2018

Mrub_2836, Mrub_1595, and Mrub_1596 are orthologs of b_1857, b_1859 and b_1858 in *Escherichia Coli* Coding for a Zinc Uptake ABC Transporter System

Austin J. Dollmeyer
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_2836, Mrub_1595, and *Mrub_1596* are Orthologs of *b_1857, b_1859,* and *b_1858* in *Escherichia Coli* coding for a Zinc Uptake ABC Transporter System

By: AJ Dollmeyer

INTRODUCTION

Why Study *Meiothermus Ruber*?

Meiothermus Ruber (*M. ruber*) is a red-pigmented, Gram-negative, thermophilic bacterium (Tindall *et al*., 2010). *M. ruber* was first discovered in Kamchatka, Russia in 1975 from natural hot springs and other thermal environments and is from the *Meiothermus* genus (Tindall *et al*., 2010, Loginova *et al*., 1975). Even though *M. ruber* was discovered in 1975, it is still relatively unstudied, especially when compared to other bacteria like *E. coli.* For example, it took until 1996, 21 years after it was discovered until it was finally given its correct name and change it from *Thermus ruber* to *Meiothermus ruber*. Because of this lack of research on *M. ruber*, there is still a lot we do not know about the genome of this organism. *M. ruber* is one of many bacteria organisms that are relatively unstudied and unknown (Scott). Because of this, the Joint Genome Institute (JGI) started a program called the Genomic Encyclopedia of Bacteria and Archaea (GEBA). The purpose of the GEBA project is to study these more understudied organisms, pool newly found information about them, and filling in knowledge gaps of highly studied organisms by identifying protein families, to understanding the evolutionary history of different microbial organisms (JGI). This means that it is very important to research understudied organisms like *M. ruber*, so we can, not only better understand relatively unknown organisms, but also better understand well-studied organisms, like *E. coli*. In this study, we try to better understand the *M. ruber* genome using *E. coli* as a control organism. The three *M. ruber* genes being studied are *Mrub_1595*, *Mrub_1596*, and *Mrub_2836* which are believed to be genes for ABC transporters that uptake zinc.

E. coli as a Model Organism

As stated earlier, *E. coli* will be used as a model organism for this research. *E. coli* is a good model organism to use because it is one of the most thoroughly studied organisms to date, it is a relatively simplistic organism, and can be easily be produced and studied in laboratory settings (Cooper 2000). The *E. coli K-12* strain was also completely sequenced in 1997, making it a very reliable and useful model organism (Moussatova *et al*., 2008). By doing a quick BLAST of *Mrub_1595*, *Mrub_1596*, and *Mrub_2836* shows that these genes have similar sequences to Zinc ABC transporter genes in *E. coli*, meaning that there might be ortholog genes in *E. coli* to the *M. ruber* genes we are interested in. This makes *E. coli* a highly desirable model organism for this research.

ABC Transporters

In general, ABC transporters are membrane proteins that are constantly transporting organic and non-organic molecules in and out of a cell against a concentration gradient using ATP (Moussatova *et al*., 2008). Because of their function, ABC transporters are extremely important to the cells and in the medical field. ABC transporters are constructed of two transmembrane domains, which make up the transport channel, and two nucleotide binding domains, which bind and hydrolyze ATP, allowing for transport of molecules. In prokaryotic organisms, like *E. coli* and *M. ruber*, ABC transporters are importers which means that substrate binding proteins, which determine directionality of the transporter, are also required to recruit substrates to the system (Moussatova *et al*., 2008). More specifically, the *E. coli* genes, b_1857 (*ZnuA*), b_1858 (*ZnuC*), and b_1859 (*ZnuB*) are a part of an ATP-dependent Zn^{2+} uptake system (Patzer *et al*., 1998). More specifically, ZnuA is the periplasmic binding protein of the system (Yatsunyk *et al*., 2007). This means that the Zn^{2+} ions bind to this portion of the Zn uptake system in the periplasm. ZnuB is the inner membrane transporter in the system (Yatsunyk *et al*, 2007). This means that ZnuB's job is to transport the Zn^{2+} that binds to ZnuA and transport it across the cytoplasmic membrane. ZnuC is the ATP-binding subunit of the Zn uptake system, meaning that it will couple ATP hydrolysis to the system allowing for the transport of Zn^{2+} ions. Because all three of these genes are a part of the same transporter system, it is believed that these three genes in both *E. coli* and *M. ruber* are also apart of operons with each other. As you can see in Figure 1, Zn^{2+} binds to the ZnuA subunit of the system. When this occurs, ZnuC will hydrolyse ATP, which in turn will give off energy for the system, and will produce ADP, inorganic phosphate, and H^+ ions. When that occurs, ZnuB will transport the Zn^{2+} ion from the periplasm to the cytosol in the cell. Because of the way the system moves the Zn^{2+} ions, from periplasm to cell cytosol, it is clearly an importer and not an exporter ABC transporter.

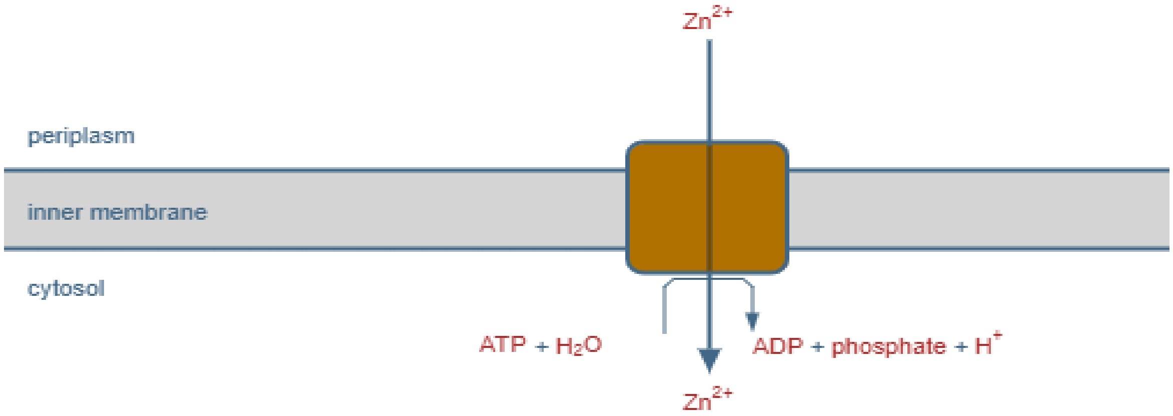


Figure 1. Zn^{2+} transport from periplasm to cytosol by the Znu transport system. Image taken from https://ecocyc.org/gene?orgid=ECOLI&id=ZNUC-MONOMER#tab=RXNS

Zinc is a very important element to all organisms. It is a very essential element because it acts a catalytic cofactor for hundreds of enzymes and proteins with many different functions (Patzer *et al*., 1998). Zinc has long been believed to be critical in all organisms in growth and reproductive factors, but that is not all. Zinc has also been found to protect biological structures, like DNA, from oxidative stress (Stefanidou *et al*., 2005) This is very important to note because of the harsh conditions that Meiothermus ruber has been found in, could cause oxidative stress, and these zinc transporters could indicate the reason *Meiothermus Ruber* is able to survive in these harsh conditions. But these are only a few roles zinc has been identified to play in organisms. Zinc has also been thought to play roles in immune responses, ageing, apoptosis, and even an antioxidant (Stefanidou *et al*., 2005). Zinc plays so many roles in organisms, is vital in almost all organisms, therefore, it is important we study zinc transporters to see more uses for zinc. This paper will focus specifically on *Meiothermus Ruber* because it may indicate the roles of zinc in this organism and how it is able to survive in stressful environments. Figure 2 shows possible paralogs of our *M. ruber* genes of interest. The closer the color is to red, the lower the E-value indicating very similar sequences. This figure shows that a similar sequence and a low E-score is not enough to prove genes are orthologs. Therefore, we will be using many bioinformatics tools to confirm that.

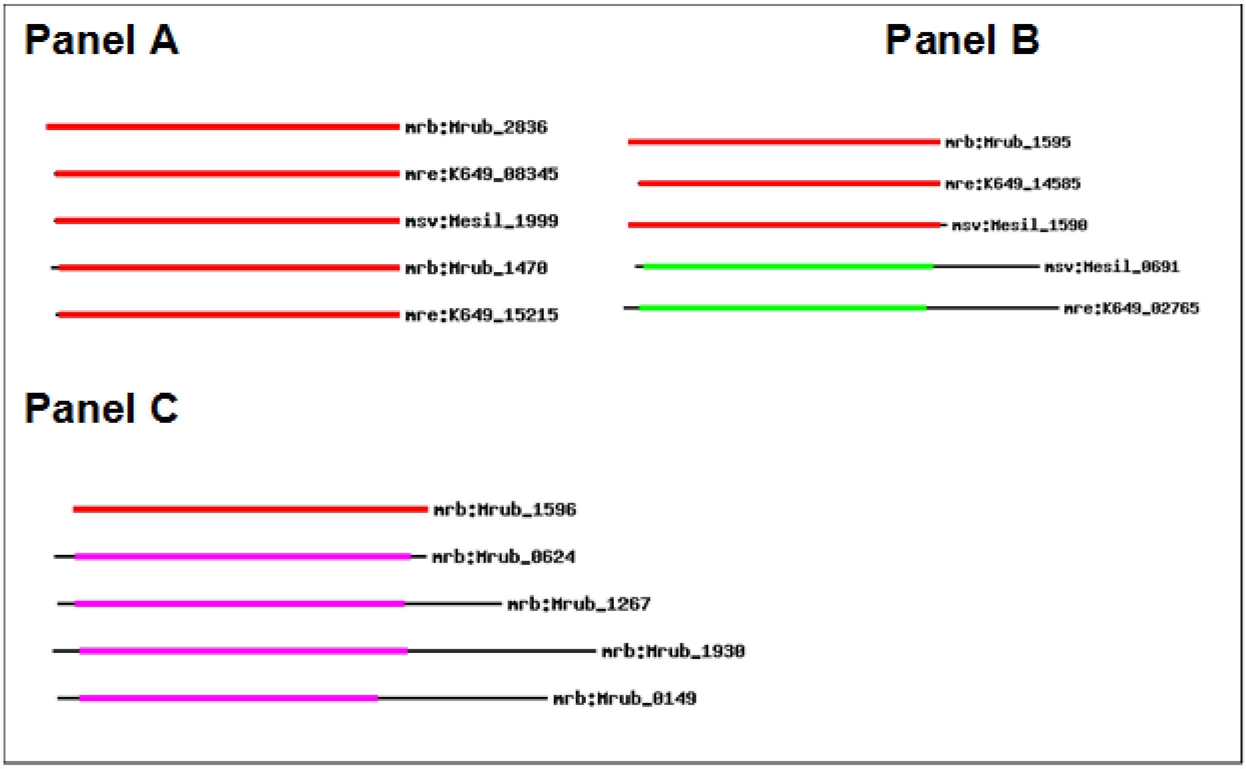


Figure 2: Paralogs of *M. ruber* genes of Interest *Mrub_2836* (Panel A), *Mrub_1595* (Panel B), and *Mrub_1596* (Panel C). Images taken from http://www.genome.jp/kegg/.

Purpose/Hypothesis:

In this paper, determination of *Mrub_2836*, *Mrub_1595*, and *Mrub_1596* genes as orthologs of the *E. coli* genes b_1857 (ZnuA), b_1859 (ZnuB), and b_1858 (ZnuC) respectively. The use of bioinformatics tools such as BLAST, KEGG, EcoCyc, and many others will be used to indicate similarities and differences in the suspected orthologs and help determine if this hypothesis is correct or not. This hypothesis came from an initial BLAST search of ZnuA, ZnuB, and ZnuC against the Meiothermus ruber genome. When this was done, BLAST indicated very low e-values of 6e-19, 6e-31, and 9e-39 respectively. Because of these low e-values, we hypothesized that these genes are orthologs of one another, but more research is needed to confirm this hypothesis.

Materials/Methods

To start the research, a KEGG pathway search (Kanehisa et al., 2016) was done on ABC transporters focusing on both Meiothermus ruber and *E. coli*. This gave a starting point with the names, protein sequences, and pathways of our genes of interest. Now that we had our genes of interest for *M. ruber* and *E. coli*, a BLAST search (Madden et al., 2003) of each *E. coli* gene protein sequence against the Meiothermus ruber genome. This gave genes in the Meiothermus ruber genome that had similar protein sequences to the *E. coli* gene. The lowest E-value genes were taken as this indicated similar genes. Once *M. ruber* genes with similar sequences to our *E. coli* genes were identified, an IMG/M bioinformatics tool search (Markowitz et al., 2012) was done using gene locus tags. Once this was done, the sequence viewer for alternate ORF search was done to confirm the correct starting amino acid. To further confirm this, another BLAST search was done of our *M. ruber* gene of interest was done and 15 of the top 20 results protein sequences were taken and input into the T-Coffee bioinformatics tool. This aligns all the different protein sequences as closely as possible. The output from T-Coffee was then input into WebLogo (Crooks et al., 2004). This tool outputs an image of each amino acid and how retained every amino acid was retained throughout the different species. A good retention value at the first amino acid site indicates a correct start point in our gene of interest. This was not done for *E. coli* genes as they are very well studied and starting points are already confirmed. Next a series of cellular localization data was taken. The protein sequence of all *E. coli* and *M. ruber* genes were input into TMHMM (Krogh et al., 2016), to find any transmembrane helices, SignalP (Petersen et al., 2011), to indicate a signal peptide probability, LipoP (Juncker et al., 2016), looking for signal peptidases, PSORTb (Yu et al., 2010), indicates likelihood of gene localization, and Phobius (Käll et al., 2007), which is a combined transmembrane and signal peptide predictor. Once the location of the genes was identified, the structure of the proteins must be confirmed. A conserved domain database (CDD) search (Marchler-Bauer et al., 2014) was done by doing a BLAST search of the gene of interest and clicking on the superfamilies' link at the top of the screen. This will give a list of domain hits and we are looking for the lowest e-value of a Cluster of orthologous Genes (COG) hit. If a *E. coli* and a *M. ruber* gene have the same COG number, this indicates the genes come have similar domains. Next, a TIGRFAM (Haft et al., 2016) and PFAM (Finn et al., 2014, Finn et al., 2016) search was done using the gene of interest's protein sequences. These bioinformatics tools classify the proteins into similar groups. If *E. coli* and *M. ruber* genes have the same TIGRFAM and PFAM number, it is an indication they are very similar in structure and function and therefore possibly orthologs. The last structure-based bioinformatics tool used was PDB (Berman et al., 2016, Berman et al., 2000). This is a protein database that gives a crystalized protein pictures. It is a relatively small database, so a hit with your gene is highly unlikely, but if the *E. coli* and *M. ruber* genes output the same crystal structure, it is a strong indication of similar structure and function. Now that the structure has been identified, the indication of an operon for our genes of interest was next. An EcoCyc search (Keseler et al., 2013) of the *E. coli* genes of interest was done by entering in the locus tags. This gave a summary of the genes and the operon tab was clicked on next. This gave an image of in which direction the gene was transcribed and any genes very close upstream or down stream was near it. Two or more genes next to each other being transcribed in the same direction is a good indication the gene is part of an operon. The IMG/M tool was used next. A search for all *E. coli* and *M. ruber* genes was done and a Chromosome viewer colored by KEGG search was done which gives a large portion of our gene is given with a similar look as what EcoCyc operon tab gave, but with our specific gene indicated with a red line underneath it. An out look like the EcoCyc test is looked for and indicates an operon. Lastly, another IMG/M search was done, but this time looking for "Show neighborhood regions with the same top COG hit" was chosen. This gives different organisms gene output of our gene of interest's region. If the different organisms' gene is also in an operon, it is a strong indication of an operon. All these results for both *E. coli* and *M. ruber* genes of interest were analyzed and compared to indicate orthologous genes. These bioinformatics tools are all free to use by anyone and great tools for other gene-based research.

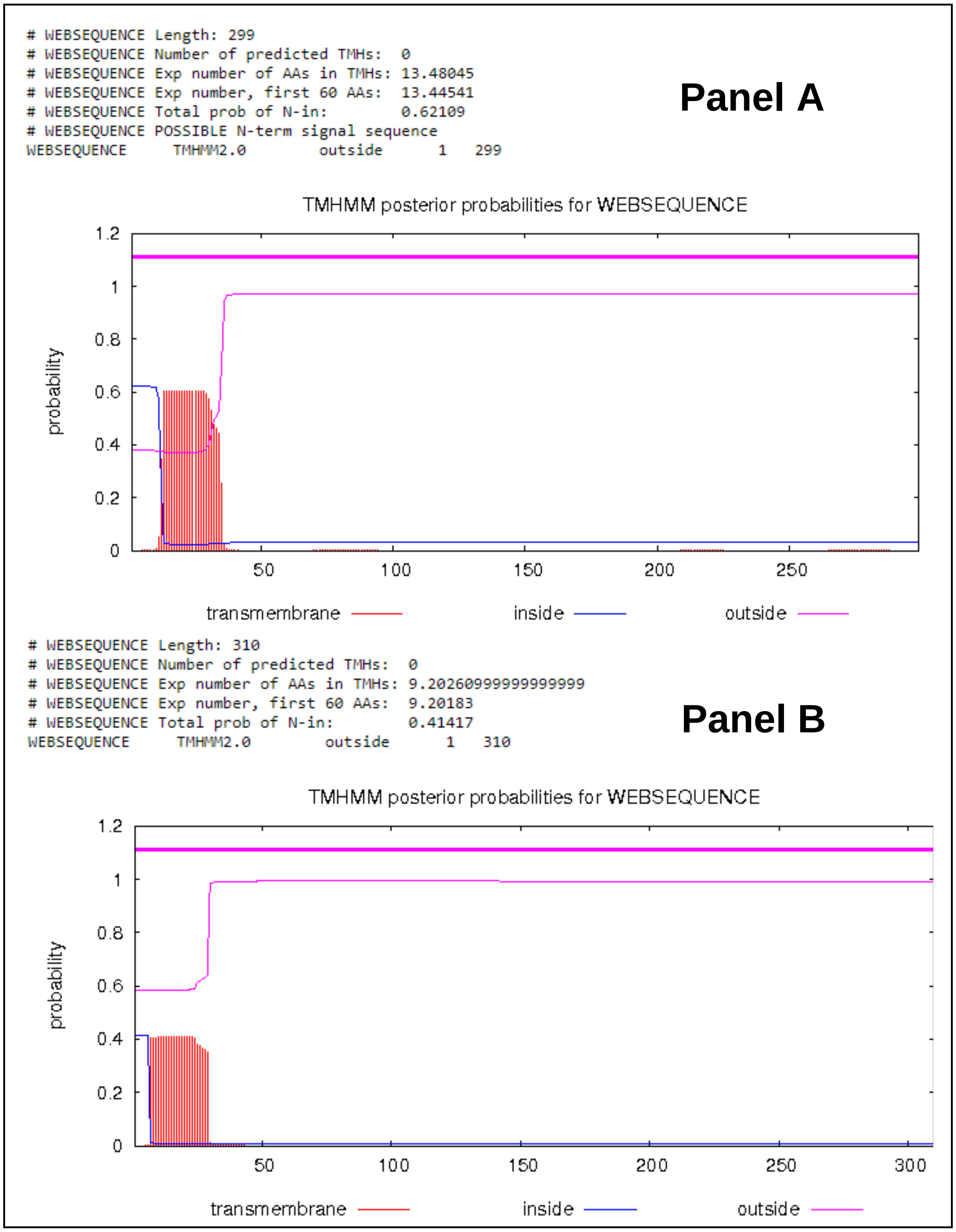
Results:

Table 2: *E. coli b_1857* and *Mrub_2836*

Bioinformatics tool used	***E. coli b_1857***	***M. ruber Mrub_2836***
BLAST *E. coli* against *M. ruber*	No match for each other.	
CDD Data (COG category)	COG #: COG4531	Cog #: COG0803
	E-value:1.67e-180	E-value:7.49e-63
Cellular Localization	Periplasm	
TIGRfam – protein family	TIGRFAM Number: TIGR03772	
	E-value:0.00058	E-value: 3.2e-10
Pfam – protein family	PFAM Number: PF01297	
	E-value: 2.3e-57	E-value: 7.3e-69
PDB	2OGW	2OGW
	E-value: 9.69e-153	E-value:9.3e-15
KEGG pathway map	Zinc ABC Transporters	

Table 1 Summarizes results from bioinformatics tools used to compare *E. coli ZnuA* gene to *Mrub_2836*. The first row of data was a BLAST result of *E. coli ZnuA* against the *M. ruber* genome. Interestingly, even though *E. coli b_1857* and *Mrub_2836* code for their version of ZnuA, there was no *Mrub_2836* locus tag from the BLAST search. There were a few *M. ruber* genes that matched, but with low e-values. Because there was no match of these genes from the BLAST search, it could mean these two genes are not related by this gene and are not orthologs of each other. The CDD search pulled up different COG numbers for the two genes and gave E-values very close to zero indicating that those genes do not belong to the same CDD family and probably are not orthologs of each other. So far, all the results have shown these two genes are not orthologs of each other, but when looking at the cellular localization bioinformatics data, (TMHMM, SingalP, LipoP, PSORTB, and Phobius) the results for both genes are identical throughout every bioinformatics tool and both indicate a cellular localization in the periplasm, and both lack a cleavage site. PSORTB gives a very strong indication of this with a periplasm localization score of 10, which is the highest score possible. This makes sense because these two genes were hypothesized to code for ZnuA protein which is a Zn^{2+} periplasmic binding protein and the results point towards these two genes having similar function and possibly being orthologs. The TIGRFAM data is relatively inconclusive. They both have the same TIGR number, TIGR03772, and name but the E-values for them are both extremely high for E-values indicating this result isn't the most accurate or reliable. PFAM, on the other hand, gave very positive results. Both proteins had the same PFAM number, PF01297, and both had very low E-values indicating both proteins are part of the zinc-uptake complex component A periplasmic. Additionally, the protein database (PDB) pulled the same name for these proteins, a high-affinity zinc uptake system protein, but the e-value for *M. ruber*, though close to zero, is very high compared to *E. coli*. Lastly, both genes were found in the same pathway for Zinc ABC transport system. Overall, the results from this table are back and forth. BLAST and CDD results indicate these genes are not orthologs but cellular localization, TIGRFAM, PFAM, PDB, and KEGG pathway all point towards these genes being orthologs of each other. The lack of any BLAST match for these genes is extremely discouraging though.

The images in Figure 3 are from the TMHMM bioinformatics tool. Panel A is from *E. coli ZnuA* and panel B is from *Mrub_2836*. Both plots do have red peaks that would indicate transmembrane helices, but these peaks are not significant enough to indicate transmembrane helices have formed. Because these genes were localized in the periplasm but are a part of a system that does have parts in the membrane, some of the amino acids of this gene are slightly embedded in the membrane, but not enough to form transmembrane helices. In saying that, both plots for *E. coli* and *M. ruber* are consistent with each other indicating the proteins from these genes are not in the membrane.



```
# WEBSEQUENCE Length: 299
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 13.48045
# WEBSEQUENCE Exp number, first 60 AAs:  13.44541
# WEBSEQUENCE Total prob of N-in:        0.62109
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        outside      1   299
```

```
# WEBSEQUENCE Length: 310
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 9.20260999999999999
# WEBSEQUENCE Exp number, first 60 AAs:  9.20183
# WEBSEQUENCE Total prob of N-in:        0.41417
WEBSEQUENCE     TMHMM2.0        outside      1   310
```

Figure 3. (Panel A): TMHMM plot for *E. coli ZnuA* indicating no transmembrane helices. (Panel B): TMHMM plot for *M. ruber*_2836 indicating no transmembrane helices. TMHMM Server v. 2.0 was used to create these plots http://www.cbs.dtu.dk/services/TMHMM/.

Figure 4 shows what the entire Zinc uptake ABC transporter system for both *E. coli* (Panel A) and for *M. ruber* (Panel B). The red circles in both panels indicates which part of the system *E. coli b_1857* and *Mrub_2836* genes code for. The fact that both genes code for the same portion of the zinc uptake transporter, ZnuA, it strongly points to the fact that these genes have the same function in these systems and, therefore, could possibly be orthologs.

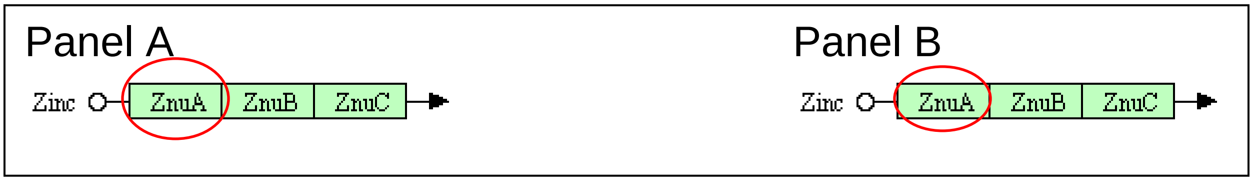
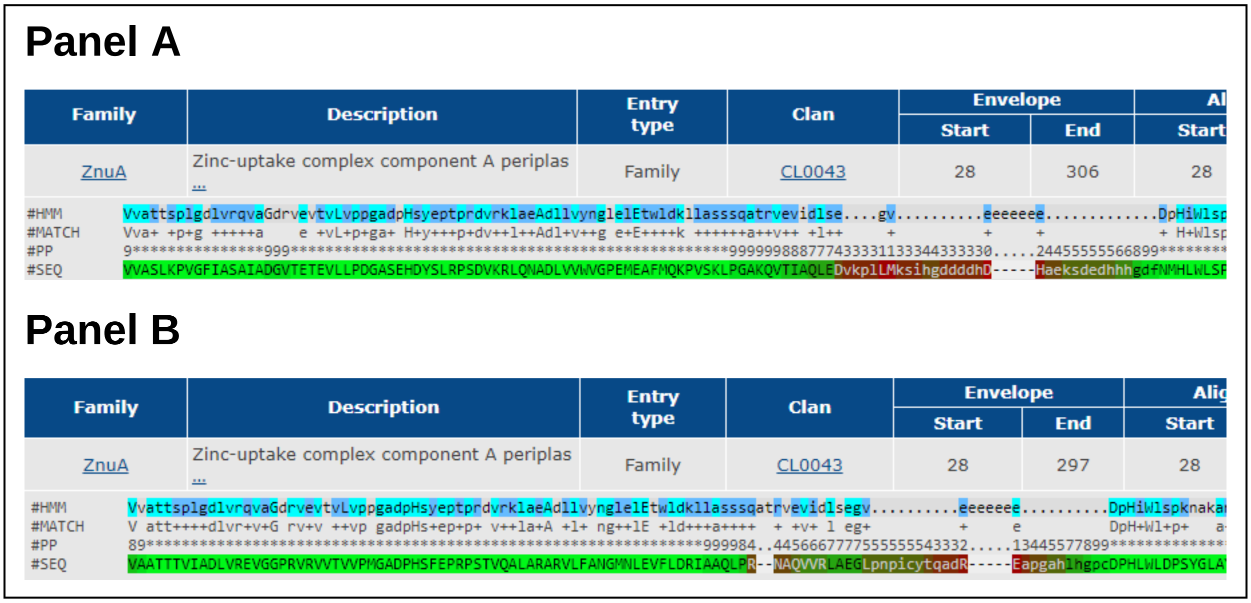


Figure 4. (Panel A): Indicates the entire *E. coli* zinc ABC transporter system. The red circle indicates what part *E. coli b_1857* codes for. (Panel B): This shows the *M. ruber* zinc ABC transporter system. The red circle indicates what part the *Mrub_2836* gene codes for. This image was taken from Kyoto Encyclopedia of Genes and Genomes (KEGG) database http://www.genome.jp/kegg-bin/show_pathway?org_name=eco&mapno=02010&mapscale=&show_description=show.

Figure 5 shows results from PFAM after entering both *E. coli b_1857* (Panel A) and *Mrub_2836* (Panel B) protein sequences. As the figure shows, both genes belong to the same family and clan. Being in the same family and clan indicate that the two proteins coded from *E. coli b_1857* and *Mrub_2836* have very similar structure and function. Below that in each panel is a pairwise alignment. #SEQ shows the sequence of the gene of interest. #HMM is a consensus sequence pulled from many different organisms with a gene coding for this. The fact that the #HMM sequence for both *E. coli b_1857* and *Mrub_2836* is the same indicates these genes are coding for proteins of the same function. The green highlighted amino acids in #SEQ indicates highly conserved amino acids when compared to the consensus sequence where red indicates non-highly conserved amino acids. Both *E. coli b_1857* and *Mrub_2836* have very similar highly conserved amino acids when compared to each other. This further indicates that these genes are in fact orthologs.



Panel A

Family	Description	Entry type	Clan	Envelope		Al
				Start	End	Start
ZnuA	Zinc-uptake complex component A periplas ...	Family	CL0043	28	306	28

```
#HMM      VvattsplgdlvrqvaGdrvevtvLvppgadpHsyeptprdvrklaeAdllvynglelEtwldkllasssqatrvevidlse....gv..........eeeeeee.............DpHiWlsp
#MATCH    Vva+ +p+g +++++a    e +vL+p+ga+ H+y+++p+dv++l++Adl+v++g e+E++++k ++++++a++v++ +l++     +          +     +            + H+Wlsp
#PP       9***************999*************************************************9999998887774333311333443333330.....24455555566899********
#SEQ      VVASLKPVGFIASAIADGVTETEVLLPDGASEHDYSLRPSDVKRLQNADLVVWVGPEMEAFMQKPVSKLPGAKQVTIAQLEDvkplLMksihgddddhD-----HaeksdedhhhgdfNMHLWLSP
```

Panel B

Family	Description	Entry type	Clan	Envelope		Alig
				Start	End	Start
ZnuA	Zinc-uptake complex component A periplas ...	Family	CL0043	28	297	28

```
#HMM      VvattsplgdlvrqvaGdrvevtvLvppgadpHsyeptprdvrklaeAdllvynglelEtwldkllasssqatrvevidlsegv..........eeeeeee..........DpHiWlspknaka
#MATCH    V att++++dlvr+v+G rv+v ++vp gadpHs+ep+p+ v++la+A +l+ ng++lE +ld+++a++++  + +v+ l eg+          +      e          DpH+Wl+p+   a
#PP       89*****************************************************************999984..44566677775555555543332.....13445577899*************
#SEQ      VAATTTVIADLVREVGGPRVRVVTVVPMGADPHSFEPRPSTVQALARARVLFANGMNLEVFLDRIAAQLPR--NAQVVRLAEGLpnpicytqadR-----EapgahlhgpcDPHLWLDPSYGLA
```

Figure 5. (Panel A): This is the PFAM results for *E. coli b_1857*. (Panel B): This panel shows the PFAM results for *Mrub_2836*. Both sequences in the pairwise alignment show similar highly conserved amino acids, same family, and same clan. Result images obtained from PFAM http://pfam.xfam.org/.

Figure 6 is a chromosome viewer colored by KEGG of *E. coli b_1857* (Panel A) and *Mrub_2836* (Panel B). Each arrow in the figures indicates a different gene and the direction the arrow points indicates the direction that gene is transcribed. The color of the arrow indicates the function that gene provides. Genes that are right next to each other, the same color, and are being transcribed in the same direction, indicates the genes are part of an operon. As you can see, our gene of interest is marked by the red bar, neither *E. coli b_1857* or *Mrub_2836* are part of an operon. This is a good sign of these two genes being orthologs. Unfortunately, *Mrub_2836* is transcribed in the opposite direction as *E. coli b_1857*, and the genes upstream and downstream of *Mrub_2836* are different from the genes upstream and downstream of *E. coli b_1857*. Because of this, this evidence refutes that these two genes are orthologs.

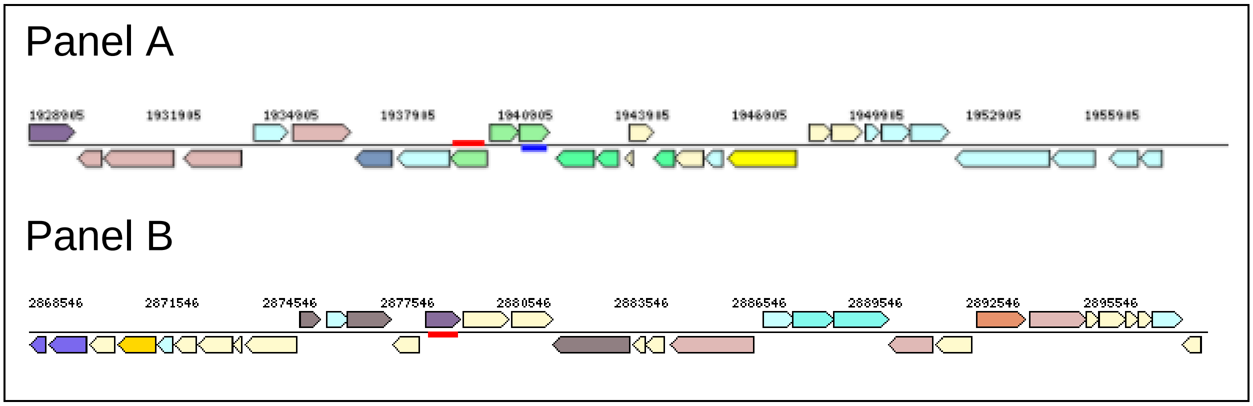


Figure 6. (Panel A): Chromosome Viewer colored by KEGG of *E. coli b_1857*. (Panel B): Chromosome Viewer colored by KEGG of *Mrub_2836*. Red bar in both panels indicates gene of interest. Indicates that neither gene is part of an operon. Images taken from IMG/M database https://img.jgi.doe.gov/cgi-bin/m/main.cgi.

Table 2: *E. coli b_1859* and *Mrub_1595*

Bioinformatics tool used	***E. coli b_1859***	***M. rub_1595***
BLAST *E. coli* against *M. ruber*	Score:112 E-value:6e-31	
CDD Data (COG category)	COG Number: COG1108	
	E-value: 3.5e-69	E-value: 2.4e-44
Cellular Localization	Cytoplasmic Membrane	
TIGRfam – protein family	TIGRfam Number: TIGR03770	
	E-value: 1.6e-9	E-value: 5.1e-20
Pfam – protein family	PFAM Number: PF00950	
	E-value:1e-89	E-value: 1.8e-65
PDB	No Results Found	
	E-value: NA	E-value: NA
KEGG pathway map	Zinc ABC Transporter System	

Table 2 summarizes the data collected from various bioinformatics tools to compare *E. coli b_1859* and *Mrub_1595*. The first row of data shows results from a BLAST search of *E. coli b_1859* against the *M. ruber* genome. The results gave a high bit score of 112 and a very E-value of 6e-31. The *M. ruber* gene that gave these values had a locus tag of *Mrub_1595*. This is the gene we were suspecting to be related to *E. coli* b_1859. The low E-value between these two genes indicates that these two genes share many of the same amino acids and that they are closely related. The CDD data gave the same COG number, COG1108, for *E. coli b_1859* and *Mrub_1595*, and both genes had very low E-values of 3.5e-69 and 2.4e-44 respectively indicating these genes code for the same enzyme in the zinc ABC transporter system. The bioinformatics tools for cellular localization, (TMHMM, SingalP, LipoP, PSORTB, and Phobius) indicated that both genes code for proteins that are located in the cytoplasmic membrane. Both gene proteins were found to have multiple transmembrane helices by TMHMM and Phobius, and PSORTB gave a cytoplasmic membrane localization score of 10 which is the highest it can be. Interestingly, SignalP and LipoP indicate no cleavage sites or signal peptides located in these proteins. This evidence can be refuted though as not all cytoplasmic membrane species have signal peptides and the PSORTB score of 10 for the cytoplasmic membrane over rules this refuting data. The TIGRFAM database gave the same TIGRFAM number of TIGR03770. Both *E. coli b_1859* and *Mrub_1595* have E-values close to zero indicating this data is significant and the proteins from these genes come from the same family, and therefore have the same function. The PFAM data confirms that the genes' proteins are very similar in structure as they have the same PFAM number and E-values close to zero. Strangely, the protein database came back with zero results for both *E. coli b_1859* and *Mrub_1595*. This evidence will also be refuted as PDB is a relatively small database could possibly not have any uploaded proteins similar enough to our *E. coli b_1859* or *Mrub_1595* proteins. Finally, both of these genes were predicted to be an integral part of the zinc uptake ABC transporter system.

Figure 7 shows the zinc uptake ABC transporter system for *E. coli* (Panel A) and *M. ruber* (Panel B). The red circles in both panels show which proteins are coded for by *E. coli b_1859* and *Mrub_1595*. Both genes code for the same protein ZnuB. This is strong indication that these genes are orthologs as they code for the same protein in the same system and, therefore, have the same function.

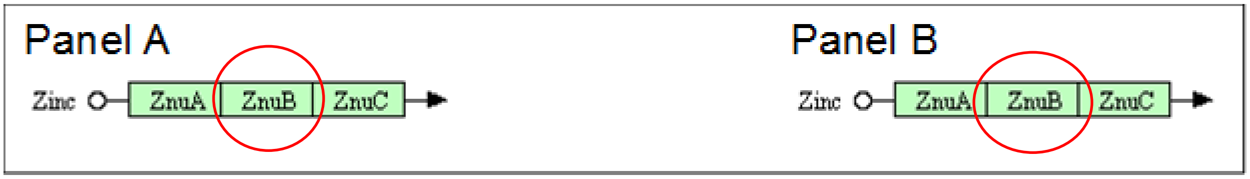
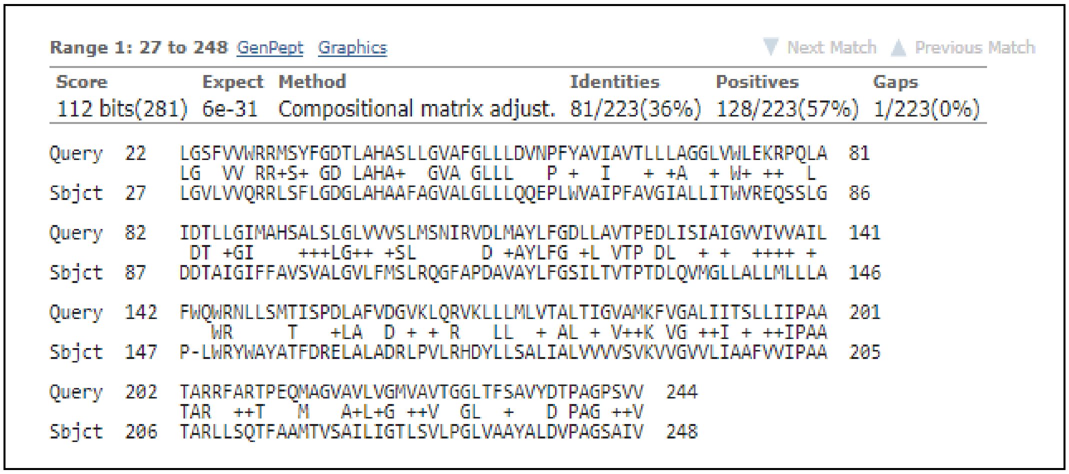


Figure 7. (Panel A): Indicates the entire *E. coli* zinc ABC transporter system. The red circle indicates what part *E. coli b_1859* codes for. (Panel B): This shows the *M. ruber*

zinc ABC transporter system. The red circle indicates what part the *Mrub_1595* gene codes for. Image taken from KEGG http://www.genome.jp/kegg-bin/show_pathway?mrb02010.

Figure 8 shows the results from a BLAST search of *E. coli b_1859* against *Mrub_1595*. This was the very first step in the research and gave the first indication that these genes are orthologs. About 36% of the amino acids were identical when comparing the two amino acid sequences, while 128 of the amino acids were very similar in their properties. The part of this figure that is really important is the E-value though. The E-value is 6e-31 which is very close to zero. This small E-value indicates that *Mrub_1595* and *E. coli b_1859* have similar structures because of similar amino acids, indicating these two genes might be orthologs.



Range 1: 27 to 248 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
112 bits(281)	6e-31	Compositional matrix adjust.	81/223(36%)	128/223(57%)	1/223(0%)

```
Query  22   LGSFVVWRRMSYFGDTLAHASLLGVAFGLLLDVNPFYAVIAVTLLLAGGLVWLEKRPQLA  81
            LG  VV RR+S+ GD LAHA+  GVA GLLL   P +  I   + +A  + W+ ++  L 
Sbjct  27   LGVLVVQRRLSFLGDGLAHAAFAGVALGLLLQQEPLWVAIPFAVGIALLITWVREQSSLG  86

Query  82   IDTLLGIMAHSALSLGLVVVSLMSNIRVDLMAYLFGDLLAVTPEDLISIAIGVVIVVAIL  141
             DT +GI    +++LG++ +SL       D +AYLFG +L VTP DL  + +  ++++ + 
Sbjct  87   DDTAIGIFFAVSVALGVLFMSLRQGFAPDAVAYLFGSILTVTPTDLQVMGLLALLMLLLA  146

Query  142  FWQWRNLLSMTISPDLAFVDGVKLQRVKLLLMLVTALTIGVAMKFVGALIITSLLIIPAA  201
               WR     T   +LA  D + + R   LL  + AL + V++K VG ++I + ++IPAA
Sbjct  147  P-LWRYWAYATFDRELALADRLPVLRHDYLLSALIALVVVVSVKVVGVVLIAAFVVIPAA  205

Query  202  TARRFARTPEQMAGVAVLVGMVAVTGGLTFSAVYDTPAGPSVV  244
            TAR  ++T   M   A+L+G ++V  GL  +   D PAG ++V
Sbjct  206  TARLLSQTFAAMTVSAILIGTLSVLPGLVAAYALDVPAGSAIV  248
```

Figure 8. BLAST search of *E. coli b_1859* against the *M. ruber* genome. Query sequence is *E. coli* b_1859. Subject sequence is *Mrub_1595*. Indicates *Mrub_1595* and *E. coli b_1859* have very similar protein sequences. + indicates similar proteins. Analysis was performed using BLAST bioinformatics tool https://blast.ncbi.nlm.nih.gov/Blast.cgi.

Figure 9 shows the TMHMM graph results for *E. coli b_1859* (Panel A) and *Mrub_1595* (Panel B). When comparing these two graphs, they are almost identical. The TMHMM graphs for both genes indicates that these proteins have 7 or 8 transmembrane helices and they are in the exact same locations in both proteins. These transmembrane helices predict that these proteins are located in the cytoplasmic membrane. This is the result we would expect as PSORTB gave a cytoplasmic membrane localization score of 10 and the protein ZnuB is known to be the inner membrane transport part of the system. The proteins coded for both *Mrub_1595* and *E. coli b_1859* have the same cell localization, therefore, we have more evidence that these two genes are orthologs.

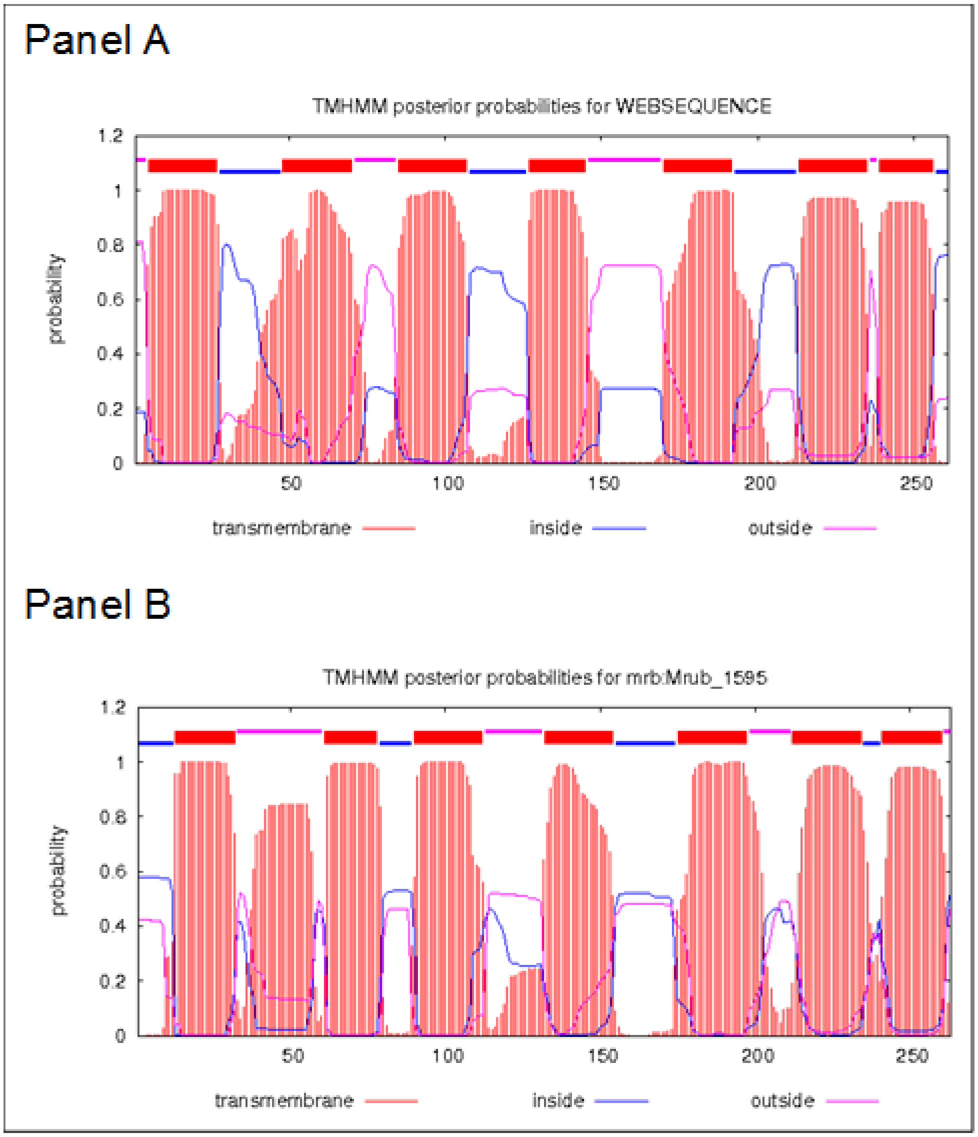
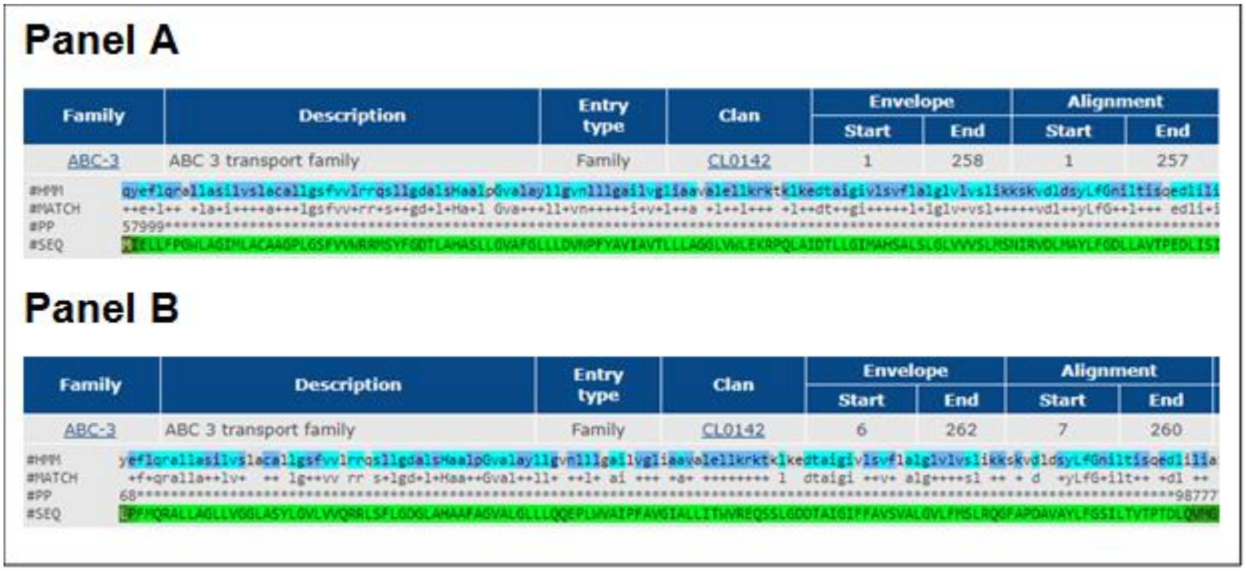


Figure 9. (Panel A). TMHMM graph of *E. coli b_1859* indicating transmembrane helices present. (Panel B). TMHMM graph of *Mrub_1595* indicating transmembrane helices present. Graphs created using TMHMM Server v. 2.0 http://www.cbs.dtu.dk/services/TMHMM/.

Figure 10 is a pairwise alignment from PFAM for *E. coli b_1859* (Panel A) and *Mrub_1595* (Panel B). As stated earlier, the #HMM sequence is the consensus sequence obtained from hundreds of different organisms that have similar protein sequences, and the #SEQ is our gene of interest's protein sequence. As you can see by the bright green colored amino acids in our genes sequences, they have highly conserved amino acids when compared to the consensus sequence, and the consensus sequence is the exact same for both *E. coli b_1859* and *Mrub_1595* meaning they both have highly conserved amino acids when compared to each other. Furthermore, both of these genes proteins are a part of the same family of proteins and clan of proteins indicating similar function and structure. This data gives us even more information that our genes are orthologous to one another.



Panel A

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
ABC-3	ABC 3 transport family	Family	CL0142	1	258	1	257

```
#HMM     qyeflqrallasilvslacallgsfvvlrrqsllgdalsHaalpGvalayllgvnlllgailvgliaavalellkrktklkedtaigivlsvflalglvlvslikkskvdldsyLfGnlltisqedlili
#MATCH   ++e+l++ +la+i++++a+++lgsfvv+rr+s++gd+l+Ha+l Gva+++ll+vn+++++i+v+l++a +l++l+++ +l++dt++gi+++++l+lglv+vsl+++++vdl++yLfG++l+++ edli+i
#PP      57999****************************************************************************************************************************
#SEQ     MIELLFPGWLAGIMLACAAGPLGSFVVWRRMSYFGDTLAHASLLGVAFGLLLDVNPFYAVIAVTLLLAGGLVWLEKRPQLAIDTLLGIMAHSALSLGLVVVSLMSNIRVDLMAYLFGDLLAVTPEDLISI
```

Panel B

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
ABC-3	ABC 3 transport family	Family	CL0142	6	262	7	260

```
#HMM     yeflqrallasilvslacallgsfvvlrrqsllgdalsHaalpGvalayllgvnlllgailvgliaavalellkrktklkedtaigivlsvflalglvlvslikkskvdldsyLfGnlltisqedllia
#MATCH    +f+qralla++lv+  ++ lg++vv rr s+lgd+l+Haa++Gval++ll+ ++l+ ai +++ +a+ ++++++++ l  dtaigi ++v+ alg++++sl ++ + d  +yLfG+ilt++ +dl ++
#PP      68*****************************************************************************************************************************98777
#SEQ     LPFMQRALLAGLLVGGLASYLGVLVVQRRLSFLGDGLAHAAFAGVALGLLLQQEPLWVAIPFAVGIALLITHVREQSSLGDDTAIGIFFAVSVALGVLFMSLRQGFAPDAVAYLFGSILTVTPTDLQVMG
```

Figure 10. (Panel A): This shows PFAM data from *E. coli* b_1859. (Panel B): This panel indicates PFAM data from *Mrub_1595*. Both panels indicate the genes have highly conserved amino acids. Results analyzed using PFAM http://pfam.xfam.org/.

Figure 11 shows the Chromosome Viewer Map Colored by KEGG for *E. coli b_1859* (Panel A) and for *Mrub_1595* (Panel B). As stated earlier, this can be an indicator for operons. Our gene of interest in each panel is indicated by a red bar. If our gene is next to another gene, is pointing in the same direction, meaning it is transcribed in the same direction, and is the same color, representing similar function, then that gene is part of an operon. As you can see from the two graphs, both *E. coli b_1859* and *Mrub_1595* are part of operons. They are also both in an operon with the ZnuC protein of their respective zinc-uptake ABC transporter system. The fact that both of these genes are operons gives more evidence that these genes are orthologs.

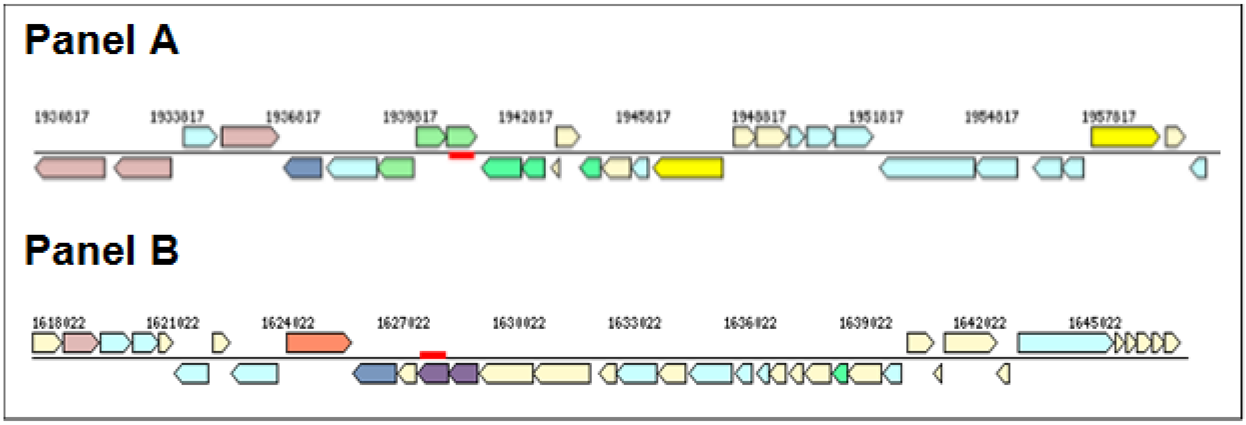


Figure 11. (Panel A): Chromosome Viewer Map Colored by KEGG for *E. coli* b_1859. (Panel B): Chromosome Viewer Map Colored by KEGG for *Mrub_1595*. Results indicate that both *E. coli b_1859* and *Mrub_1595* are part of operons. Images taken from IMG/M https://img.jgi.doe.gov/cgi-bin/m/main.cgi.

Table 3: *E. coli b_1858* and *Mrub_1596*

Bioinformatics tool used	***E. coli b_1858***	***Mrub_1596***
BLAST *E. coli* against *M. ruber*	Score: 132 E-value: 9e-39	
CDD Data (COG category)	COG Number: COG1121	
	E-value: 2.76e-106	E-value: 8.67e-82
Cellular Localization	Cytoplasmic Membrane	
TIGRfam – protein family	TIGRfam Number: TIGR03771	
	E-value: 5.7e-24	E-value: 7.1e-28
Pfam – protein family	PFAM number: PF00005	
	E-value: 2.8e-28	E-value: 1.2e-6
PDB	4YER	
	E-value: 1.06e-20	E-value: 6.22e-16
KEGG pathway map	Zinc ABC Transporter System	

Table 3 summarizes the combined data collected from various bioinformatics tools and compare two suspected orthologous genes, *E. coli b_1858* and *Mrub_1596*. The first set of data is a BLAST search of the *E. coli b_1858* against the *Mrub_1596*. This came back to give a very high Bit score as well as a very small E-value output of 9e-39. When this E-value gets that close to zero, the data is very significant and shows here that these two genes share a lot of similar proteins possibly meaning they are related. The next data provided is the CDD data. This gave the same COG number, COG1121 for both *E. coli b_1858* and *Mrub_1596* as well as very small E-values of 2.76e-106 and 8.67e-82 respectively. This indicates that both of these genes code for the same protein in the zinc uptake ABC transporter system. As for the localization bioinformatics tools, TMHMM and Phobius indicated that neither *E. coli b_1858* or *Mrub_1596* have any transmembrane helices. Signalp and LipoP both indicate that there was no signal peptide in either gene. All of this data is pointing to localization in the cytoplasm. But PSORTB only gives the cytoplasm localization score a 2.11 and a cytoplasmic membrane score of 7.88. This is not understood why, as this is a cytoplasmic ATP-binding protein. Because we know this gene codes for an ATP-Binding Protein, we know that this protein is located in the cytoplasm and not in the cytoplasmic membrane. Because we know this, and all other evidence pointing towards localization in the cytoplasm, PSORTB data can be ignored. The TIGRFAM database gave a result of both genes with the same TIGRFAM number, TIGR03771 and with low E-values. This indicates that the proteins from these genes are very similar and giving evidence that they are orthologs. Both *E. coli b_1858* and *Mrub_1596* have the same PFAM number of PF00005 and both have E-values close to zero. This shows that both of these genes proteins have a similar structure. The PDB for these genes was very promising. Both genes yielded the same protein code 4YER, the code of an ABC ATP-binding protein, with a description of being an ABC Transporter ATP-binding protein while both giving E-values close to zero showing that both *E. coli b_1858* and *Mrub_1596* have similar proteins and assemble their complex very similar to the 4YER protein. This is very strong evidence that these genes are orthologs.

Figure 12 shows the zinc uptake ABC transporter system for *E. coli* (Panel A) and *M. ruber* (Panel B). The red circle in each panel indicates the protein in the system that each gene codes for. Both *E. coli b_1858* and *Mrub_1596* genes code for the ZnuC protein in their respective system. This is strong information that our genes are orthologs because they code for similar proteins in the same ABC transporter system.

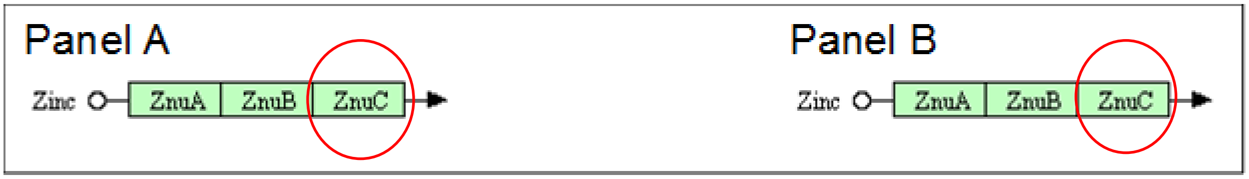
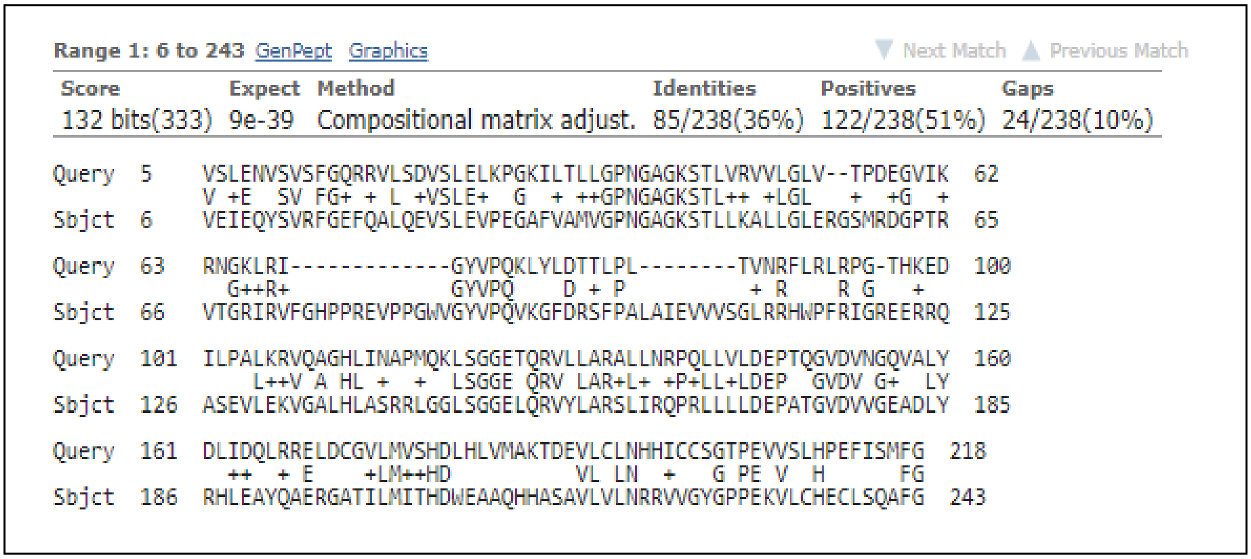


Figure 12. (Panel A): This is the zinc uptake ABC transporter system in *E. coli*. The red circle indicates the protein that the *E. coli b_1858* gene codes for. (Panel B): This is the

zinc uptake ABC transporter system in *M. ruber*. The red circle indicates the protein that the *Mrub_1596* gene codes for.

Figure 13 shows BLAST results of *E. coli b_1858* against the *Mrub_1596* gene. This BLAST search was the first indication that these genes were related to each other. As shown in the figure, 36% of the proteins are similar between these two genes, and 122 of them have similar properties between them. The major indicators that these proteins are related are the bit score and the E-score. The BLAST search gives a Bit score of 132 and an E-value of 9e-39. With this E-value that is close to zero, it indicates that *E. coli b_1858* and *Mrub_1596* have similar protein sequences and therefore similar structure. This information gives evidence that these two genes are orthologs.



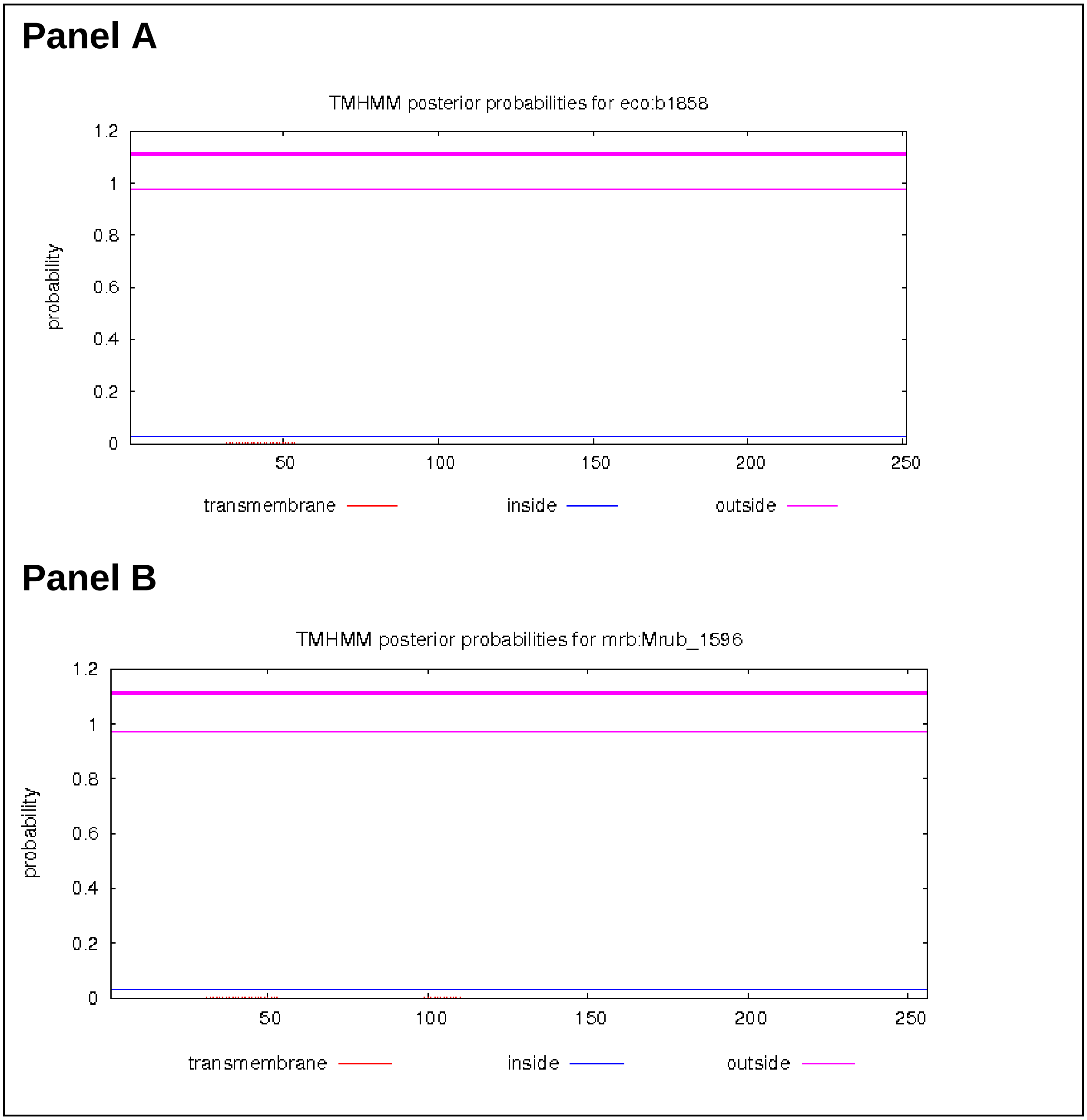
Range 1: 6 to 243 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
132 bits(333)	9e-39	Compositional matrix adjust.	85/238(36%)	122/238(51%)	24/238(10%)

```
Query  5    VSLENVSVSFGQRRVLSDVSLELKPGKILTLLGPNGAGKSTLVRVVLGLV--TPDEGVIK  62
            V +E  SV FG+ + L +VSLE+  G  + ++GPNGAGKSTL++ +LGL   +  +G  +
Sbjct  6    VEIEQYSVRFGEFQALQEVSLEVPEGAFVAMVGPNGAGKSTLLKALLGLERGSMRDGPTR  65

Query  63   RNGKLRI-------------GYVPQKLYLDTTLPL--------TVNRFLRLRPG-THKED  100
              G++R+             GYVPQ    D + P          + R    R G   +
Sbjct  66   VTGRIRVFGHPPREVPPGWVGYVPQVKGFDRSFPALAIEVVVSGLRRHWPFRIGREERRQ  125

Query  101  ILPALKRVQAGHLINAPMQKLSGGETQRVLLARALLNRPQLLVLDEPTQGVDVNGQVALY  160
                L++V A HL +  +  LSGGE QRV LAR+L+ +P+LL+LDEP  GVDV G+  LY
Sbjct  126  ASEVLEKVGALHLASRRLGGLSGGELQRVYLARSLIRQPRLLLLDEPATGVDVVGEADLY  185

Query  161  DLIDQLRRELDCGVLMVSHDLHLVMAKTDEVLCLNHHICCSGTPEVVSLHPEFISMFG  218
              ++  + E    +LM++HD          VL LN  +   G PE V  H      FG
Sbjct  186  RHLEAYQAERGATILMITHDWEAAQHHASAVLVLNRRVVGYGPPEKVLCHECLSQAFG  243
```

Figure 13. BLAST search of *E. coli b_1858* against the *M. ruber* genome. Query sequence is *E. coli* b_1858. Subject sequence is *Mrub_1596*. Indicates *Mrub_1596* and *E. coli b_1858* have very similar protein sequences. + indicates similar proteins. Analysis was performed using BLAST bioinformatics tool https://blast.ncbi.nlm.nih.gov/Blast.cgi.

Figure 14 shows TMHMM results for *E. coli b_1858* (Panel A) and *Mrub_1596* (Panel B). As you can see, neither of the proteins have any transmembrane helices in them. This not only indicates that both of these proteins lie in the cytoplasm and not in the membrane, but since both proteins of *E. coli b_1858* and *Mrub_1596* have the same localization, it provides more evidence that these genes are orthologs.

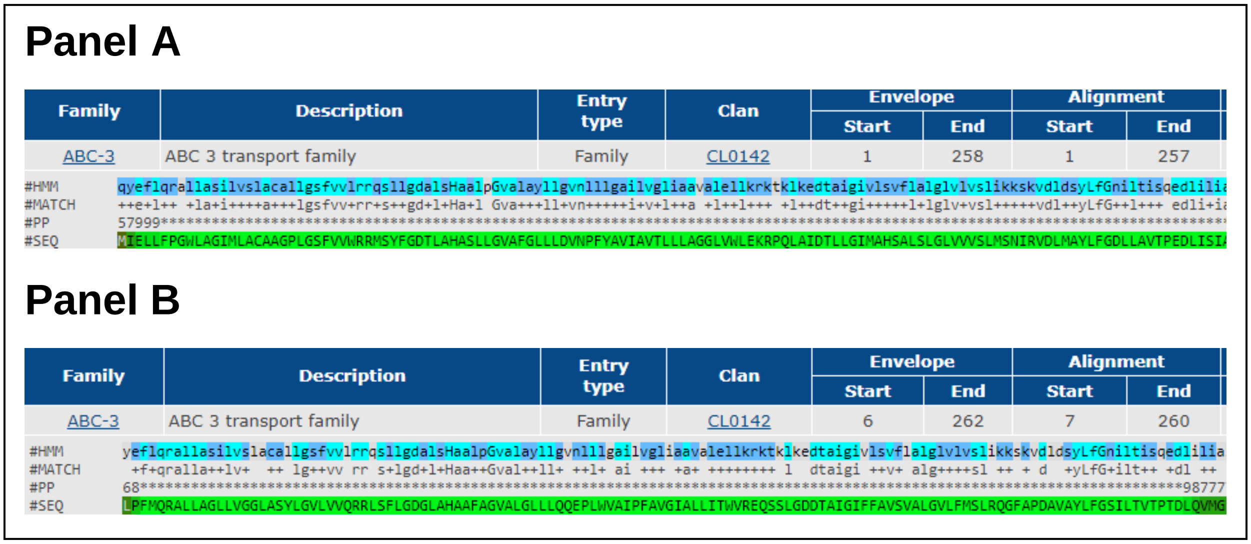


TMHMM posterior probabilities for eco:b1858

TMHMM posterior probabilities for mrb:Mrub_1596

Figure 14. (Panel A): TMHMM graph of *E. coli* b_1858. (Panel B): TMHMM graph of *Mrub_1596*. Both panels indicate no transmembrane helices in either protein. Graph analyzed by TMHMM Server v. 2.0 http://www.cbs.dtu.dk/services/TMHMM/.

Figure 15 shows the PFAM results of *E. coli b_1858* (Panel A) and *Mrub_1596* (Panel B). The consensus sequence in both panels are almost identical, and the gene sequence for both genes of interest are highly conserved when compared to the consensus sequence. Because the consensus sequence is identical in both panels, this means that the sequences for *E. coli b_1858* and *Mrub_1596* are highly conserved when compared to each other. Therefore, this gives more evidence that these two genes are orthologs.



Panel A

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
ABC-3	ABC 3 transport family	Family	CL0142	1	258	1	257

```
#HMM       qyeflqrallasilvslacallgsfvvlrrqsllgdalsHaalpGvalayllgvnlllgailvgliaavalellkrktklkedtaigivlsvflalglvlvslikkskvdldsyLfGniltisqedlilia
#MATCH     ++e+l++ +la+i++++a+++lgsfvv+rr+s++gd+l+Ha+l Gva+++ll+vn+++++i+v+l++a +l++l+++ +l++dt++gi+++++l+lglv+vsl+++++vdl++yLfG++l+++ edli+ia
#PP        57999******************************************************************************************************************************
#SEQ       MIELLFPGWLAGIMLACAAGPLGSFVVWRRMSYFGDTLAHASLLGVAFGLLLDVNPFYAVIAVTLLLAGGLVWLEKRPQLAIDTLLGIMAHSALSLGLVVVSLMSNIRVDLMAYLFGDLLAVTPEDLISIA
```

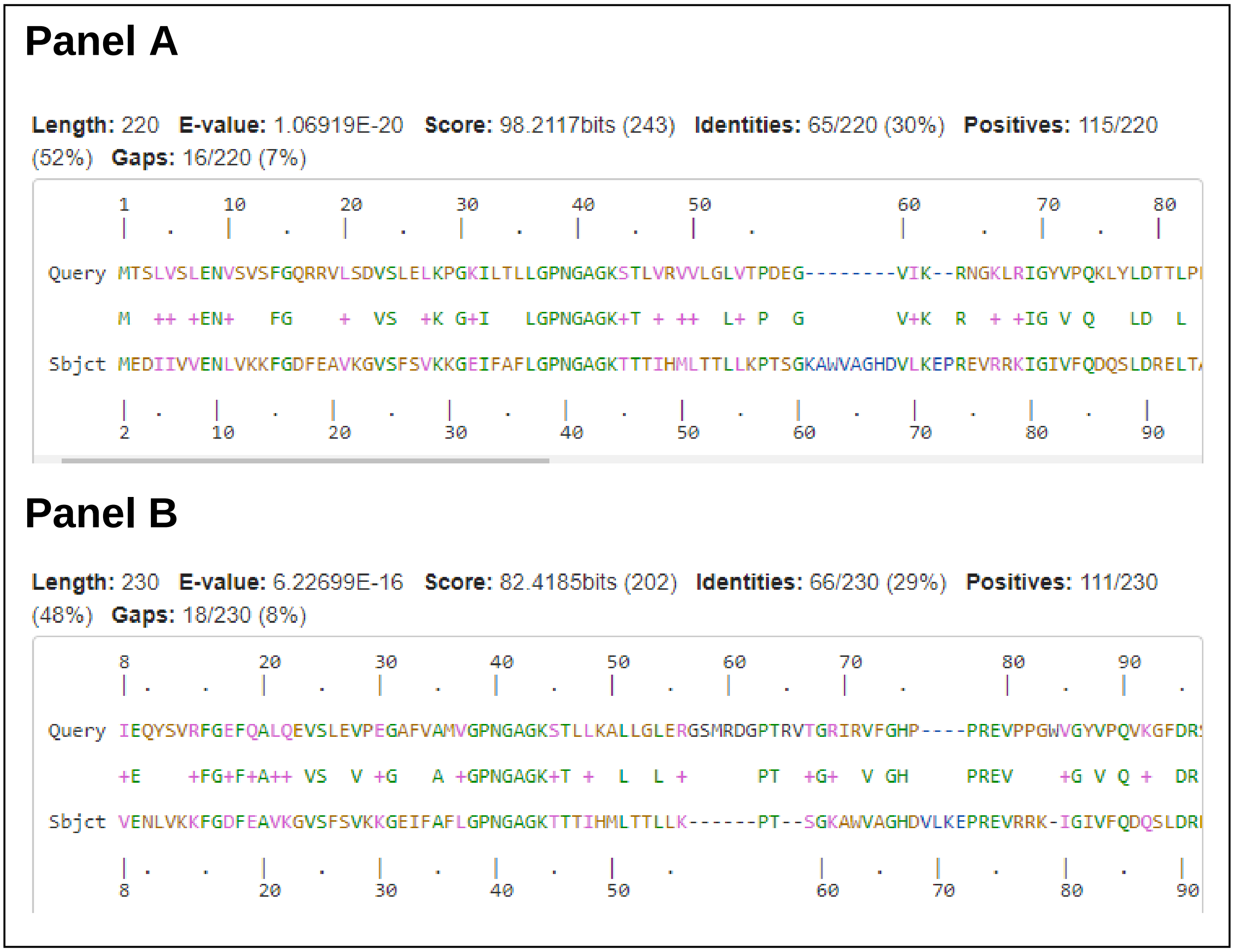
Panel B

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
ABC-3	ABC 3 transport family	Family	CL0142	6	262	7	260

```
#HMM       yeflqrallasilvslacallgsfvvlrrqsllgdalsHaalpGvalayllgvnlllgailvgliaavalellkrktklkedtaigivlsvflalglvlvslikkskvdldsyLfGniltisqedlilia
#MATCH      +f+qralla++lv+  ++ lg++vv rr s+lgd+l+Haa++Gval++ll+ ++l+ ai +++ +a+ ++++++++ l  dtaigi ++v+ alg++++sl ++ + d  +yLfG+ilt++ +dl ++
#PP        68***************************************************************************************************************************98777
#SEQ       LPFMQRALLAGLLVGGLASYLGVLVVQRRLSFLGDGLAHAAFAGVALGLLLQQEPLWVAIPFAVGIALLITWVREQSSLGDDTAIGIFFAVSVALGVLFMSLRQGFAPDAVAYLFGSILTVTPTDLQVMG
```

Figure 15. (Panel A): This shows PFAM data from *E. coli* b_1859. (Panel B): This panel indicates PFAM data from *Mrub_1595*. Both panels indicate the genes have highly conserved amino acids. Results analyzed using PFAM http://pfam.xfam.org/.

Figure 16 shows the pairwise alignment of *E. coli b_1858* (Panel A) and *Mrub_1596* (Panel B) when compared to 4YER protein which is an ABC transporter ATP-binding protein. The *E. coli b_1858* gene has 30% of its proteins are exactly the same, with 52% of them being similar to 4YER protein. The E-value is also 1.06e-20 which is very close to zero. This indicates *E. coli b_1858* is very similar in structure and function to 4YER protein. The *Mrub_1596* gene has 29% of its amino acids identical to 4YER with 48% of them similar. It also gives an E-value of 6.22e-16. This is very close to zero indicating *Mrub_1596* is similar to 4YER in function and structure. Because both *E. coli b_1858* and *Mrub_1596* are structurally and functionally similar to 4YER, they are both structurally and functionally similar to each other. This indicates that both of these genes are orthologous to each other.



Panel A

Length: 220 **E-value:** 1.06919E-20 **Score:** 98.2117bits (243) **Identities:** 65/220 (30%) **Positives:** 115/220 (52%) **Gaps:** 16/220 (7%)

```
      1        10        20        30        40        50                60          70        80
      |   .    |    .    |    .    |    .    |    .    |    .            |      .    |    .    |
Query MTSLVSLENVSVSFGQRRVLSDVSLELKPGKILTLLGPNGAGKSTLVRVVLGLVTPDEG--------VIK--RNGKLRIGYVPQKLYLDTTLP
      M  ++ +EN+   FG    +  VS  +K G+I   LGPNGAGK+T + ++  L+ P  G        V+K  R  + +IG V Q   LD  L
Sbjct MEDIIVVENLVKKFGDFEAVKGVSFSVKKGEIFAFLGPNGAGKTTTIHMLTTLLKPTSGKAWVAGHDVLKEPREVRRKIGIVFQDQSLDRELT
      |  .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
      2       10        20        30        40        50        60        70        80        90
```

Panel B

Length: 230 **E-value:** 6.22699E-16 **Score:** 82.4185bits (202) **Identities:** 66/230 (29%) **Positives:** 111/230 (48%) **Gaps:** 18/230 (8%)

```
      8           20        30        40        50        60        70            80        90
      | .    .    |    .    |    .    |    .    |    .    |    .    |    .        |    .    |    .
Query IEQYSVRFGEFQALQEVSLEVPEGAFVAMVGPNGAGKSTLLKALLGLERGSMRDGPTRVTGRIRVFGHP----PREVPPGWVGYVPQVKGFDR
      +E    +FG+F+A++ VS  V +G   A +GPNGAGK+T +  L  L +      PT  +G+  V GH     PREV    +G V Q +  DR
Sbjct VENLVKKFGDFEAVKGVSFSVKKGEIFAFLGPNGAGKTTTIHMLTTLLK------PT--SGKAWVAGHDVLKEPREVRRK-IGIVFQDQSLDR
      | .    .    |    .    |    .    |    .    |    .            |    .    |    .     |    .    |
      8           20        30        40        50                60        70         80        90
```

Figure 16. (Panel A): A portion of the pairwise alignment of *E. coli b_1858* compared to 4YER. (Panel B): A portion of the pairwise alignment of *Mrub_1596* compared to 4YER. These images were taken from PDB bioinformatics tool https://www.rcsb.org/.

Figure 17 shows the Chromosome Viewer Colored by KEGG for *E. coli b_1858* (Panel A) and *Mrub_1596* (Panel B). Both panels show our gene of interest with a red bar. Both panels also indicate they are part of an operon as they are next to another gene that are the same color and are transcribed in the same direction. Therefore, both *E. coli b_1858* and *Mrub_1596* are parts of operons giving more information that these are orthologs.

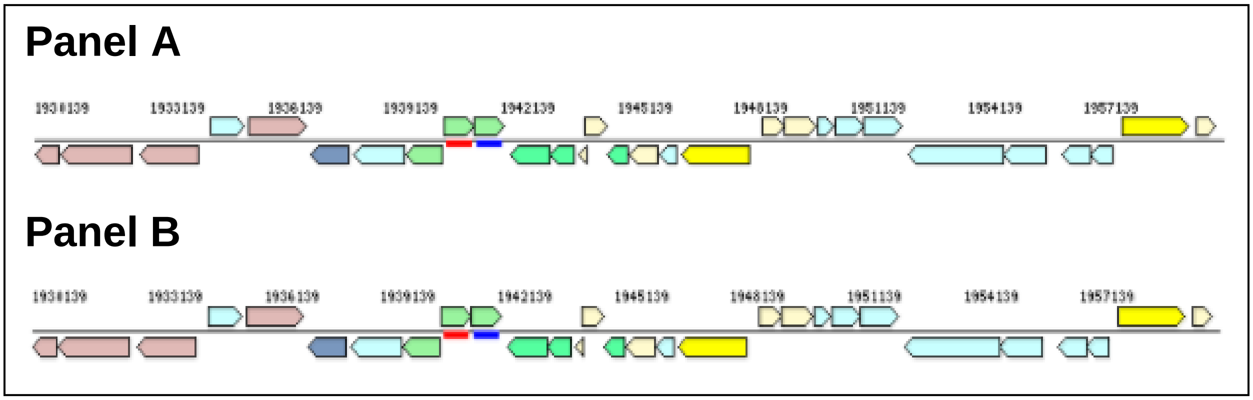


Figure 17. (Panel A): Chromosome Viewer Colored by KEGG for *E. coli* b_1858. (Panel B): Chromosome Viewer Colored by KEGG for *Mrub_1596*. Both panels indicate the genes are part of operons. Images taken from IMG/G https://img.jgi.doe.gov/cgi-bin/m/main.cgi.

Conclusion:

The results obtained for *E. coli b_1857* and *Mrub_2836* in this study gave mixed results as to if these are orthologs or not. All cellular localization bioinformatics tools like TMHMM, SignalP, LipoP, PSORTB, and Phobius all pointed towards a similar localization for both of these genes. Additionally, the PFAM and PDB results were the same for both genes while having E-values close to zero indicating that these genes would be orthologs. There was no match when doing BLAST search to compare the protein sequences of *E. coli b_1857* and *Mrub_2836*, but when doing a BLAST search of the *E. coli b_1857* protein sequence against the *Mrub_2836* protein sequence, it gives an E-value of 3e-24 which is close to zero which indicates these are orthologs. There was a TIGRFAM match but the E-value for both genes, especially *E. coli b_1857*, were very high. In saying that, the E-value was still below the threshold and therefore indicates orthologous genes. Finally, IMG/G indicated that these genes were not part of an operon, but the genes located upstream and downstream of these genes are completely different. In saying this tough, this still indicates orthologous genes as they are both not part of an operon. At first, because of the no match on the BLAST search, and the low E-values for the TIGRFAM results, it seemed these were not orthologous genes, but at further investigation *Mrub_2836* is an ortholog of E. *coli b_1857.*

The results obtained for *E. coli b_1859* and *Mrub_1595* in this study indicate that these genes are orthologs, meaning these genes have a common ancestry. This link was first obtained by doing a BLAST search to compare the two protein sequences of the genes giving a desired low E-value. This localization bioinformatics tools all indicated that the *E. coli b_1859* and *Mrub_1595* are located in the cytoplasmic membrane. Additionally, TIGRFAM and PFAM matched the protein sequences of *E. coli b_1859* and *Mrub_1595* with low E-values indicating similar structure and functions. The PDB bioinformatics tool strangely gave no results, and therefore had to be ignored, but *E. coli b_1859* and *Mrub_1595* are orthologous genes.

The results obtained for *E. coli b_1858* and *Mrub_1596* in this study indicate these genes are orthologs of each other. The first evidence found to support this was the BLAST search comparing *E. coli b_1858* against *Mrub_1596* giving a low E-value. Additionally, the localization bioinformatics all gave evidence of the localization of these genes to be in the cytoplasm, except for PSORTB. It only gave a cytoplasm score of 2.11 while giving a 7.88 score to cytoplasmic membrane. I believe this is because these genes are partially in both the cytoplasm and the cytoplasmic membrane, and it doesn't extend all the way through the membrane, therefore no transmembrane helices are needed. This would explain why TMHMM, SingalP, LipoP and Phobius all gave evidence of localization being in the cytoplasm. Also, the TIGRFAM and PFAM results matched in both *E. coli b_1858* and *Mrub_1596* all while showing E-values close to zero showing both of these genes are similar in structure. The PBD results was a very strong indicator of these genes being orthologs. Both *E. coli b_1858* and *Mrub_1596* were very similar to the 4YER protein indicating they are similar in function. Finally, IMG/G indicated that both of these genes are part of operons and have similar genes upstream and downstream of them.

In conclusion, *Mrub_1596* and *Mrub_1595* are orthologous to *E. coli b_1858* and *E. coli b_1859* respectively based on consistent evidence from various bioinformatics tools. In saying that, there is too much refuting evidence to confirm that *Mrub_2836* and *E. coli b_1857* are orthologs.

Works Cited

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed

2. 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/.

3. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

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

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

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

7. Finn RD, Bateman A, Clements J, *et al*. Pfam: the protein families database. Nucleic Acids Research. 2014;42 (Database issue):D222-D230. doi:10.1093/nar/gkt1223.

8. 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/

9. 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.

10. Juncker, A., 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/

11. Kall L, Krough A, Sonnhammer E. 2004. A combined transmembrane topology and signal Lukas Käll, Anders Krogh and Erik L. L. Sonnhammer. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of*

Molecular Biology, 338(5):1027-1036, May 2004.
(doi) (PubMed)

12. Käll,L., Anders Krogh and Erik L. L. Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.*, 35:W429-32, July 2007 (doi) (PubMed)

13. 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/

14. 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.

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

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

17. Lori Scott, personal communication.

18. 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/ BLAST tool: BLASTp tool from https://blast.ncbi.nlm.nih.gov/Blast.cgi

19. 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. 2015 Jan 28;43(Database issue):D222-2. doi: 10.1093/nar/gku1221. Epub 2014 Nov 20. [PubMed PMID: 25414356]

20. 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

21. Moussatova A, Kandt C, O'Mara ML, Tieleman DP. 2008. ATP-binding cassette transporters in escherichia coli. Biochimica Et Biophysica Acta (BBA) - Biomembranes 1778(9):1757-71.

22. Patzer SI and Hantke K. 1998. The ZnuABC high-affinity zinc uptake system and its regulator zur in escherichia coli. Mol Microbiol 28(6):1199-210.

23. Petersen, Thomas, 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

24. Phylogenetic Diversity [Internet] [cited 2018 Feb 9,]. Available from: https://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/.

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

26. Stefanidou M, Maravelias C, Dona A, Spiliopoulou C. 2006. Zinc: A multipurpose trace element. Arch Toxicol 80(1):1-9.

27. Tindall, B. J., Sikorski, J., Lucas, S., Goltsman, E., Copeland, A., Glavina Del Rio, T., … Lapidus, A. (2010). Complete genome sequence of Meiothermus ruber type strain (21T). Standards in Genomic Sciences, 3(1), 26–36. http://doi.org/10.4056/sigs.1032748

28. Yatsunyk LA, Easton JA, Kim LR, Sugarbaker SA, Bennett B, Breece RM, Vorontsov II, Tierney DL, Crowder MW, Rosenzweig AC. 2008. Structure and metal binding properties of ZnuA, a periplasmic zinc transporter from escherichia coli. United States: USDOE. Report nr 13. 271 p.

29. Yu, NY, 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