonhammer et al., 1997). Next, the Protein Data Bank (PDB) is accessed which gives 3-D models of protein structures our gene is found in, it helps portray how the protein is folded which helps predict function (Bernman et al., 2003). The recommended Metacyc tool was not used, but its sister site, Ecocyc which is devoted to E. coli K12 MG1655 (Kesler et al., 2017). This is a highly curated site, which provided pathway information. E.C. numbers (ExPASy enzyme numbers) were then determined to help identify which enzymes are important in the starch and sucrose metabolism of M. ruber and E. coli (http://www.expasy.ch/enzyme/enzyme-search-ec.html). Paralogs are then determined by comparing NCBI’s BLAST amino acid sequences which appear to be extremely similar. These paralogs are the result of a gene duplication event within a genome (Koury, 2010). KEGG maps were created for starch and sucrose metabolism and colored to determine where M. ruber and E. coli to help determine if any of the selected genes belonged to an operon (Kanehisa et al., 2016). Furthermore, to determine if there was horizontal gene transfer (which would hint to a shared ancestry amongst the selected organisms), a phylogenetic tree was created (http://www.phylogeny.fr) of M. ruber and E. coli using the sequences collected for T-COFFEE alignment.

Results

Table 1 provides a summary of some of the most helpful information provided from the bioinformatics tools used to compare E. coli b0394 (mak) to Mrub_0759. The BLAST information in the first row shows that this comparison has an extremely low bit score (17.7) which is not the only data that should be considered to write off the genes as non-orthologous. However, the Expect-value (0.37) is extremely high which indicates that any alignments between the sequences are most likely due to chance. Thus far, it appears that these two genes have no relation, because when the CDD was run they each received different COG numbers.
(b0394(mak)-COG1940) and (Mrub_0759-COG0524) however they are predicted to belong to a similar sugar-kinase family. With this information, it is more likely that the genes code for a similar enzyme in starch and sucrose metabolism. According to the bioinformatics tools used to determine cellular location such as, THM, SignalP, LipoP, and PSORT-B, the enzyme (fructokinase) can be found in the cytoplasm of the cell. Data acquired from TIGRfam showed separate identification number for each organism. Furthermore, the Pfam test rendered that the two protein sequences belong to two separate domains (PF00480, ROK family) and (PF00294, PfkB family). The enzyme commission number (E.C. 2.7.1.4) also points at the two genes being similar in purpose. They were predicted to be part of the same step of starch and sucrose metabolism in the KEGG pathway information.

<table>
<thead>
<tr>
<th>Bioinformatics Tool used</th>
<th>E. coli b0394(mak) gene</th>
<th>M. ruber Mrub_0759 gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST E. coli against M. ruber</td>
<td>Score: 17.7 bits E-value: 0.37</td>
<td></td>
</tr>
<tr>
<td>CDD Data (COG category)</td>
<td>COG Number: COG1940</td>
<td>COG Number: COG0524</td>
</tr>
<tr>
<td>Sugar kinase family</td>
<td>E-value: 1.19e-78</td>
<td>E-value: 2.98e-55</td>
</tr>
<tr>
<td>Cellular Localization</td>
<td>Cytoplasm of the cell</td>
<td></td>
</tr>
<tr>
<td>TIGRfam-protein family</td>
<td>TIGR00744</td>
<td>TIGR02152</td>
</tr>
<tr>
<td>E-value: 8.6e-14</td>
<td>E-value: 8.3e-11</td>
<td></td>
</tr>
<tr>
<td>Pfam- protein family</td>
<td>PF00480 (ROK family)</td>
<td>PF00294 (PfkB)</td>
</tr>
<tr>
<td>E-value: 2.2e-100</td>
<td>E-value: 6.3e-52</td>
<td></td>
</tr>
<tr>
<td>Protein Database</td>
<td>(4U7X) Crystal structure of Fructokinase from Brucella abortus 2308</td>
<td></td>
</tr>
<tr>
<td>E-value: 1.91715E-49</td>
<td>E-value: 9.56355E-48</td>
<td></td>
</tr>
<tr>
<td>Enzyme Commission number</td>
<td>2.7.1.4–fructokinase</td>
<td></td>
</tr>
<tr>
<td>KEGG pathway map</td>
<td>Starch and sucrose metabolism pathway</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5 (below) represents the BLAST analysis between gene sequences of E.coli b0394 and Mrub_0759 it provides some of the most important information for this project such as amino acid match identity and Expect value. 32% of the amino acids were found to be the same between these two sequences (however, it is important to note that the subject’s sequence is at a greater length than the query sequence). This information along with the large E-value of 0.37 indicates that there is quite a large probability that these two sequences aligned due to chance. With this knowledge, it is important to thoroughly investigate all further evidence because just this first set of data indicates that these two genes are not orthologous.
Figure 5. *Mrub_0759* and *E. coli b0394* (mak) do not appear to have any significant similarities among their protein sequences. Query sequence is *Mrub_0759*; Subject sequence is *E. coli b0394*. Sequence analysis was performed using NCBI BLAST analyzer at https://blast.ncbi.nlm.nih.gov

Figure 6 (below) displays transmembrane-helices hydropathy results of *E. coli b0394* and *Mrub_0759*. As described in the legends of these graphs, red lines would represent the appearance of transmembrane helices proteins. Obviously, there are no vertical red lines occurring in these two graphs, which indicates that no TMHs are present in either *E. coli b0394* or *Mrub_0759*. This information indicates a commonality between the proteins that these two genes code for, because the proteins are both found in the cytoplasm rather than the membrane.

Panel A

| No topology graph was able to be formed because of insufficient data | |

```plaintext
# E-coli-manno(fructo)kinase-protein Length: 302
# E-coli-manno(fructo)kinase-protein Number of predicted TMHs: 0
# E-coli-manno(fructo)kinase-protein Exp number of AAs in TMHs: 1.43379
# E-coli-manno(fructo)kinase-protein Exp number, first 60 AAs: 0.00389
# E-coli-manno(fructo)kinase-protein Total prob of N-in: 0.02712
E-coli-manno(fructo)kinase-protein TMHM2.0 outside 1 302
```
Panel B

Figure 6. Panel A shows the TMHMM results for *E. coli b0394*; Panel B shows the TMHMM results for *Mrub_0759*. *E. coli b0394* and *Mrub_0759* do not contain any Transmembrane-helices; data prediction asserts that the proteins are both found in the cytoplasm of the cells. TMHMM Server v 2.0 was used to determine these two Transmembrane-helices graphs found at [http://www.cbs.dtu.dk/services/TMHMM](http://www.cbs.dtu.dk/services/TMHMM)

Figure 7 contains SignalP graphs for *E. coli b0394* and *Mrub_0759*. In SignalP server v. 4.1, the selected proteins are determined a “D-value”. This value is calculated by using the S-score, Y-score, and cutoff value(portrayed by the horizontal purple line). Panel A (*E. coli b0394*) has a D-value of (0.102) which is far below the cutoff of (0.570). This is indicative of the protein, *b_0394* codes for, does not contain any cleavage sites. Panel B (*Mrub_0759*) has similar data showing a D-value of (0.166) but having a cutoff value of (0.570). This information helps conclude that both genes are similar to one another in that they code for proteins which lack cleavage sites.
Figure 7. *E. coli* b0394 and *Mrub_0759* are predicted to not contain any cleavage sites according to the bioinformatics tool, SignalP server v. 4.1 found at the following link; http://www.cbs.dtu.dk/services/SignalP The program uses a cut-off measurement (D-value of 0.570) and both plots show that the genes’ values were below the cut-off. Panel A shows the plot for *E. coli* b0394; Panel B shows the plot for *Mrub_0759*.

Figure 8 displays the starch and sucrose metabolism biochemical pathway, that both *E. coli* b0394 and *Mrub_0759* are involved in. Enzymes which are highlighted in green are predicted to be present in the selected organism. The pathways show that *E. coli* b0394 and *Mrub_0759* are
both genes involved in the second to last step of fructose biosynthesis, both coding for the enzyme fructokinase in the starch and sucrose metabolism pathway. This similarity between the genes provides evidence that they are predicted orthologs.

Figure 8. *E. coli* b0394 and *Mrub_0759* are found in the same position in the same biochemical pathway. Panel A displays *b0394* in the KEGG pathway after selecting for *Escherichia coli* in the starch and sucrose metabolism pathway. Panel B displays *Mrub_0759* in the KEGG pathway after selecting for *Meiothermus ruber* in the starch and sucrose metabolism pathway. Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which was used for locating the two selected genes in starch and sucrose metabolism, can be found at http://www.genome.jp/kegg/pathway.html

From the pairwise alignments displayed in Figure 9, *E. coli* b0394 and *Mrub_0759* do not share any of the same highly conserved amino acids throughout their sequences. This pairwise alignment compares the provided sequence to a consensus sequence created by pulling many other protein sequences. Since *E. coli* b0394 and *Mrub_0759* have few similarities in their conserved amino acids, it adds to the evidence that these two genes are not orthologous. Furthermore, this is to be expected since b0394 appears to belong to the ROK family while *Mrub_0759* belongs to the PfkB family.

Panel A
Figure 9. *E. coli* b0394 and Mrub_0759 have several of the same highly conserved amino acids, but not a significant amount. Also it is a larger indicator that the two genes are not orthologous because they do not code for the same domain. Panel A (*E. coli* b0394) codes for the ROK family domain; Panel B (Mrub_0759) codes for the PfkB family domain. These pairwise alignments were created by using the Pfam website at http://pfam.sanger.ac.uk/search

Figure 10 provides more information that these two genes are not orthologous to one another. The differences in color of the genes around them indicate that neither *E. coli* b0394 nor Mrub_0759 belong to an operon. This information is important to note because each gene has the same light-purple color coding which flags them both as being part of carbohydrate metabolism which may lead one to believe they are orthologous.

Figure 10. *E. coli* b0394 and Mrub_0759 genes are not part of an operon. Chromosome Viewer maps, colored by KEGG information displays this evidence. Panel A: *E. coli* b0394 Chromosome Viewer; Panel B: Mrub_0759 Chromosome Viewer. Images were adopted from https://img.jgi.doe.gov

In figure 11 (below) all species displayed in Panel A belong to the phylum, Proteobacteria, and therefore the model gene of *E. coli* b0394 shows no sign of Horizontal Gene Transfer (HGT). In Panel B, the species closest to Mrub_0759 also belong to the same phylum, Deinococcus-
Thermus. However, other distant species are found to belong to different phyla such as Proteobacteria, Dictyoglomi, and Actinobacteria. This indicates that there is HGT between the species closely related to *Meiothermus ruber* which helps indicate a common ancestor.

Panel A

![Phylogenetic tree for E. coli b0394](image)

Panel B

![Phylogenetic tree for Mrub_0759](image)

Figure 11. E. coli b0394 appears to have no Horizontal Gene Transfer (HGT), whereas, Mrub_0759 appears to have (HGT). Panel A: phylogenetic tree for E. coli b0394; Panel B: phylogenetic tree for Mrub_0759. Phylogenetic trees produced by [http://www.phylogeny.fr](http://www.phylogeny.fr)

Table 2 provides a summary of some of the most helpful information provided from the bioinformatics tools used to compare *E. coli b0394* (*mak*) to *Mrub_2365*. The BLAST information in the first row shows that this comparison has an extremely low bit score (16.5) which is not the only data that should be considered to write off the genes as non-orthologous. However, the Expect-value (0.83) is extremely high which indicates that any alignments between the sequences are most likely due to chance. Thus far, it appears that these two genes have no relation, because when the CDD was run they each received different COG numbers (*b0394*(mak)-COG1940) and (*Mrub_2365*-COG0524) however they are predicted to belong to a similar sugar-kinase family. With this information, it is more likely that the genes code for a similar enzyme in starch and sucrose metabolism. According to the bioinformatics tools used to
determine cellular location such as, THM, SignalP, LipoP, and PSORT-B, the enzyme (fructokinase) can be found in the cytoplasm of the cell. Data acquired from TIGRfam had conflicting identification numbers because *M. ruber’s* was unable to be identified due to a high E-value. However, the Pfam test rendered that the two protein sequences belong to two separate domains (PF00480, ROK family) and (PF00294, PfkB family). The enzyme commission number (E.C. 2.7.1.4) also points at the two genes being similar in purpose. They were predicted to be part of the same step of starch and sucrose metabolism in the KEGG pathway information.

**Table 2: b0394 (mak) and Mrub_2365 are not orthologs**

<table>
<thead>
<tr>
<th>Bioinformatics Tool used</th>
<th>E. coli b0394(mak) gene</th>
<th>M. ruber Mrub_2365 gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST <em>E. coli</em> against <em>M. ruber</em></td>
<td>Score: 16.5 bits E-value: 0.83</td>
<td></td>
</tr>
<tr>
<td>CDD Data (COG category)</td>
<td>COG Number: COG1940</td>
<td>COG Number: COG0524</td>
</tr>
<tr>
<td>Cellular Localization</td>
<td>Cytoplasm of the cell</td>
<td></td>
</tr>
<tr>
<td>TIGRfam-protein family</td>
<td>TIGR00744</td>
<td>Indeterminate TIGRfam(E-value too high)</td>
</tr>
<tr>
<td></td>
<td>E-value: 8.6e-14</td>
<td>E-value: 6.7x10^-5</td>
</tr>
<tr>
<td>Pfam- protein family</td>
<td>PF00480 (ROK family)</td>
<td>PF00294 (PfkB)</td>
</tr>
<tr>
<td></td>
<td>E-value: 2.2e-100</td>
<td>E-value: 2.7e-40</td>
</tr>
<tr>
<td>Protein Database</td>
<td>(4U7X) Crystal structure of Fructokinase from <em>Brucella abortus</em> 2308</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E-value: 1.91715E-49</td>
<td>E-value: 6.38167E-20</td>
</tr>
<tr>
<td>Enzyme Commission number</td>
<td>2.7.1.4 –fructokinase</td>
<td></td>
</tr>
<tr>
<td>KEGG pathway map</td>
<td>Starch and sucrose metabolism pathway</td>
<td></td>
</tr>
</tbody>
</table>

Figure 12 (below) represents the BLAST analysis between gene sequences of *E.coli b0394* and *Mrub_2365*. It provides some of the most important information for this project such as amino acid match identity and Expect value. 34% of the amino acids were found to be the same between these two sequences (however, it is important to note that the subject’s sequence is at a shorter length than the query sequence). This information along with the large E-value of 0.83 indicates that there is quite a large probability that these two amino acid sequences aligned due to chance. With this knowledge, it is important to thoroughly investigate all further evidence because just this first set of data indicates that these two genes are not orthologous.
Mrub_2365 and E. coli b0394(mak) do not appear to have any significant similarities among their protein sequences. Query sequence is E. coli b0394(mak); Subject sequence is Mrub_2365. Sequence analysis was performed using NCBI BLAST analyzer at https://blast.ncbi.nlm.nih.gov

The Figure 13 graphs portrayed below, display transmembrane-helices hydropathy results of E. coli b0394 and Mrub_2365. The noticeable red lines in Panel B’s graph represent the appearance of transmembrane helices proteins. However, the heights of these peaks indicate that these possible TMHs are not significant enough to be considered. Therefore, E. coli b0394 or Mrub_2365 have no predicted transmembrane helices which indicate a commonality between the proteins that these two genes code for, because the proteins are both found in the cytoplasm rather than the membrane.

Panel A

No topology graph was able to be formed because of insufficient data
Panel B

Figure 13. Panel A shows the TMHMM results for *E. coli b*0394; Panel B shows the TMHMM results for *Mrub_2365*. *E. coli b*0394 and *Mrub_2365* do not contain any transmembrane-helices; data prediction asserts that the proteins are both found in the cytoplasm of the cells. TMHMM Server v 2.0 was used to determine these two Transmembrane-helices graphs found at http://www.cbs.dtu.dk/services/TMHMM

Figure 14 below, contains SignalP graphs for *E. coli b*0394 and *Mrub_2365*. In SignalP server v. 4.1, the selected proteins are determined a “D-value”. This value is calculated by using the S-score, Y-score, and cutoff value (portrayed by the horizontal purple line). Panel A (*E. coli b*0394) has a D-value of (0.102) which is far below the cutoff of (0.570). This is indicative of the protein, *b_0394* codes for, does not contain any cleavage sites. Panel B (*Mrub_0759*) has similar data showing a D-value of (0.210) but having a cutoff value of (0.570). This information helps conclude that both genes are similar to one another in that they code for proteins which lack cleavage sites.
Figure 14. *E. coli b0394* and *Mrub_2365* are predicted to not contain any cleavage sites according to the bioinformatics tool, SignalP server v. 4.1 found at the following link; http://www.cbs.dtu.dk/services/SignalP. The program uses a cut-off measurement (D-value of 0.570) and both plots show that the genes’ values were below the cut-off. Panel A shows the plot for *E. coli b_0394*; Panel B shows the plot for *Mrub_2365*. 
Figure 15 displays the starch and sucrose metabolism biochemical pathway, that both \textit{E. coli} b0394 and Mrub_2365 are involved in. Enzymes which are highlighted in green are predicted to be present in the selected organism. The pathways show that \textit{E. coli} b0394 and Mrub_2365 are both genes involved in the second to last step of fructose biosynthesis, both coding for the enzyme fructokinase in the starch and sucrose metabolism pathway. This similarity between the genes provides evidence that they are predicted orthologs.

Panel A

Panel B

Figure 15. \textit{E. coli} b0394 and Mrub_2365 are found in the same position in the same biochemical pathway. Panel A displays b0394 in the KEGG pathway after selecting for \textit{Escherichia coli} in the starch and sucrose metabolism pathway. Panel B displays Mrub_2365 in the KEGG pathway after selecting for \textit{Meiothermus ruber} in the starch and sucrose metabolism pathway. Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which was used for locating the two selected genes in starch and sucrose metabolism, can be found at http://www.genome.jp/kegg/pathway.html

From the pairwise alignments displayed in Figure 16 (displayed below), \textit{E. coli} b0394 and Mrub_2365 share none of the same highly conserved amino acids throughout their sequences. This pairwise alignment compares the provided sequence to a consensus sequence created by pulling many other protein sequences. Since \textit{E. coli} b0394 and Mrub_2365 have few similarities in their conserved amino acids, it adds to the evidence that these two genes are not orthologous. Panel A

<table>
<thead>
<tr>
<th>PKD</th>
<th>PKD family</th>
<th>Family</th>
<th>Cl0108</th>
<th>2</th>
<th>302</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>#HHK</td>
<td>#HATCH</td>
<td>#PP</td>
<td>#SEQ</td>
<td>PKD</td>
<td>PKD</td>
<td>PKD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PKD</td>
<td>PKD</td>
<td>PKD</td>
</tr>
<tr>
<td>PKD</td>
<td>PKD family</td>
<td>Family</td>
<td>Cl0108</td>
<td>2</td>
<td>302</td>
<td>3</td>
</tr>
<tr>
<td>PKD</td>
<td>PKD family</td>
<td>Family</td>
<td>Cl0108</td>
<td>2</td>
<td>302</td>
<td>3</td>
</tr>
</tbody>
</table>
Figure 16. *E. coli b0394* and *Mrub_0759* contain none of the same highly conserved amino acids. Also it is a larger indicator that the two genes are not orthologous because they do not code for the same domain. Panel A (*E. coli b0394*) codes for the ROK family domain; Panel B (*Mrub_0759*) codes for the PfkB family domain. These pairwise alignments were created by using the Pfam website at [http://pfam.sanger.ac.uk/search](http://pfam.sanger.ac.uk/search).

Figure 17 (below) provides more information that these two genes are not orthologous to one another. The differences in color of the genes around them indicate that neither *E. coli b0394* nor *Mrub_2365* belong to an operon. This information is important to note because each gene has the same light-purple color coding which flags them both as being part of carbohydrate metabolism which may lead one to believe they are orthologous.

Figure 18, shown below provides essential information on the phylogenetic trees of species most closely related to *Escherichia coli* and *Meiothermus ruber*. All species displayed in Panel A belong to the phylum, Proteobacteria, and therefore the model gene of *E. coli b0394* shows no sign of Horizontal Gene Transfer (HGT). In Panel B, the species closest to Mrub_2365 also belong to the same phylum as itself, Deinococcus-Thermus, however, other closely related
species belong to varying phyla, such as; Firmicutes, Thermotogae, Actinobacteria, and Proteobacteria. This indicates that there is HGT between the species closely related to *Meiothermus ruber*, which means there is a possibility of a common ancestor.

Panel A

![Phylogenetic tree for E. coli b0394](image1.png)

Panel B

![Phylogenetic tree for Mrub_2052](image2.png)

Figure 18. E. coli b0394 appears to have no Horizontal Gene Transfer (HGT), whereas, and Mrub_2365 does appear to have (HGT). Panel A: phylogenetic tree for E. coli b0394; Panel B: phylogenetic tree for Mrub_2365. Phylogenetic trees produced by [http://www.phylogeny.fr](http://www.phylogeny.fr)

Table 3 provides a summary of some of the most helpful information provided from the bioinformatics tools used to compare *E. coli b0688 (mak)* to *Mrub_2052*. The BLAST information in the first row shows that this comparison has a bit score of (698) which is not the
only data that should be considered to write off the genes as orthologous. Furthermore, the Expect-value (0.00) indicates that any alignments between the sequences have a high probability that they are not due to chance. Thus far, it appears that these two genes have a relation, because when the CDD was run they each received the same COG number (COG0033) and are predicted to have the same COG name of Phosphoglucomutase meant for Carbohydrate transport and metabolism. With this information, it is more likely that the two genes code for a similar enzyme in starch and sucrose metabolism. According to the bioinformatics tools used to determine cellular location such as, THM, SignalP, LipoP, and PSORT-B, the enzyme (phosphoglucomutase) can be found in the cytoplasm of the cell. Data acquired from TIGRfam had the same identification number (TIGR01132) for each protein sequence. The Pfam test rendered that the two protein sequences belong to the same domain (PF02878, Phosphoglucomutase, alpha/beta/alpha domain I). The enzyme commission number (E.C. 5.4.2.2) also points at the two genes being similar in purpose. They were predicted to be part of the same secondary step of D-Glucose production in starch and sucrose metabolism in the KEGG pathway information.

**Table 3: b0688(pgm) and Mrub_2052 are orthologs**

<table>
<thead>
<tr>
<th>Bioinformatics Tool used</th>
<th>E. coli b0688(pgm)</th>
<th>M. ruber Mrub_2052</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST E. coli against M. ruber</td>
<td>Score: 698 bits E-value: 0.0</td>
<td></td>
</tr>
<tr>
<td>CDD Data (COG category)</td>
<td>COG Number: COG0033 Phosphoglucomutase [Carbohydrate transport and metabolism]</td>
<td>E-value: 0.0 E-value: 0.0</td>
</tr>
<tr>
<td>Cellular Localization</td>
<td>Cytoplasm of the cell</td>
<td>TIGR01132</td>
</tr>
<tr>
<td>TIGRfam-protein family</td>
<td></td>
<td>E-value: 0.00 E-value: 0.00</td>
</tr>
<tr>
<td>Protein Database</td>
<td>(2FUV) Phosphoglucomutase from Salmonella typhimurium</td>
<td>E-value: 0.00 E-value: 0.00</td>
</tr>
<tr>
<td>Enzyme Commission number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEGG pathway</td>
<td>Starch and sucrose metabolism pathway</td>
<td></td>
</tr>
</tbody>
</table>

Figure 19 (below) represents the BLAST analysis between gene sequences of *E.coli b0688* and *Mrub_2052*. It provides some of the most important information for this project such as amino acid match identity and Expect value. 63% of the amino acids were found to be the same
between these two sequences (it also should be noted that the subject’s sequence length is almost identical to that of the query sequence). This information along with the E-value of 0.00 indicates that it is highly probable that these two gene sequences did not align solely due to chance. With this knowledge, it is important to thoroughly investigate all further evidence because just this first set of data indicates that these two genes are orthologous.

Figure 19. Mrub_2052 and E. coli b0688(pgm) do appear to have significant similarities among their protein sequences. Query sequence is E. coli b0688; Subject sequence is Mrub_2052. Sequence analysis was performed using NCBI BLAST analyzer at https://blast.ncbi.nlm.nih.gov

The Figure 20 graphs portrayed below, display transmembrane-helices hydrophathy results of E. coli b0688 and Mrub_2052. Both graphs in Panels A and B have no noticeable vertical red lines. These seemingly blank graphs indicate that neither E. coli b0688 nor Mrub_2052 have predicted transmembrane helices. This evidence indicates a commonality between the proteins that these two genes code for, because the proteins are both found in the cytoplasm rather than the membrane.
Panel A

Panel B

Figure 20. Panel A shows the TMHMM results for *E. coli* b0688; Panel B shows the TMHMM results for *Mrub_2052*. *E. coli* b0688 and *Mrub_2052* do not contain any Transmembrane-helices; data prediction asserts that the proteins are both found in the cytoplasm of the cells. TMHMM Server v.2.0 was used to determine these two transmembrane-helices graphs found at [http://www.cbs.dtu.dk/services/TMHMM](http://www.cbs.dtu.dk/services/TMHMM)

Figure 21 below, contains SignalP graphs for *E. coli* b0688 and *Mrub_2052*. In SignalP server v. 4.1, the selected proteins are given a “D-value”. This value is calculated by using the; C-score(signal peptide cleavage site), S-score(signal peptide position), Y-score(combined C and S scores), and cutoff value (portrayed by the horizontal purple line). Panel A (*E. coli* b0688) has a D-value of (0.096) which is far below the cutoff of (0.570). This is indicative that the protein, b_0688 codes for, does not contain any cleavage sites or signal peptides. Panel B (*Mrub_2052*)
has similar data showing a D-value of (0.164) but having a cutoff value of (0.570). This information helps conclude that the two genes are potential orthologs because the proteins they both code for, lack signal peptides and cleavage sites.

Panel A

Figure 21. *E. coli b*0688 and *Mrub_2052* are predicted to not contain any cleavage sites according to the bioinformatics tool, SignalP server v. 4.1 found at the following link; http://www.cbs.dtu.dk/services/SignalP. The program uses a cut-off measurement (D-value of 0.570) and both plots show that the genes’ values were below the cut-off. Panel A shows the plot for *E. coli b*_0688; Panel B shows the plot for *Mrub_2052*. 
Figure 22 displays the starch and sucrose metabolism biochemical pathway, that both *E. coli* b0688 and *Mrub_2052* are involved in. Enzymes which are highlighted in green are predicted to be present in the selected organism. The pathways show that *E. coli* b0688 and *Mrub_2052* are both genes involved in the second to last step of D-Glucose-6P biosynthesis, both coding for the enzyme, phosphoglucomutase, in the starch and sucrose metabolism pathway indicated by the E.C. number 5.4.2.2. This similarity between the genes provides evidence that they are predicted orthologs.

Panel A

Panel B

Figure 22. *E. coli* b0688 and *Mrub_2052* are found in the same position in the same biochemical pathway. Panel A displays *b_0688* in the KEGG pathway after selecting for *Escherichia coli* in the starch and sucrose metabolism pathway. Panel B displays *Mrub_2052* in the KEGG pathway after selecting for *Meiothermus ruber* in the starch and sucrose metabolism pathway. The genes were found at the same E.C. number (5.4.2.2) indicating that they help code for a similar enzyme (phosphoglucomutase) and its production of D-Glucose-6P. Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which was used for locating the two selected genes in starch and sucrose metabolism, can be found at [http://www.genome.jp/kegg/pathway.html](http://www.genome.jp/kegg/pathway.html)

From the pairwise alignments displayed in Figure 23 (displayed below), it can be seen that *E. coli* b0688 and *Mrub_2052* share regions of the same highly conserved amino acids throughout their sequences. This pairwise alignment compares the provided sequence to a consensus sequence created by pulling many other protein sequences. Since *E. coli* b0688 and *Mrub_2052* have similarities in their regions of conserved amino acids (regions outlined in red), it adds to the evidence that these two genes are predicted to be orthologous.
Figure 23. *E. coli* b0688 and Mrub_2052 have regions of the same highly conserved amino acids (highlighted in red). This analysis is a large indicator that the two genes are orthologous because they also code for proteins in the same domain, Phosphoglucomutase/phosphomannomutase, alpha/beta/alpha domain I. These pairwise alignments were created by using the Pfam website at [http://pfam.sanger.ac.uk/search](http://pfam.sanger.ac.uk/search).

Figure 24 (below) provides more information that these two genes are orthologous to one another. The differences in color of the genes around them indicate that neither *E. coli* b0688 nor Mrub_2052 belong to an operon. This information is important to note because each gene has the same light-purple (salmon) color coding which flags them both as being part of carbohydrate metabolism, which is evidence for them being orthologous.
Figure 24. *E. coli* b0688 and *Mrub_2052* genes are not part of an operon. However, salmon color indicates both genes are involved in carbohydrate metabolism. This evidence helps with the prediction that the genes are orthologous. Chromosome Viewer maps, colored by KEGG information displays this evidence. Panel A: *E. coli* b0688 Chromosome Viewer; Panel B: *Mrub_2052* Chromosome Viewer. Images were adopted from https://img.jgi.doe.gov

Figure 25, shown below, provides essential information on the phylogenetic trees of species most closely related to *Escherichia coli* and *Meiothermus ruber*. All species displayed in Panel A belong to the phylum, Proteobacteria, and therefore the model gene of *E. coli* b0688 shows no sign of Horizontal Gene Transfer (HGT). In Panel B, the species closest to *Mrub_2052* also belong to the same phylum as itself, Deinococcus-Thermus. There are only slight differences in nearby species belonging different phylum, such as Chloroflexi and Proteobacteria. This indicates that there is also no existent HGT between the species most closely related to *Meiothermus ruber*.
Panel B


Table 4 provides a summary of some of the most helpful information provided from the bioinformatics tools used to compare *E. coli* b0688 (ycjM) to *Mrub_3029*. The BLAST information in the first row shows that this comparison has a bit score of (486) which is not the only data that should be considered to verify the genes as orthologous. Furthermore, the Expect-value (9e-172) indicates that any alignments between the sequences have a high likelihood that they are not due to chance. Thus far, it appears that these two genes have a relation, because when the CDD was run they each received the same COG number (COG0366) and are predicted to have the same COG name of Glycosidase meant for Carbohydrate transport and metabolism. With this information, it is more likely that the two genes code for a similar enzyme in starch and sucrose metabolism. According to the bioinformatics tools used to determine cellular location such as, THM, SignalP, LipoP, and PSORT-B, the enzyme (Sucrose phosphorylase) can be found in the cytoplasm of the cell. Data acquired from TIGRfam had the same identification number (TIGR03852) for each protein sequence. The Pfam test rendered that the two protein sequences belong to the same domain (PF00128, Alpha-Amylase). The enzyme commission number (E.C. 2.4.1.7) also points at the two genes being similar in purpose. They were predicted to be part of the same secondary step of D-Fructose production in starch and sucrose metabolism as determined in the KEGG pathway information.
Table 4: *b1309*(ycjM) and *Mrub_3029* are orthologs

<table>
<thead>
<tr>
<th>Bioinformatics Tool used</th>
<th><em>E. coli</em> <em>b1309</em>(ycjM) gene</th>
<th><em>M. ruber</em> <em>Mrub_3029</em> gene</th>
</tr>
</thead>
</table>
| BLAST *E. coli* against *M. ruber* | Score: 486 bits  
E-value: 9e-172 | |
| CDD Data (COG category) | COG Number: COG0366  
Glycosidase [Carbohydrate transport and metabolism] | |
| Cellular Localization | Cytoplasm of the cell | |
| TIGRfam-protein family | TIGR03852 | |
| Pfam-protein family | PF00128  
Alpha-Amylase | |
| Protein Database | (3UEQ) Crystal structure of amylosucrase  
from *Neisseria polysaccharea* in complex  
with turanose | |
| Enzyme Commission number | 2.4.1.7-Sucrose phosphorylase | |
| KEGG pathway map | Starch and sucrose metabolism pathway | |

Figure 26 (below) represents the BLAST analysis between gene sequences of *E. coli* *b1309* and *Mrub_3029*. It provides some of the most important information for this project such as amino acid match identity and Expect value between sequences. 47% of the amino acids were found to be the same between these two sequences (it also should be noted that the subject’s sequence length is almost identical to that of the query sequence). This information along with the E-value of 9e-172 indicates that it is highly probable that these two gene sequences did not align solely due to chance. With this knowledge, it is important to thoroughly investigate all further evidence because just this first set of data indicates that these two genes can be predicted as orthologous.
Figure 26. Mrub_3029 and E. coli b1309(ycjM) do appear to have significant similarities among their protein sequences. Query sequence is E. coli b1309; Subject sequence is Mrub_3029. Sequence analysis was performed using NCBI BLAST analyzer at https://blast.ncbi.nlm.nih.gov

The Figure 27 graphs portrayed below, display transmembrane-helices hydropathy results of E. coli b1309 and Mrub_3029. The graph in Panel A shows a small peak of red, which indicates there is a transmembrane protein present, however, it is not significant enough to cross the horizontal pink line(indicating the outside of the cell) therefore, it cannot be considered a TMH. Panel B has no noticeable vertical red lines. These seemingly blank graphs indicate that neither E. coli b1309 nor Mrub_3029 have predicted transmembrane helices. This evidence indicates a commonality between the proteins that these two genes code for, because the proteins are both found in the cytoplasm rather than the membrane.

Panel A
Figure 27. Panel A shows the TMHMM results for *E. coli b_1309*; Panel B shows the TMHMM results for *Mrub_3029*. *E. coli b1309* and *Mrub_3029* do not contain any Transmembrane-helices; data prediction asserts that the proteins are both found in the cytoplasm of the cells. TMHMM Server v 2.0 was used to determine these two transmembrane-helices graphs found at [http://www.cbs.dtu.dk/services/TMHMM](http://www.cbs.dtu.dk/services/TMHMM)
Figure 28 (below), contains SignalP graphs for *E. coli* *b1309* and *Mrub_3029*. In SignalP server v. 4.1, the selected proteins are given a “D-value”. This value is calculated by using the; C-score(signal peptide cleavage site), S-score(signal peptide position), Y-score(combined C and S scores), and cutoff value (portrayed by the horizontal purple line). Panel A (*E. coli b1309*) has a D-value of (0.132) which is far below the cutoff of (0.570). This is indicative that the protein *b1309* codes for, does not contain any cleavage sites or signal peptides. Panel B (*Mrub_3029*) has similar data showing a D-value of (0.101) but having a cutoff value of (0.570). This information helps conclude that the two genes are potential orthologs because the proteins they both code for, lack signal peptides and cleavage sites.

Panel A

Panel B
Figure 28. *E. coli b1309* and *Mrub_3029* are predicted to not contain any cleavage sites according to the bioinformatics tool, SignalP server v. 4.1 found at the following link; [http://www.cbs.dtu.dk/services/SignalP](http://www.cbs.dtu.dk/services/SignalP). The program uses a cut-off measurement (D-value of 0.570) and both plots show that the genes’ values were below the cut-off. Panel A shows the plot for *E. coli b1309*; Panel B shows the plot for *Mrub_3029*.

Figure 29 displays the starch and sucrose metabolism biochemical pathway, that both *E. coli b1309* and *Mrub_3029* are involved in. Enzymes which are highlighted in green are predicted to be present in the selected organism. The pathways show that *E. coli b1309* and *Mrub_3029* are both genes involved in the second to last step of D-Fructose biosynthesis, both coding for the enzyme, Sucrose phosphorylase, in the starch and sucrose metabolism pathway indicated by the E.C. number 2.4.1.7. This similarity between the genes provides evidence that they are predicted orthologs.

**Panel A**

**Panel B**

Figure 29. *E. coli b1309* and *Mrub_3029* are found in the same position in the same biochemical pathway. Panel A displays *b1309* in the KEGG pathway after selecting for *Escherichia coli* in the starch and sucrose metabolism pathway. Panel B displays *Mrub_3029* in the KEGG pathway after selecting for *Meiothermus ruber* in the starch and sucrose metabolism pathway. The genes were found at the same E.C. number (2.4.1.7) indicating that they help code for a similar enzyme (Sucrose phosphorylase) and its production of D-Fructose. Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which was used for locating the two selected genes in starch and sucrose metabolism, can be found at [http://www.genome.jp/kegg/pathway.html](http://www.genome.jp/kegg/pathway.html)
From the pairwise alignments displayed in Figure 30 (displayed below), it can be seen that *E. coli b1309* and *Mrub_3029* share regions of the same highly conserved amino acids throughout their sequences. This pairwise alignment compares the provided sequence to a consensus sequence created by pulling many other protein sequences. Since *E. coli b1309* and *Mrub_3029* have similarities in their regions of conserved amino acids (regions outlined in red), it adds to the evidence that these two genes are predicted to be orthologous.

Panel A

![Pairwise Alignment](image)

Panel B

![Pairwise Alignment](image)

Figure 30. *E. coli b1309* and *Mrub_3029* have regions of the same highly conserved amino acids (highlighted in red). This analysis is a large indicator that the two genes are orthologous because they also code for proteins in the same domain, Alpha-Amylase. These pairwise alignments were created by using the Pfam website at [http://pfam.sanger.ac.uk/search](http://pfam.sanger.ac.uk/search)

Figure 31 (below) provides more information that these two genes are orthologous to one another. The differences in color of the genes around them indicate that neither *E. coli b1309* nor *Mrub_3029* belong to an operon. This information is important to note because each gene has the same light-purple (salmon) color coding which flags them both as being part of carbohydrate metabolism, which is evidence for them being orthologous.

Panel A

![Genes Comparison](image)
Figure 31. *E. coli* b1309 and *Mrub_3029* genes are not part of an operon. However, salmon color indicates both genes are involved in carbohydrate metabolism. This evidence helps with the prediction that the genes are orthologous. Chromosome Viewer maps, colored by KEGG information displays this evidence. Panel A: *E. coli* b1309 Chromosome Viewer; Panel B: *Mrub_3029* Chromosome Viewer. Images were adopted from https://img.jgi.doe.gov

Figure 32, shown below, provides essential information on the phylogenetic trees of species most closely related to *Escherichia coli* and *Meiothermus ruber*. All species displayed in Panel A belong to the phylum, Proteobacteria, and therefore the model gene of *E. coli* b1309 shows no sign of Horizontal Gene Transfer (HGT). In Panel B, the species closest to *Mrub_3029* also belong to the same phylum as itself, Deinococcus-Thermus. However, there are differences in nearby species belonging different phylum, such as Chloroflexi and Firmicutes, and Proteobacteria. This indicates that there is HGT between the species closely related to *Meiothermus ruber* which hints to the presence of a common ancestor.
Figure 32. E. coli b1309 appears to have no Horizontal Gene Transfer (HGT), whereas, Mrub_3029 appears to have (HGT). Panel A: phylogenetic tree for E. coli b_1309; Panel B: phylogenetic tree for Mrub_3029. Phylogenetic trees produced by [http://www.phylogeny.fr](http://www.phylogeny.fr)

Conclusion

From the results of this study the following can be concluded; b0394(mak) and Mrub_0759 are not orthologous, b0394(mak) and Mrub_2365 are not orthologous, Mrub_0759 and Mrub_2365 are paralogs, b0688(pgm) and Mrub_2052 are orthologous, and b1309(ycjM) and Mrub_3029 are orthologous. Initially, evidence to determine these genetic linkages were found by using NCBI’s BLAST program to compare the amino acid sequences between each depicted E. coli and M. ruber gene pair. This program showed that it is highly unlikely that b0394 (mak) and Mrub_0759 (E-value: 0.37, Bit Score: 17.7) and b0394 (mak) and Mrub_2365 (E-value: 0.83, Bit Score: 16.5) are orthologous pairs. However, contradictory data was found in the other Bioinformatics tools that enhance cellular localization. Trans-membrane-helices program ([http://www.cbs.dtu.dk/services/TMHMM](http://www.cbs.dtu.dk/services/TMHMM)), SignalP ([http://www.cbs.dtu.dk/services/SignalP](http://www.cbs.dtu.dk/services/SignalP)), LipoP (Juncker et al., 2003) indicated that all three genes were found in the cytoplasm and none coded for TMHs or signal-peptides. This indicates a commonality in the genes because they possibly code for proteins with similar structure. Interestingly, the discrepancy in data was found upon a closer look into the KEGG website. In the pathway map for both E. coli and M. ruber The Enzyme Commission Number of each gene was found to be the same (2.7.1.4), however, when searched deeper into the KEGG database another E.C. number was found which was (2.7.1.2) which labels the glucokinase protein family. Further investigation should be performed about this discrepancy because if it is not confronted, it could lead to further information and misinterpretation of orthologous pairs such as in this study. To add a secondary source as to the purpose of b0394, a search was run through the Metacyc database. Here it was found that b0394 is responsible for coding for fructokinase which further indicates that the genes all code for the same enzyme (specifically; the biosynthesis of Fructokinase in the starch and sucrose metabolism) ([http://www.expasy.ch/enzyme/enzyme-search-ec.html](http://www.expasy.ch/enzyme/enzyme-search-ec.html)). TIGRfam results displayed
that the genes belonged to the same protein family (Haft et al., 2001). Finally, the phylogenetic
trees produced for each species *M. ruber* and *E. coli* indicate the genes are likely orthologous
because Horizontal-Gene-Transfer did not occur in these species indicating that the proteins the
desired genes code for received their functions from a common ancestor rather than receiving a
new function from outside their original lineage (http://www.phylogeny.fr). With all of this
contradictory evidence, one must be skeptical of any concrete final annotations. It would be
beneficial in future research to run more NCBI BLAST tests when a more complete *M. ruber*
genome is formed in order to obtain better amino acid sequence comparisons with *E. coli.*
However, these test results showed surprising similarity between *Mrub_0759* and *Mrub_2365,*
and identical evidence was found in the cellular localization tests, as well as protein family
analyses (TIGRfam and Pfam) to give reason to believe that these two genes were paralogs
(Lefers 2004). The two *M. ruber* genes were BLASTed against one another and received a
percent identity of 33% which is considered high in a sequence alignment of
306(http://www.geni-act.org). Knowing that these two genes are paralogs will help in future
research when further studying the biosynthesis process of Fructokinase because *Mrub_0759* and
*Mrub_2365* can be compared for similar structure and function.

The remaining gene pairs *b0688(pgm)/Mrub_2052* and *b1309(ycjM)/Mrub_3029* were
predicted as orthologs by the same process of initially using NCBI’s BLAST program to
compare amino acid sequence alignment. *Mrub_2052* and *b1309(ycjM)* (E-value:0.0, Bit Score:
698), *b1309(ycjM)* and *Mrub_3029* (E-value: 9e-172, Bit Score: 486). Having such low E-
values and adequate bit scores indicates that was highly unlikely that each amino acid sequence
pair aligned by chance, and prompts further exploration using bioinformatics to determine any
similarities the *E. coli* and *M. ruber* genes may share. All of these genes were found to code for
proteins found in the cytoplasm by using bioinformatics tools for cellular localization; TMH
(http://www.cbs.dtu.dk/services/TMHMM), SignalP (http://www.cbs.dtu.dk/services/SignalP),
and LipoP (Juncker et al., 2003). This is a commonality that indicates gene orthology because it
is likely that the genes had a common ancestor which produced proteins in the same area of the
cell. The Enzyme Commission Number of each gene was also found to be the same; pair
*b0688(pgm)/Mrub_2052* found to be 5.4.2.2 (phosphoglucomutase) and pair *b1309(ycjM)/
Mrub_3029* found to be 2.4.1.7 (Sucrose phosphorylase). This information further indicates that
the gene pairs are likely orthologous because these gene functions are found in both the *E.coli*
and *M. ruber* species (http://www.expasy.ch/enzyme/enzyme-search-ec.html). TIGRfam and
Pfam results displayed that the genes belonged to the same protein families; *b_0688(pgm)/
Mrub_2052: Phosphoglucomutase domain 1; b1309(ycjM)/ Mrub_3029: Alpha-Amylase* (Haft et
al., 2001). Lastly, the phylogenetic trees produced for each species *M. ruber* and *E. coli* indicate
the genes are likely orthologous because Horizontal-Gene-Transfer did not occur in these species
indicating that the proteins the desired genes code for received their functions from a common
ancestor rather than receiving a new function from outside their original lineage
(http://www.phylogeny.fr). None of the acquired evidence appears to be contradictory, and
based on the bioinformatics results the prediction that gene pairs *b0688(pgm)/Mrub_2052* and
*b1309(ycjM)/Mrub_3029* are orthologous pairs providing evidence that future studies on these
*M. ruber* can be compared with the *E.coli* genes when determining similar functions or structures
due to a common ancestor.
Site-Directed-Mutagenesis (*Mrub_2052*)

Figure 33 displays the most highly conserved amino acid that *Mrub_2052* codes for (which appears to be Aspartate(D)). It is known that it is the highest conserved because bioinformatics tools HMMlogo and Pfam pairwise alignment were used to compare the amino acid’s prevalence across species. Pfam pairwise alignment helps display highly conserved amino acids in the consensus sequence used to generate the HMM logo, with capital letters. If the amino acids in the query sequence are identical to those in the consensus sequence, the second row in the pairwise alignment will display these (Sonnhammer, 1997). HMMlogo helps further determine the conservation of the amino acid which hints to common ancestry because it runs the amino acid sequence with other enzymes found within the same protein family (http://pfam.sanger.ac.uk/search).

Panel A

Panel B

Figure 33. Highly conserved amino acid found in the HMMlogo data for the gene *Mrub_2052*. Panel A: (D) Aspartate is portrayed as being the most highly conserved amino acid at position 48 in the HMM alignment program. Panel B: (D) Aspartate is also found to be the most conserved at position 48 in the Pfam pairwise alignment program as displayed by the capital letter (D) in the second row.

Figure 34 (below) displays the deletion mutation of Aspartate amino acid(48) in *Mrub_2052* that is coded for by the nucleotide sequence (CGG). Aspartate was the highest conserved amino acid when run through WEBlogo (Crooks *et al*. 2004). This makes sense because it is a polar and is important for protein binding sites such as forming salt-ion bridges (Betts *et al*. 2003). Because of its importance it can be predicted that if the nucleotid sequence found at positions 142-144bp is
deleted, the protein will lose its function because Aspartate is a main factor in phosphoglucomutase’s function to facilitate the conversion of glucose-1-phosphate into glucose-6-phosphate during Glucose’s metabolic process (www. http://oregonstate.edu). The predicted primers would insert Alanine in substitution for the deletion of Aspartate to inhibit a frame-shift mutation from occurring. This enables a more distinct analysis of the loss-of-function mutation (http://nebasechanger.neb.com/).

Figure 34. Primers designed to perform site directed mutagenesis on Aspartate 48 in Mrub_2052. The primers designed would replace Aspartate with Alanine which would permit for a loss of function mutation. Images adopted from http://nebasechanger.neb.com/


Lasser JP; Beyne E; Pyndiah S; Lapaillerie D; Claverol S; Bonneu M. A complexomic study of Escherichia coli using two-dimensional blue native/SDS polyacrylamide gel electrophoresis. 2006 Aug;27(16):3306-21: [2017 Feb].

Lefers Mark. 2004. Paralogous Genes: Holmgren Labs. N/A


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:


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


Scott LR. Meiothermus ruber Genome Analysis Project. [Internet]. GENI-ACT; [2015 Dec 16].

Available from:http://geni-science.org/secure/projects/vi
