2018

Mrub_0680, Mrub_0836, and Mrub_0837 found to be orthologous to *E. coli* CcmA, CcmB, and CcmC, respectively, coding for ABC-transport proteins involved in cytochrome-C biogenesis

Sarah N. Church
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_0680, Mrub_0836, and Mrub_0837 found to be orthologous to *E. coli CcmA, CcmB, and CcmC*, respectively, coding for ABC-transport proteins involved in cytochrome-C biogenesis.

Sarah Church, Dr. Lori Scott

1 | Introduction

1.1 | Cytochrome C Biosynthesis

Cytochromes are a large group of proteins in bacteria responsible for transport of electrons between proteins. Most cytochromes are located in the cytoplasm, but a portion of them are located in the periplasm and interact with intergral membrane proteins located there (Thonyl-Meyer, 1997). Cytochromes are primarily heavily involved in aerobic and anaerobic respitation and photosynthesis (Feissner *et al.*, 2006; Thonyl-Meyer, 1997). In this study we look specifically at cytocrome-C (Cyt-C) which is a peripheral membrane protein that is associated with intergral membrane proteins (Yeagle, 2016). C-type cytochromes are characterized by their involvement in ATP-coupled reactions and heme co-factor. The heme is covalently bound to the c-type cytochrome via two thiodiester bonds (Thonyl-Meyer, 1997). In *Escherichia coli,* Cyt-C serves as an electron carrier between membrand bound proteins (Yeagle, 2016). In *E. coli* Cyt-C biosynthesis involves proteins CcmA, CcmB, CcmC, CcmD, CcmE, CcmF, CcmG, and CcmH (Kesler *et al.*, 2013). Many of these proteins are located in the cytoplasmic membrane, as seen in the proposed layout in Fig 1. below.

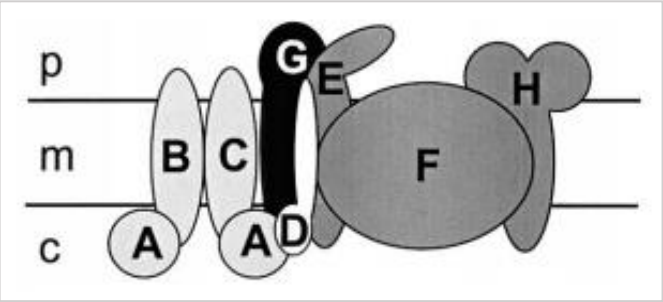

Fig 1. Proposed arrangement of proteins involved in cytochrome C biosynthesis. Different shading refers to different function of the proteins. The letters relate to the Ccm_ gene nomenclature used for *Escherichia coli*. This diagram was reproduced from Thonyl-Meyer (1997).

1.2 | ABC Transporter Function

ATP-binding cassette (ABC) transporters are very common in bacteria, comprising a significant amount of the genome of bacteria such as *E. coli* (Wilkens, 2015). For this study, the focus is on the transport complex involved in Cyt-C biosynthesis which includes CcmA, CcmB, CcmC, and CcmD; *E. coli* locus tags b2201 (2297021..2297644), b2200 (2296362..2297024), b2199 (2295583..2296320), and b2198 (2295377..2295586), respectively; these sequences were taken from KEGG (Kanehisa *et al.*, 2016). This system was initially thought to be involved with heme export as listed in The UniProt Consortium (2016) and Ecocyc (Kesler *et al.*, 2013). This transport was of particular interest because the production of Cyt-C requires both heme and the Cyt-C protein to be in the periplasm. To have this happen, heme needs to be exported into the periplasm (Fergusson *et al.*, 2007). The CcmBC complex was the primary target for heme transport but has since been disproven by several studies (Thonyl-Meyer, 1997; Goldman *et al.*, 2001). In *E. coli* CcmB and CcmA make up the transport complex of the system with interacting domains however since heme as a substrate has been disproved, researchers are unsure of what the complex exports (Fergusson *et al.*, 2007; Goldman *et al.*, 2001). CcmA is known to contain an ATP-binding domain and has been found to interact with the CcmCB complex and is involves in heme transport as it has a heme binding site as well (Thonyl-Meyer, 1997; Ferguson *et al.*, 2007). CcmA transports heme to CcmE, a chaperone protein, which is involved in the further steps of the Cyt-C biosynthesis mechanism. This step is thought to be assisted by CcmD in *E. coli* (Ferguson *et al.*, 2007). These four genes are all part of an operon in *E. coli* as suggested by the consecutive locus tags. This is confirmed by the operon map found through Ecocyc (Kesler *et al.*, 2013). Transcription begins at b2201 and the operon codes for Cyt-C related genes starting with CcmA through CcmH (Kesler *et al.*, 2013).

1.3 | *Meiothermus ruber* as Study Organism

Meiothermus ruber is a red pigmented gram-negative bacterium closely related to the genus *Thermus*. The specific organism examined in this study is *M. ruber* DSM 1279 as sequenced by Tindall *et al.* (2010). This organism was isolated from Russian hot springs and then further isolated in lab. Additionally, it has been found in many other European and Asian countries giving rise to several species within the genus *Meiothermus* (Tindall *et al.*, 2010). One reason *Meiothermus* bacteria have been studied is due to their ability to form biofilms on certain surfaces through the excretion of products from adhesion organelles. Studies have found *M. ruber* and *M. silvanus* to make colored biofilms on machine equipment and also from colonies on produced paper and boards (Huarng *et al.*, 2011).

1.4 | Purpose and Hypothesis

This study specifically seeks to map out the functions of three genes proposed to be involved in the ABC transport complex involved in Cyt-C biosynthesis. These genes are found at locus tags Mrub_0680 (659484..660071), Mrub_0836 (823734..824399), and Mrub_0837 (824570.825262), as taken from KEGG database (Kanehisa *et al.*, 2016). We hypothesize that Mrub_0680 is orthologous to b2201, Mrub_0836 is orthologous to

b2200, and Mrub_0837 is orthologous to b2199. Additionally, these genes are involved in the ABC transport complex involved in Cyt-C biosynthesis.

2 | Materials and Methods

2.1 | Sequences and Orthologs

Nucleotide and amino acid sequences for Mrub_0680, Mrub_0836, Mrub_0837, b2198, b2199, b2200 and b2201 were obtained from KEGG database from the ABC transporter map (KEGG map number 02010) (Kanehisa *et al.*, 2016). *E. coli* genes were run against the *M. ruber* genome using amino acid sequences in BLASTp to identify orthologs between the two bacteria (Altschul *et al.*, 1990).

2.2 | Alternate Start Codons

Start codons were identified and confirmed for all *M. ruber* genes using bioinformatics tools. IMG/M database was used to look at the amino acid 50 positions upstream of the suggested start codon to identify possible alternate start codons and Shine-Dalgarno sequences (Markowitz 2012). Multiple sequence alignment was generated in T-Coffee using orthologs of 10-15 closely related species to the gene of interest (GOI) taken from a BLASTp search (Madden *et al.*, 2002; Notredame *et al.*, 2000). Weblogo was also generated using the same closely related orthologs to examine conserved amino acids (Crooks *et al.*, 2004). These steps were completed for all *M. ruber* genes. Start codons for the *E. coli* genes have already been confirmed (Keseler *et al.*, 2013).

2.3 | Cell Localization

Amino acid sequence was run through TMHMM, PSORTb, and Phobius to determine the number of transmembrane helices and where in the cell it was most likely to be located (Krogh *et al.*, 2016; Krogh *et al.*, 2001; Sonnhammer *et al.*, 1998; Yu *et al.*, 201; Kall *et al.*, 2007; Kall *et al.*, 2004). The amino acid sequence was also run through bioinformatics tools SignalP and LipoP to determine if a signal peptide was present in the protein (Petersen *et al.*, 2011; Junker *et al.*, 2003). This process was followed for all *M. ruber* and *E. coli* GOIs.

2.4 | Protein Function Similarities

Domains of all *M. ruber* and *E. coli* GOIs were determined by running the amino acid sequence through the CDD database of NCBI BLASTp, TIGRfam, and Pfam returning significant hits for COG, TIGR, and PF domains, respectively (Marchler-Bauer *et al.*, 2014; Haft *et al.*, 2001; Finn *et al.*, 2014; Finn *et al.*, 2016). For CDD and TIGRfam, a simple output was given and data collected. For Pfam, pairwise alignment between the domain and the GOI was recorded and the HMM logo for the PF domain was saved (Finn *et al.*, 2014; Finn *et al.*, 2016). Additionally, the sequences were run through PDB to see if the protein, or any of its domains, had been crystalized (Bergman *et al.*, 2000; Bergman *et al.*, 2000). If there was a PDB match, pairwise alignment and 3D protein image was recorded.

2.5 | Operons

Next, genes were analyzed on a wider scale to examine the possibility it was part of an operon. The locus tag of the GOI was run through IMG/M and then images of nearby genes was obtained using the Color-by-KEGG imaging option and the genes upstream and downstream of the GOI were identified (Markowitz *et al.*, 2012). Using the function "Show neighborhood regions with the same top COG hit," an image of 5 related species was generated of the area of the chromosome containing the GOI (Markowitz *et al.*, 2012).

2.6 | Paralogs

Paralogs for all GOIs were determined using KEGG (Kanehisa *et al.*, 2016). Once the gene was located from the pathway a database search was done from KEGG which ran the amino acid sequence against the GOI's own genome only (Mrub genes run against *M. ruber* genome and *E. coli* genes run through the *E. coli* genome) to determine if there were any significantly comparable genes. If there were significant paralogs an alignment was drawn to see how closely related the genes were (Kanehisa *et al.*, 2016).

3 | Results

3.1 | KEGG and BLASTp Analysis

The genes of interest were found on KEGG to be part of the heme transport system within the ABC Transporter map. *E. coli* has genes that code for proteins CcmA (b2201), CcmB (b2200), CcmC (b2199), and CcmD (b2198) as seen in Fig. 1A. *M. ruber* has genes that code for CcmA (Mrub_680), CcmB (Mrub_836), and CcmC (Mrub_837), but does not code for CcmD as seen in Fig 1B (Kanehisa *et al.*, 2016). The BLASTp search of

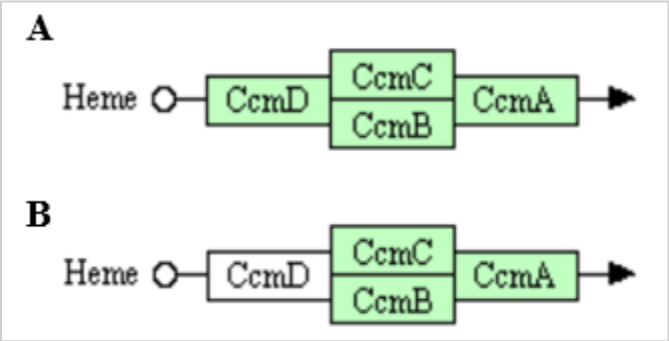

Fig 2. KEGG pathway maps for (A) *Escherichia Coli* K-12 and (B) *Meiothermus ruber* DSM 1279 for heme ABC Transport map. ID: 02010

E. coli GOIs against the *M. ruber* genome confirmed that the suggested *M. ruber* and *E. coli* genes were orthologs. The top hit for b2201 was Mrub_0680 with an expect value of 2e-25 and bit score of 94.7, the alignment can be seen in Fig 4A. The top hit for b2200 was Mrub_0836 with an expect value of 3e-09 and bit score of 52.0, the pairwise alignment can be seen in Fig 4B. The top hit for b2199 was Mrub_0837 with an expect value of 2e-28 and bit score of 105, the alignment can be seen in Fig 4C. BLASTp search of b2198 against *M. ruber* DSM 1279 yielded no significant hits within the genome.

3.2 | CcmA: Mrub_0680 and b2201

Pairwise alignment of multiple related orthologs of Mrub_0680 showed that M5 is the most likely start codon of the CcmA gene in *M. ruber* (Fig 3). IMG/M identified a Shine-Dalgarno sequence 9 to 14 positions upstream of the initial start codon of M1, which is then 11 to 26 amino acids away from the new proposed start codon of M5. The WebLogo in Fig 6 confirms the conserved M5 start codon.

Fig 5 shows that neither Mrub_0680 nor b2201 contains transmembrane helices as determined using TMHMM and Phobias tools. Mrub_0680 had a signal peptide probability of 0.135 on SignalP and was predicted by LipoP to be a cytosolic protein not containing a signal peptide. b2200 has a signal peptide probability of 0.120 on SignalP and

```
Bacillus_sonorensis              MI-----AELHGVQKRFKGKRVLEDIN
Chloroflexi_bacterium            M------ISVTKLTKRFGLKTILRNLD
Clostridium                      M------FKLNNVSKTIKKQNVLKELN
Euryarchaeota_archaeon           M------IEAKGISKTFGRRTVLDKVD
Geobacillus                      MK----AIVLSNVSKTIKGREVLRHIN
Geobacillus_thermocatenulatus    MK----AIVLSNVSKTIKAREVLRNIN
Marinithermus_hydrothermalis     MSAGTGIAELVGVWKRYGREWILKDLN
Meiothermus_cerbereus            M-----LIEAVSVSKRYGRDWVLRNLD
Meiothermus_ruber                MVCS-MLIEAIAVSKRYGRDWVLRNLD
Oceanithermus_profundus          MAD--PFVEVQNVWKRFGRQWVLRDLS
Thermanaerothrix_daxensis        M------IEVHHLTKRYGPKVVLRRLD
Thermus_aquaticus                M-----LLRLLGVSKRFGRDWVLKDLD
Thermus_tengchongensis           M-----LLRLQGISKRFGRDWVVRDLS
                                 *          : *     ::  :.
```

Fig 3. Multispecies pairwise alignment of Mrub_0680 and related orthologs showing likely start codon. Created using T-Coffee, described in methods.

A

Score	Expect	Method	Identities	Positives	Gaps
94.7 bits(234)	2e-25	Compositional matrix adjust.	64/167(38%)	85/167(50%)	18/167(10%)

```
Query  22   LSFTLNAGEWVQITGSNGAGKTTLLRLLTGLSRPDAGEVLWQGQPLHQVRDSYHQNLLWI  81
            L F L    E V + G NG GKTTLLR+L GL RP  G V   G+             + ++
Sbjct  25   LDFQLAQHEAVALVGPNGVGKTTLLRVLAGLVRPTQGSVKLSGR------------VGFL  72

Query  82   GHQPGIKTRLTALENLHFYHR-DGDT---AQCLEALAQAGLAGFEDIPVNQLSAGQQRRV  137
             + P       T  ENLH+  R DG T    ++   AL Q GL    D PV   S+G ++R+
Sbjct  73   ANPPAFHRHFTGAENLHYALRLDGKTGGRSEIRAALTQFGLP--HDKPVLSYSSGMKKRL  130

Query  138  ALARLWLTRATLWILDEPFTAIDVNGVDRLTQRMAQHTEQGGIVILT  184
             A+ARL L    +W+LDEP  A+D  G   L   + Q   +GG+VI T
Sbjct  131  AMARLHLQNPDIWLLDEPEAALDAQGRGLLENLVQQARSRGGVVIAT  177
```

B

Score	Expect	Method	Identities	Positives	Gaps
94.7 bits(234)	2e-25	Compositional matrix adjust.	64/167(38%)	85/167(50%)	18/167(10%)

```
Query  22   LSFTLNAGEWVQITGSNGAGKTTLLRLLTGLSRPDAGEVLWQGQPLHQVRDSYHQNLLWI  81
            L F L    E V + G NG GKTTLLR+L GL RP  G V   G+             + ++
Sbjct  25   LDFQLAQHEAVALVGPNGVGKTTLLRVLAGLVRPTQGSVKLSGR------------VGFL  72

Query  82   GHQPGIKTRLTALENLHFYHR-DGDT---AQCLEALAQAGLAGFEDIPVNQLSAGQQRRV  137
             + P       T  ENLH+  R DG T    ++   AL Q GL    D PV   S+G ++R+
Sbjct  73   ANPPAFHRHFTGAENLHYALRLDGKTGGRSEIRAALTQFGLP--HDKPVLSYSSGMKKRL  130

Query  138  ALARLWLTRATLWILDEPFTAIDVNGVDRLTQRMAQHTEQGGIVILT  184
             A+ARL L    +W+LDEP  A+D  G   L   + Q   +GG+VI T
Sbjct  131  AMARLHLQNPDIWLLDEPEAALDAQGRGLLENLVQQARSRGGVVIAT  177
```

C

Score	Expect	Method	Identities	Positives	Gaps
105 bits(261)	2e-28	Compositional matrix adjust.	65/188(35%)	106/188(56%)	10/188(5%)

```
Query  44   APADYQQGNSYRIIYLHVPAAIWSMGIYASMAVAAF--IGLVWQMKMANLAVAAMAPIGA  101
            +P D  QG   RI ++HVP A W M   AS    A+    LV +    +   AA+  +G
Sbjct  37   SPPDQSQGFVARIFHMHVPTA-W-MAYLASFGALAYSVAYLVRRRAHHDRVAAAVVEVGL  94

Query  102  VFTFIALVTGSAWGKPMWGTWWVWDARLTSELVLLFLYVGVIALWHAFDDRRLAGRAAGI  161
            +F  +AL+TG  W +P WG +W W+ RLT+  +L  +YVG + +  A +D  L  +AA
Sbjct  95   IFMALALLTGMLWARPTWGVYWDWEPRLTTTAILFAIYVGYVVVRGAIEDPELRAKAAAG  154

Query  162  LVLIGVVNLPIIHYSVEWWNTLHQGS----TRMQQSIDPAMRSPLRWSIFGFLLLSATLT  217
            + ++G +N+PI + SV+WW +LHQ      T  + ++D A+  P+  ++  F LL   +
Sbjct  155  VAILGSINVPISYMSVKWWRSLHQTQSIDLTTGKINVDAAILIPMLVNLAAFTLL--FIG  212

Query  218  LMRMRNLI  225
            L+R+R++I
Sbjct  213  LVRLRSII  220
```

Fig 4. BLASTp alignment of orthologs of *Escherichia coli* GOIs when against the *Meiothermus ruber* genome; Panel A shows b2201 alignment with Mrub_680, panel B shows b2200 alignment with Mrub_836, panel C shows b2199 alignment with Mrub_837. BLASTp program described in methods.

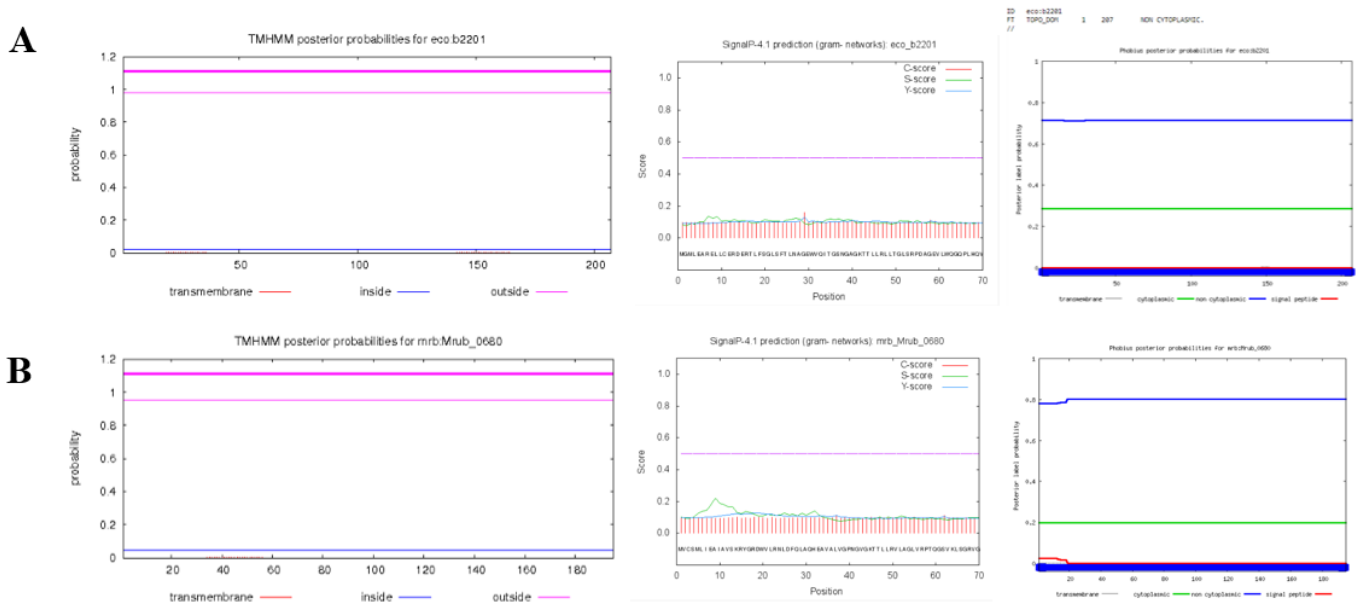

Fig 5. Cellular localization data from (right to left) TMHMM, SignalP, and Phobius confirming that both b2201 (A) and Mrub_0680 (B) have no transmembrane helices, are located in the cytoplasm, likely attached to the membrane and do not have any signal peptides. Programs are explained in Methods.

was predicted to be a cytosolic protein not containing any signal peptide by LipoP. PSORTb predicted that Mrub_0680 and b2201 are located in the cytoplasmic membrane with scores of 9.82 and 7.88 respectively. b2201 had a cytoplasmic score of 2.11, all other location scores were insignificant. The best prediction for the location of b2201 and Mrub_0680 is the cytoplasm attached to the inner membrane. The domains were found to be conserved between b2201 and Mrub_0680.

The top COG and TIGRfam hits for b2201 and Mrub_0680 were matches for CcmA with e-values as shown in Table 1. The top Pfam hit was for a family of ABC transporters, e-values shown in Table 1. b2201 and Mrub_0680 also had the same clan hit within the Pfam database representing the P-loop NTPase domain (CL0023). All top hits matched between b2201 and Mrub_0680 searches and had significant e-values suggesting similar function. When run in the PDB database, several domains were listed among both genes as having significant alignment. The top hit present in both was 1Z47, an ATP binding subunit of CysA ABC protein. Though it is not the exact protein in question, it is likely that the crystalized structure (Fig 7) shares similarities with the structure of the CcmA protein in question. TIGRfam, Pfam, and PDB data conclude that b2201 and Mrub_0680 contain an ATP binding domain, suggesting CcmA is directly involved in the active transport of heme via ATP hydrolysis.

Fig 6. WebLogo created from the pairwise alignment generated by T-Coffee of the orthologs of Mrub_0680.

Table 1. Description of the orthologs E. Coli b2201 and Mrub_0680. All programs described in methods.

Bioinformatics tool used	***E. coli* b2201**	***M. ruber* Mrub_0680**
BLAST *E. coli* against *M. ruber*	Score: 94.7 bits E-value: 2e-25	
CDD Data (COG category)	COG1131 CcmA; ABC-type multidrug transport system	
	E-value: 1.9e-101	E-value: 4.20e-51
Cellular Localization	Periplasmic space, attached to inner membrane	
TIGRfam – protein family	TIGR01189 CcmA: heme ABC exporter, ATP-binding protein	
	E-value: 1.4e-134	E-value: 6.9e-27
Pfam – protein family	PF00005 (ABC Transporter) CL0023 (P-loop NTPase)	
	E-value: 2.6e-32	E-value: 8.5e-24
PDB	1Z47 (ATPase subunit CysA: ATP-binding cassette)	
	E-value: 1.801e-5	E-value: 1.72e-20
KEGG pathway map	Heme ABC Transporter	

3.2 | CcmB: Mrub_0836 and b2200

The start codon of Mrub_0836 was confirmed as being correct from the KEGG sequence given using T-Coffee pairwise alignment. The alignment showed that across several relates species, the M1 of *M. ruber* was highly

Fig 7. Structure of the ATPase subunit CysA of the putative sulfate ATP-binding cassette (ABC) transporter from Alicyclobacillus acidocaldarius as provided by PDB.

```
Anaerolineae_bacterium          MNF-------------------WRSVMAIVWKDIRAELR
Deinococcus_misasensis          MR----------------------EALLIALKDLRLEGR
Gemmatimonadetes_bacterium      M-----------------------KLAWVVARKDLLLEFR
Marinithermus_hydrothermalis    ----------------------MRRVWALAARDLKLELR
Meiothermus_cerbereus           ----------------------MQRIFWLAWRDLVLEFR
Meiothermus_ruber               ----------------------MQRIFWLAWRDLVLEFR
Meiothermus_silvanus            ----------------------MRRIFWLAWRDLTLELR
Nitrospina_gracilis             MNGY------------------FQQIGAIAAKDFSTEFK
Oceanithermus_profundus         ----------------------MR-ALVLALRDLRLEWR
Thermus_antranikianii           MKLPV---EGWPEVSGQGRVGAVRRVVLLALRDLRLEVR
Thermus_thermophilus            ----------------------MRRVWLLALRDLRLEVR
Truepera_radiovictrix           MVKARTKVASRP-ASGSTPWGDAVAVWAVARKDLLLELR
                                                       :. :*:  * :
```

Fig 8. Multispecies pairwise alignment of Mrub_0836 and related orthologs showing likely start codon. Created using T-Coffee, described in methods.

Fig 9. WebLogo created from the pairwise alignment generated by T-Coffee of the orthologs of Mrub_0836.

conserved in relation to the other species (Fig 8). A few species, *D. misasensis* and *G. bacterium*, were misaligned at the start. The MR and M at the far left of the alignment could have been shifted to the right of the gaps to have the M line up with the majority of the other species.

Alternative start codons proposed by IMG/M were analyzed and determined not to be convincing based on the T-Coffee alignment data being so significant for the M1 start codon. The Weblogo associated with these species reinforced the data that the correct start codon was chosen (Fig 9).

Fig 10 shows the proposed transmembrane domains of b2200 and Mrub_0836. These proteins were found to be transmembrane proteins with six transmembrane helices. No signal peptide was found for b2200 or Mrub_836 through either SignalP or LipoP. The best prediction for cellular location is the cytoplasmic membrane with a PSORT score of 10.0 for both b2200 and Mrub_0836. LipoP also predicted that b2200 and Mrub_836 are located in the membrane.

Additional functionality research identified common COG, TIGR, and Pfam domains for b2200 and Mrub_0836. These genes contained domains relating to ABC-transport and those previously identified as being part of heme transporter complex. All e-values and domains are listed in Table 2. Both b2200 and Mrub_0836 belong to the same clan, ABC-2 (CL0181). No significant matches were obtained from the PBD search for either gene. Data from Pfam and TIGRfam confirm that Mrub_0836 is correctly determined as protein CcmB involved in heme transport. CDD data indicated that this particular protein in involved in transport. Together these data suggest that Mrub_0836 codes for CcmB which is involved in heme export to the periplasmic space.

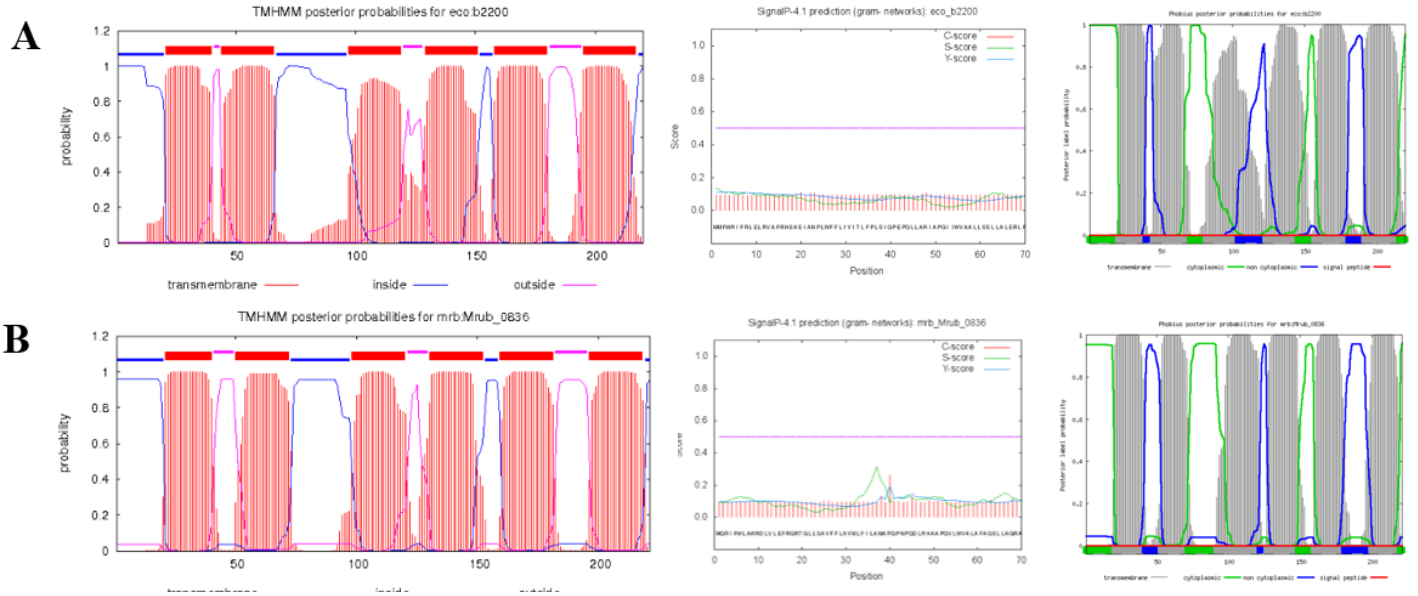

Fig 10. Cellular localization data from (right to left) TMHMM, SignalP, and Phobius confirming that both b2200 (A) and Mrub_0836 (B) are transmembrane proteins that have 6 transmembrane helices indicated by red and gray peaks on the TMHMM and Phobius plots, respectively, and do not have any signal peptides. Programs are explained in Methods.

3.3 | CcmC: Mrub_0837 and b2199

The start codon for Mrub_0837 was confirmed using pairwise alignment of multiple species (Fig 11) which showed all related organisms having the same M conserves within the orthologs to Mrub_0837. This was further supported by WebLogo data (Fig 12) which showed the highly conserved start codon M1. Additionally, a Shine-Dalgarno sequence was identified 9 positions upstream of the start codon by IMG/M.

Table 2. Description of the orthologs *E. Coli* b2200 and Mrub_0836. All programs described in methods.

Bioinformatics tool used	***E. coli* b2200**	***M. ruber* Mrub_0836**
BLAST *E. coli* against *M. ruber*	Score: 52.0 bits E-value: 3e-09	
CDD Data (COG category)	COG2386 ABC-type transport system involved in cytochrome c biogenesis	
	E-value: 4.93e-76	E-value: 1.83e-24
Cellular Localization	Cytoplasmic/inner membrane	
TIGRfam – protein family	TIGR01190 CcmB: heme exporter protein CcmB	
	E-value: 1.9e-145	E-value: 8.3e-08
Pfam – protein family	PF03379 CcmB protein	
	E-value: 4.9e-92	E-value: 2.7e-39
PDB	No significant matches found	
KEGG pathway map	Heme ABC Transporter	

```
Deinococcus_gobiensis          MTRD--L--------TTRVLGGLTLLLLLIA
Deinococcus_pimensis           MKD---R--------LTLPLGLLTLAAFAVG
Marinithermus_hydrothermalis   MMNR--VASDRGLDRLTLGLLVAGFGVLAVG
Meiothermus_cerbereus          MQLAHSNQTN-RLDGLTLGILGLGVVGAVVG
Meiothermus_ruber              MQLARSNQTN-RLDGLTLGLLGLGVVGAVVG
Myxococcus_stipitatus          MNKL----------VKWGLPIVGLAVLGFG
Oceanithermus_profundus        MNHA--EKTS-RLDSASRALLTLALVVFAVG
Pseudonocardia                 MTDH--AVTVSARAVFGHRLPVAAGLAGAAG
Streptomyces_zinciresistens    MTL----------FGRRLPIATALVTAIA
Thermus_aquaticus              MLKA--ANPD-RPDLLTWAFLALGL-ALPVG
Thermus_arciformis             MLNA--AHPE-RPDALTWVFLGLGLLLLPLG
Thermus_brockianus             MLKA--AHPE-RPDALTWAFLGLGVLLLPVG
Thermus_caliditerrae           MLKT--AQPE-RPDTLTWVFLGLGLLLLPVG
```

Fig 11. Multispecies pairwise alignment of Mrub_0837 and related orthologs showing likely start codon. Created using T-Coffee, described in methods.

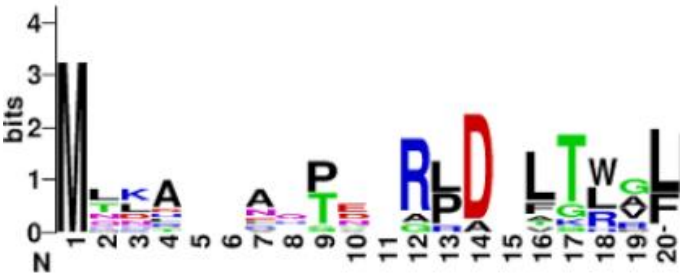

Fig 12. WebLogo created from the pairwise alignment generated by T-Coffee of the orthologs of Mrub_0837.

Fig 13 shows the transmembrane domains of b2199 (13A) and Mrub_0837 (13B). The data shows that these proteins each contain six transmembrane helices and lack signal peptides. Additional data from LipoP confirmed the lack of signal peptides for both genes and suggested the cellular membrane as a possible location. The best prediction for cellular location is the cytoplasmic membrane with a PSORT score of 10.0 for both b2200 and Mrub_0836. LipoP also predicted that b2199 and Mrub_0837 are located in the membrane.

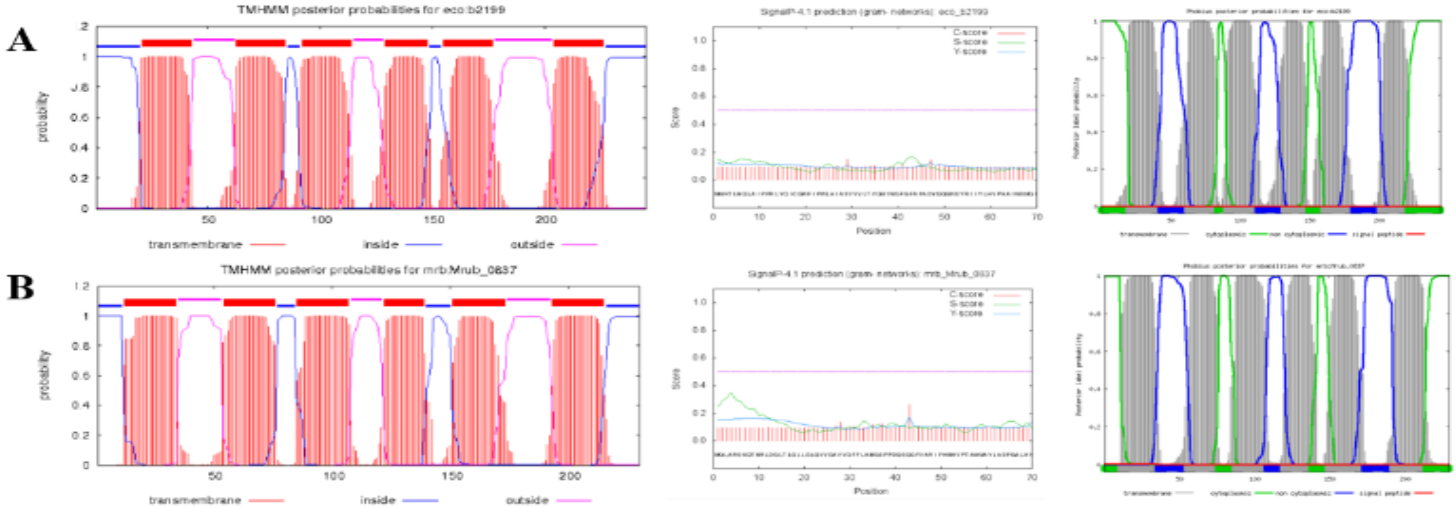

Fig 13. Cellular localization data from (right to left) TMHMM, SignalP, and Phobius confirming that both b2199 (A) and Mrub_0637 (B) are transmembrane proteins that have 6 transmembrane helices indicated by red and gray peaks on the TMHMM and Phobius plots, respectively, and do not have any signal peptides. Programs are explained in Methods.

Looking into the functional families of b2199 and Mrub_0837, matching hits were found for COG, TIGRfam, and Pfam families. The e-values for each can be seen in Table 3. In addition to the same Pfam match, b2199 and Mrub_0837 were also found to be part of the same clan, 2heme cytochrome (CL0328). All categorizations suggest that b2199 and Mrub_0837 code CcmC and are integral membrane proteins involved in transport. No PDB matches were found to be significant.

Table 3. Description of the orthologs *E. Coli* b2199 and Mrub_0837. All programs described in methods.

Bioinformatics tool used	***E. coli* b2199**	***M. ruber* Mrub_0837**
BLAST *E. coli* against *M. ruber*	Score: 105 bits E-value: 2e-28	
CDD Data (COG category)	COG0755 CcmC, permease component	
	E-value: 7.77e-93	E-value: 1.67e-27
Cellular Localization	Cytoplasmic/inner membrane	
TIGRfam – protein family	TIGR01191 CcmC: heme exporter protein CcmC	
	E-value: 4.9e-135	E-value: 1.3e-40
Pfam – protein family	PF01578 Cytochrome C assembly protein	
	E-value: 1.3e-32	E-value: 5.8e-27
PDB	No significant matches found	
KEGG pathway map	Heme ABC Transporter	

Table 4. Description of *E. Coli* b2201 which does not have an ortholog in *M. ruber*. All programs described in methods.

Bioinformatics tool used	***E. coli* b2198**
BLAST *E. coli* against *M. ruber*	No paralog found in *M. ruber*
CDD Data (COG category)	COG3114 CcmD, Heme exporter protein D
	E-value: 1.17e-21
Cellular Localization	Cytoplasmic/inner membrane
TIGRfam – protein family	TIGR03141 heme exporter protein CcmD
	E-value: 4.2e-21
Pfam – protein family	PF04995 CcmD; Heme exporter protein D
	E-value: 1.9e-16
PDB	No significant matches found
KEGG pathway map	Heme ABC transporter

3.4 | CcmD: b2198

As already mentioned, b2198 codes for CcmD and has no ortholog in *M. ruber*. Fig 14 shows data collected from TMHMM and Phobius indicate that b2198 contains a single transmembrane domain but no signal peptide as indicated by the SignalP plot. LipoP analysis confirmed the lack of a signal peptide and suggested the membrane as the protein's location within the cell. PSORT analysis also suggested the cytoplasmic membrane as the location with a score of 9.82, all other scores were insignificant.

Table 4 shows the best hits for COG, TIGRfam, and Pfam families with their corresponding e-values. No PDB matches were found for b2198. All data link b2198 to CcmD, a heme exporter protein.

3.5 | Paralogs

When examined for paralogs, only b2201 and Mrub_0680 came back with significant results as visualized by KEGG in Fig 15. The top hit for b2201 was b0262 which codes for AfuC, an ATP-binding component

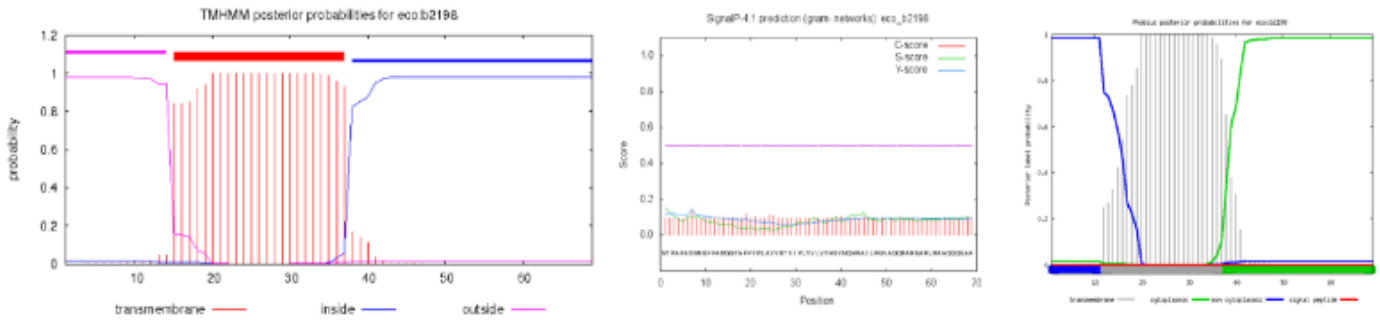

Fig 14. Cellular localization data from (right to left) TMHMM, SignalP, and Phobius confirming that is transmembrane proteins containing one transmembrane helix indicated by red and gray peaks on the TMHMM and Phobius plots, respectively, and does not have any signal peptides. Programs are explained in Methods.

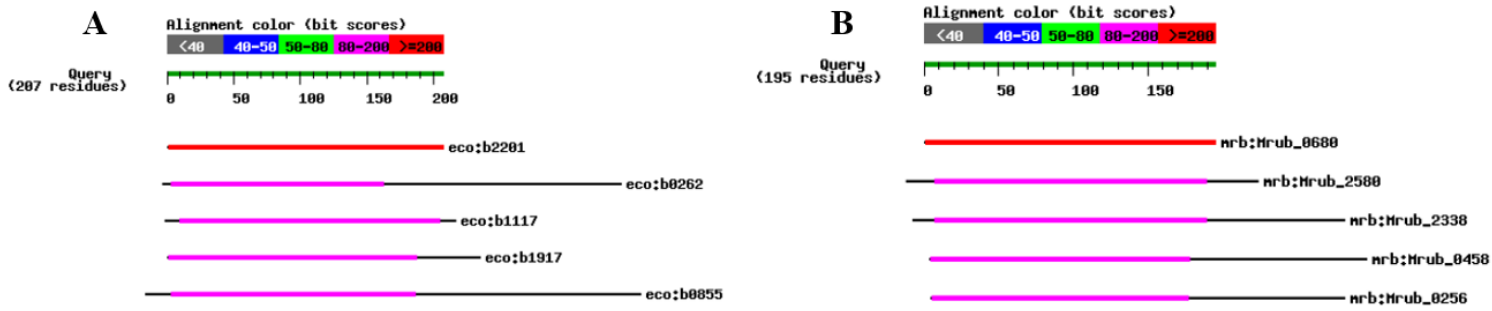

Fig 15. Alignment of b2201 (Panel A) and Mrub_0680 (Panel B) in red with multiple possible paralogs in pink for each GOI. Alignment performed using BLAST tool from KEGG as described in methods.

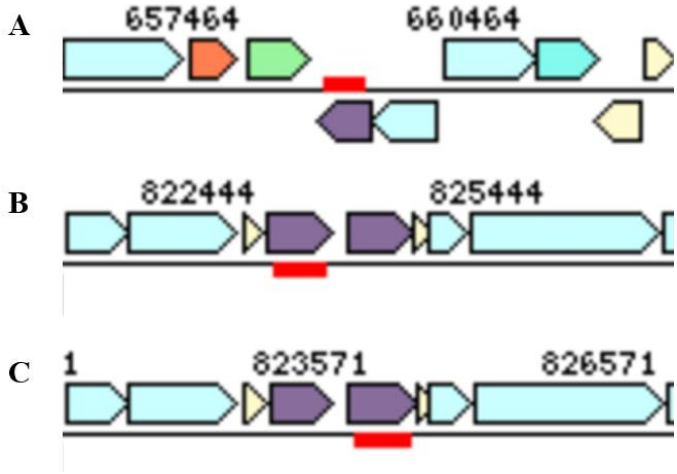

Fig 16. Nearby areas of Mrub_0680 (A), Mrub_0836 (B), and Mrub_0837 (C) colored by KEGG; genes are indicated by the red bar.

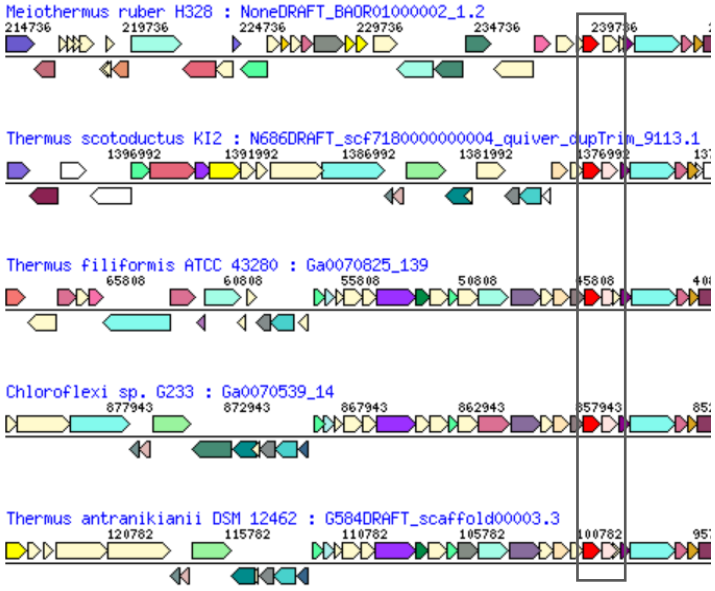

Fig 17. Neighborhood regions with the same top COG hit for multiple species closely related to *M. ruber* as taken from IMG/M. Operon for Mrub_0836 (red) and Mrub_0837 (white) are visibly conserved as shown by the box.

of an ABC transporter, with an e-value of 1e-21. The top hit for Mrub_0680 was Mrub_2580, an ATP-binding protein involved in liposaccharide export, with an e-value of 2e-31. These results are not surprising given that ATP-binding domains are not overly specific and are very common within ABC transport systems.

3.6 | *Meiothermus ruber* Operon

Gene locations were examined using Color by KEGG to see the relationship to genes upstream and downstream of the GOI (Fig 16). Mrub_680 was bordered by Cytochrome c biogenesis protein CcmA (Mrub_0681) and short-chain dehydrogenase/reductase SDR (Mrub_0679), both unrelated to the function of Mrub_680. Mrub_0836 and Mrub_0837 are next to each other and directly upstream and downstream is located a hypothetical protein (white in Fig 16). This relationship of Mrub_0836 and Mrub_0837 is conserved between many different related species as seen in Fig 17, highlighted within the box.

4 | Conclusions

4.1 | Mrub_0680 codes for CcmA and is orthologous to b2201

The locus Mrub_0680 was determined to code for protein CcmA which is a component of the transport complex previously linked to heme transport in Cyt-C biosynthesis. CcmA contains an ATP-binding domain and assists in the transfer of a heme cofactor to CcmE with the assistance of energy from ATP. This same gene is found in *E. coli* at locus b2201 and performs the same function in the Cyt-C biosynthesis pathway. CcmA does not contain any transmembrane helices or signal peptides and is found in the cytoplasm attached to the inner side of the cytoplasmic membrane. The start codon chosen by KEGG for Mrub_0680 was shown to be incorrect. The alignment from T-Coffee indicated that most other species have a different start codon that matches up better with the M5 residue of the *M. ruber* sequence rather than the M1 initially chosen. Additionally, there is evidence that suggests that CcmA may bind to the CcmBC transporter complex during the mechanism, but research has yet to show conclusive evidence. CcmA contains an ATP-binding domain recognized by PDB that was present in both orthologs and is very common in ABC-transport complexes since they are driven by energy input in the form of ATP hydrolysis.

4.2 | Mrub_0836 codes for CcmB and is orthologous to b2200

Mrub_0836 codes for CcmB, a subunit of an ABC transporter associated with Cyt-C biosynthesis. CcmB is also found in *E. coli* at locus b2200. This protein contains six transmembrane helices and forms a complex with CcmC to make a complete export, originally associated the heme factor of Cyt-C. This has since been disproven but no substrate has been determined. The function of CcmB as encoded by both Mrub_0836 and b2200 has been confirmed with the low e-values of the Pfam, TIGRFam, and COG hits that were the same for both genes. The start codon for Mrub_0836 is correct as taken from KEGG. Though there is no Shine-Dalgarno sequence in the near vicinity upstream of the start codon, the M chosen aligns well with most of the paralogs in other species.

4.3 | Mrub_0837 codes for CcmC and is orthologous to b2199

It was found that Mrub_0837 codes for the protein CcmC involved as half of the ABC transporter complex involved in Cyt-C biosynthesis. This gene is also found in *E.* coli at locus tag b2199. All protein function matches were consistent between Mrub_0837 and b2199 with very low e-values indicating a significant match. CcmC contains six transmembrane helices and no signal peptide, making its home in the cytoplasmic membrane. CcmC forms a complete exported with CcmB and is linked to heme expor. However, more recent studies have determined that the CcmBC complex does not transport heme and they are uncertain what its substrate it. More research is required to determine this. Additionally, the start codon was confirmed for Mrub_0837 as being M1 as taken from KEGG database. This M1 was highly conserved between several closely related species to *M. ruber*.

4.4 | There was no ortholog in *M. ruber* for b2198 which codes for CcmD

The data in Fig 14. shows conclusively that b2198 codes for a transmembrane protein with a singular trans-membrane helix. It does not contain any signal peptide and lives in the cytoplasmic membrane. b2198 codes for CcmD which assists CcmA with the heme transfer to CcmE in the biosynthetic pathway of Cyt-C. All protein functional matches described in results has highly significant e-values (all < 1e-15) which strongly confirms the association with CcmE and the pathway of Cyt-C biosynthesis. As mentioned already, there was no ortholog for CcmD found in *M. ruber*. This suggests that though CcmD is helpful in the heme transfer from CcmC to CcmE, it is not crucial for function of this complex. *M. ruber* likely lost this gene in the past and since it was a nonessential protein, survived with the incomplete system. There were no paralogs found for b2198 in *E. coli*.0020

4.5 | Overview

Our initial hypothesis was confirmed showing that Mrub_0680, Mrub_0836, and Mrub_0837 code for proteins CcmA, CcmB, and CcmC respectively. Additionally, these proteins have orthologs in *E. coli* that represent the same system. It is known that these genes are all part on a single operon in *E. coli*, however it was determined that this is not the case for *M. ruber*. For *M. ruber*, the genes for CcmB and CcmA are part of an operon that is present in closely related species. However, the locus for CcmA is much further upstream and separate from the other two proteins in the system. This could be explained by a duplication in the gene followed by the deletion at the original locus, or the addition of several genes between the two locations. Research has yet to determine the cause of the separation of these genes. The only paralogs found for all GOIs were for CcmA in both *E. coli* and *M. ruber*. This is not surprising since the ATP-binding domain of CcmA is very common to find in ABC transporters and the transporter complex, CcmBC, is highly specified so no paralogs is expected.

4.6 | Future Directions

For future research, we suggest looking into the key amino acids in the mechanism of the protein. One possible target is H53 of Mrub_0837. This residue has been shown to be highly conserved by the HMM logo created based on data using the Pfam domain (PF01578) as visualized in Fig 18.

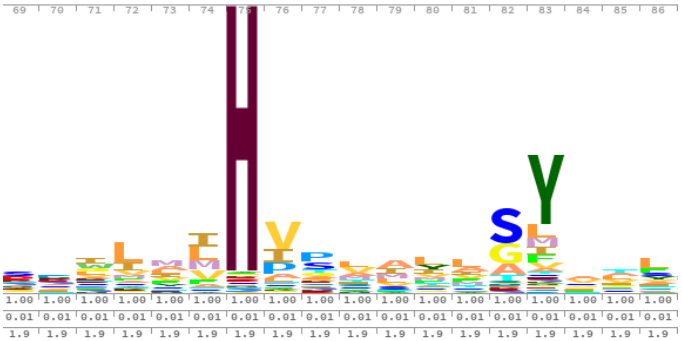

Fig 18. HMM logo generated for Pfam Cytochrome C assembly protein (PF01578) showing the highly conserved H residue which corresponds to H53 of Mrub_0837.

To do this, primers would be obtained using NEBase-Changer in order to change H53 to F; these primers are shown in Fig 19A (NE Biolabs). However, for Mrub_0837, there is another H at position 51 which may take over the role that H53 fills in the mechanism if only H53 is mutated. We propose that another mutated form be induced and tested that replaces both H51 and H53 with F to determine is the loss of the hypothesized catalytic H results in loss of function of the CcmC protein encoded by Mrub_0837; primers for this are listed in 19B (NE Biolabs). A

A

Name (F/R)	Oligo (Uppercase = target-specific primer)	Len	% GC	Tm	Ta *
Q5SDM_2/16/2018_F	CTTCCATATGttcGTCCCAACGGCCTGGATG	31	55	69°C	70°C
Q5SDM_2/16/2018_R	ATGCGGGCCACAAAGCCC	18	67	72°C	

B

Name (F/R)	Oligo (Uppercase = target-specific primer)	Len	% GC	Tm	Ta *
Q5SDM_2/16/2018_F	gttcGTCCCAACGGCCTGGATG	22	64	69°C	68°C
Q5SDM_2/16/2018_R	ataaaGAAGATGCGGGCCACAAAG	24	46	67°C	

Fig 19. Forward and reverse primers for site directed mutagenesis of Mrub_0837 substituting H53 for F (A) and Substituting H51 for F and H53 for F. Primers obtained from NEBaseChanger (NE Biolabs).

substitution from H to F was chosen because both residues have similar shape and size so not it interfere with the stereochemistry of the possible active site, and also F is not likely to act as an acid and donate a proton which is the most likely function of the H residue in the protein (Betts and Russel, 2003). The primers would be used to perform site directed mutagenesis in hope to determine the role and importance of the H53 residue.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. [cited 18 Jan 2018];215:403-410.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000 [updated 2016 Dec 6]. The Protein Data Bank. [cited 16 Jan 2018]. Available from: http://www.rcsb.org/

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000 The Protein Data Bank Nucleic Acids Research, 28: 235-242.

Betts MJ, Russell RB. 2003. Amino-Acid Properties and Consequences of Substitutions. Bioinformatics for Geneticists [Internet]. [cited 15 Nov 2017]:311–342.

Biolabs NE. Reagents for the Life Sciences Industry | NEB. New England Biolabs: Reagents for the Life Sciences Industry. [accessed 2018 Feb 8]. https://www.neb.com/

Crooks, GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator, Genome Research. [cited 18 Jan 2017];14:1188-1190.

Feissner RE, Richard-Fogal CL, Frawley ER, Kranz RG. 2006. ABC transporter-mediated release of a haem chaperone allows cytochrome c biogenesis. Mol Microbiol. [cited 8 Feb 2018];61(1):219-231. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2006.05221.x/abstract

Ferguson SJ, Stevens MJ. 2007. Loss of ATP hydrolysis activity by CcmAB results in loss of c-type cytochrome synthesis and incomplete processing of CcmE. The FEBS Journal. [cited 2018 Feb 6];274(9):2322–2332.

Finn RD, Bateman A, Clements J, et al. 2014. Pfam: the protein families database. Nucleic Acids Research. [cited 10 Jan 2018];42 (Database issue):D222-D230. doi:10.1093/nar/gkt1223.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future: Nucleic Acids Res. [cited 15 Jan 2018];44:D279-D285. Available from: http://pfam.xfam.org/

Goldman BS, Kranz RG. 2001. ABC transporters associated with cytochrome c biogenesis. Res Microbiol. [cited 8 Feb 2018];152(3-4):323-329. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11421279

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. [cited 6 Feb 2018];29(1):41-3.

Huarng M, Hunyh S. 2011. Meiothermus [Internet]. Microbe Wiki. [cited 7 Feb 2018]. Available from: https://microbewiki.kenyon.edu/index.php/Meiothermus#Description_and_Significance

Juncker A, Willenbrock H, von Heijne G, Nielsen H, Brunak S, Krogh A. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. [cited 16 Dec 2017];12(8):1652-62. Available at: http://www.cbs.dtu.dk/services/LipoP/

Kall L, Krough A, Sonnhammer E. 2004. A combined transmembrane topology and peptide prediction method. 2004. Journal of Molecular Biology. [cited 16 Jan 2017];338(5):1027-1036.

Käll L, Krogh A, Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res., [cited 16 Jan 2018];35:W429-32.

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., [cited 13 Dec 2017];44, D457–D462. Available from: http://www.genome.jp/kegg/

Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, et al. 2013. EcoCyc: fusing model organism databases with systems biology Nucleic Acids Research [cited 11 Dec 2017];41:D605-612.

Krogh, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology. [cited 20 Jan 2018];305(3):567-580.

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

Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US). 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

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. 2014. CDD: NCBI's conserved domain database. Nucleic Acids Res. [cited 15 Jan 2015];43:D222-2. doi: 10.1093/nar/gku1221. Epub. [PubMed PMID: 25414356]

Markowitz VM, Chen IA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, et al. 2012. IMG: The integrated microbial genomes database and comparative analysis system. Nucleic Acids Research [cited 1 Feb 2018];40(D1):D115-22. Available from: http://nar.oxfordjournals.org/content/40/D1/D115.full

Notredam C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for multiple sequence alignments. J Mol Biol. [cited 8 Feb 2018];302:205-217.Available from: https://www.ncbi.nlm.nih.gov/pubmed/10964570

Petersen T, Brunak S, von Heijne G, Nielsen H. 2011. Discriminating signal peptides from transmembrane regions. Nature Methods, [cited 1 Feb 2018];8:785-786, 2011 Available from: http://www.cbs.dtu.dk/services/SignalP

Sonnhammer ELL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. In: Glasgow J, Littlejohn T, Major F, Lathrop

Tindall BJ, Sikorski J, Lucas S, Goltsman E, Copeland A, Del Rio TG, Nolan M, Tice H, Cheng JF, Han C, et. al. 2010. Complete genome sequence of Meiothermus ruber type strain (21T). Standards in Genomic Sciences [Internet]. EBSCO Industries; [cited 15 Nov 2017];3:26–36.

Thony-Meyer L. 1997. Biogenesis of respiratory cytochromes in bacteria. PubMed [cited 8 Feb 2018];61(3):337-376.

The UniProt Consortium. 2016. UniProt: the universal protein knowledgebase. [cited2 Feb 2018];45:D158-D169 Available from: https://academic.oup.com/nar/article/45/D1/D158/2605721

Wilkens S. 2015. Structure and mechanism of ABC transporters. F1000Prime Reports. [cited 3 Feb 2018];7:1-9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4338842/

Yeagle PL. 2016. Membrane proteins. The Membranes of Cells. 3rd Ed. Storrs(CT): Academic Press. P. 219-268.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster JL, Brinkman FSL. 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. [cited 18 Jan 2018];26(13):1608-1615.