2018

Mrub_1325, Mrub_1326, Mrub_1327, and Mrub_1328 are orthologs of B_3454, B_3455, B_3457, B_3458, respectively found in *Escherichia coli* coding for a Branched Chain Amino Acid ATP Binding Cassette (ABC) Transporter System

Bennett Tomlin
Augustana College, Rock Island Illinois

Adam Buric
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_1325, Mrub_1326, Mrub_1327, and Mrub_1328 are orthologs of B_3454, B_3455, B_3457, B_3458, respectively found in *Escherichia coli* coding for a Branched Chain Amino Acid ATP Binding Cassette (ABC) Transporter System

Bennett Tomlin and Adam Buric

Introduction:

All of the genes in this paper are a part of a class of proteins called ABC transporters. ABC stands for ATP Binding Cassette, and they are called this because in order to transport anything ATP binds at the "cassette". All life depends on transporters in order to maintain the appropriate levels of different materials. ABC transporters are the largest family of proteins discovered so far representing approximately 2% of the bacterial genomes. (Tomil and Kenhisa 1998) Their ubiquity, including among higher organisms, has led to them being implicated in several diseases including cystic fibrosis. (Vasiliou and Vasiliou and Nebert 2009) ABC transporters are divided into three main categories including two classes of importers and a single class of exporter. (Wilkens 2015) Exporters tend to be closely related phylogenetically to other exporters, and importers to other importers, as those two types likely diverged once a long time ago in the evolution of these systems. (Saurin, Hofnung, and Dassa 1999) There are several highly conserved features for these transporters including: a consensus motif (LSGGQ), and a phosphate binding loop. (Wilkens 2015) There are also several prototypical structural features including: two binding "cassettes", two transmembrane domains, and up to twenty transmembrane helices. (Holland and Blight 1999)

The specific genes discussed in this study are a series of proteins involved in branched-chain amino acid transport. This is an important set of genes to study because they

have been indicated as importance for the virulence of several strains of pathogens that can affect humans, and interference of the genes could potentially prevent infection. (Basavanna et al. 2009)

System in *E. Coli*:

The system of *E. coli* genes used in this system are B_3454, B_3455, B_3456, B_3457, and B_3458. They code for LivF, LivG, LivM, LivH, and LivK, respectively. This system of genes in *E. coli* code for a set of ABC transporters involved in branched chain amino acid transport. Figure 1 depicts the system as a whole. Figure 2 depicts the operon coding for the genes discussed in this paper in E coli.

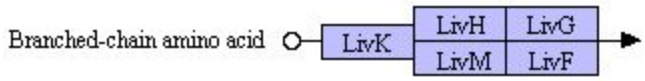

Figure 1: Depiction of the Branched Chain Amino Acid ABC Transporter system. (Kanehisa et al 2016)

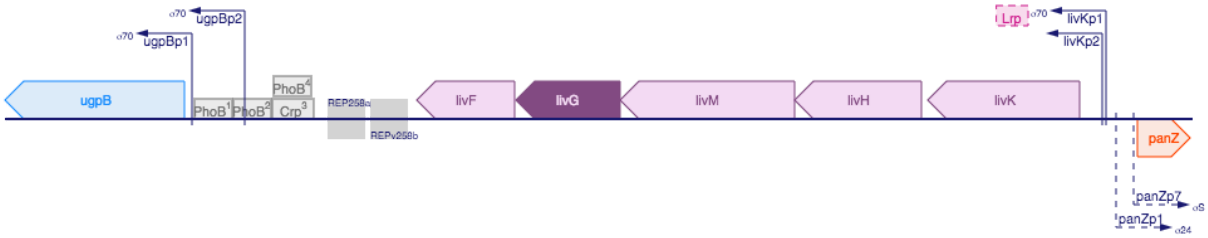

Figure 2: Depiction of the *E. coli* operon coding for this set of genes (Keseler et al. 2013). All *liv* genes are under identical transcriptional control.

***Meiothermus Ruber* as Study Organism:**

Meiothermus ruber, first discovered in 1975, is a gram negative bacteria normally found in warmer temperatures. (Loginova and Egorova 1975, Tindall et al 2010) Its shape is rod, and it produces a distinctive red pigmentation, with an optimal growth temperature of 60 degrees

celsius.(Loginova et al 1984) *Meiothermus ruber* became a much more widely studied organism after its genome sequencing in 2010. (Tindall et al 2010) *Meiothermus ruber* was discovered to have a GC (guanine and cysteine) concentration of over 63%, which would make sense in light of the greater thermostability needed for an organism in these temperature conditions. (Tindall et al 2010) The maximum temperature that it seems to be viable at seems to be around 90 degrees celsius. (Loginova and Egorova 1975) The lowest viable temperature is approximately 40 degrees celsius. (Loginova et al 1984). From our experiments in the laboratory, we know that occasionally genes inserted from *E. coli* will not function in *M. ruber* due to the higher temperature needed for incubation. (Dr. Scott Personal Communication)

Purpose: To determine if Mrub_1325, Mrub_1326, Mrub_1327, and Mrub_1328 are orthologs of B_3454, B_3455, B_3456, B_3457, or B_3458.

Hypothesis: Mrub_1325, Mrub_1326, Mrub_1327, and Mrub_1328 are orthologs of B_3454, B_3455, B_3457, B_3458, respectively.

Materials and Methods:

Initial identification of potential orthologs: Initially, the assigned system of ABC transporters was located in KEGG, and the sequences of both the *E. coli* coding genes and the *M. ruber* coding genes were determined. (Kanehisa et al. 2013). KEGG contains pathway maps along with information on both the *M. ruber* and *E. coli* variants of these genes and their sequences. These sequences were then analyzed using BLASTp in order to help further elucidate potential orthologs. (Altschul et al. 1990; Madden et al. 2002) In order to elucidate these orthologs the sequence of the *E. coli* genes were blasted against the genome of *M. ruber* or vice versa. This

allowed for similar sequences to be determined from the database in order to help identify potential orthologs.

Determination of proper start codon: Sequences were then analyzed using IMG/M in order to ensure there was an appropriately placed Shine-Dalgarno sequence before the called start codon and to identify other potential start codons. (Markowitz et al 2012) If upon analyzing the start codon a different one was selected, the sequence beginning at the start codon would become the sequence that would be analyzed for the remainder of the experiment.

Determination of highly conserved amino acids: For each of the amino acids they are again analyzed using BlastP (Altschul et al. 1990; Madden et al. 2002), with potential orthologs selected from several different genuses and species, and their sequences are downloaded. These sequences are then aligned using T-Coffee. (Notredame and Higgins and Heringa 2000) This tool outputs aligned sequences that can then be input into WebLogo which will produce a logo demonstrating the degree to which different amino acids are conserved. (Crooks et al. 2004) This provides a visual tool to better understand the important and highly conserved features.

Determination of Cellular Localization: In order to better confirm that these genes were ortholog their cellular localization was determined. This began by determining the number of transmembrane helices for each gene by using TMHMM. (Krogh et al. 2001; Krogh and Rapacki 2016; Sonnhammer and von Heijne and Krogh 1998). Orthologous genes would be expected to have the same number of transmembrane helices. This is followed by determining

whether each gene is likely a signal peptide using SignalP. (Petersen et al. 2011) LipoP can then be used if there is a signal peptide to determine what type of signal peptide it is and the most likely cleavage site. (Juncker et al. 2003) PSORTB is then used to determine the most likely cellular location of the protein product of each gene. Orthologous genes would presumably have their protein products localized to the same part of the cell. (Yu et al. 2010) The bioinformatic tool Phobius can then be used to confirm much of the earlier localization data while also showing the localization probabilities along the sequence. (Kall and Krogh and Sonnhammer 2004; Kall, Krogh, and Sonnhammer 2007)

Analysis of Structural Similarity: In order to further confirm orthology structural features are compared between the various proposed orthologous genes. This begins with returning to BLASTp and now finding what it proposes as the highly conserved domains. (Altschul et al. 1990; Madden et al. 2002; Marchler-Bauer et al. 2014) The genes are then analyzed using TIGRFam in order to see if they belong to the same protein family. (Haft et al. 2001) This is followed by a comparison made using Pfam to determine again if they belong to the same family, but also to see if they belong to the same clan. (Finn et al. 2014, Finn et al. 2016) Finally PDB is used to determine whether the same, or very similar structures are returned, as we would expect for orthologous genes. (Berman et al. 2000a; Berman et al. 2000b)

Determination of Operons: Finally in order to better understand the system of genes as a whole an attempt is made to determine whether or not they are all part of an operon. IMG/M is used again for its chromosome viewer that allows for the operon to be colored by KEGG. (Kanehisa

et al. 2013; Markowitz et al 2012) This will aid in determining whether or not the sets of genes in a single organism function together as part of an operon.

Results:

Figure 3 summarizes the proposed orthologs in this study and shows which protein is coded for by each gene and their shared identity. Important to note that no ortholog was found for livM in the *Meiothermus ruber* genome.

M. ruber locus tag	*E. coli* ortholog locus tag	Protein	Function	Percent identity
mrub1325	b3454	livF	ABC Transporter	52%
mrub1326	b3455	livG		49%
N/A	b3456	livM		N/A
mrub1327	b3457	livH		41%
mrub1328	b3458	livK		35%

Figure 3: Summary of proposed orthologs, their shared identity, and the proteins they code for. Function of all genes studied is analogous: ABC transport. Percent identity shows percent similarity in nucleotide identity.

Figure 4 is another summary table showing the different legs of the various genes and the proteins they code for.

Protein	*M. ruber*			*E. coli*		
	Locus tag	Nucleotide length	Protein length (amino acids)	Locus tag	Nucleotide length	Protein length (amino acids)
livF	mrub1325	732	243	b3454	714	237
livG	mrub1326	1773	590	b3455	768	255
livM				b3456	1278	425
livH	mrub1327	1032	343	b3457	927	308
livK	mrub1328	1164	387	b3458	1110	369

Figure 4: Comparison information about length between *M. ruber* and *E. coli* orthologs.

Figure 5 depicts a summary of the different cellular localization data collected. The largest disparities exist between the two systems for livG. This may be because in *M. Ruber* livG is performing a function closer to that of a combined livG and livM.

Protein	*M. Ruber*				*E. Coli*			
	Locus Tag	# TMH	Signal Peptide	Cellular Localization	Locus Tag	#TMH	Signal Peptide	Cellular Localization
livF	mrub 1325	0	N/A	Cytoplasmic Membrane	b3454	0	N/A	Cytoplasmic Membrane
livG	mrub 1326	9	SP1	Cytoplasmic Membrane	b3455	0	N/A	Cytoplasm
livM					b3456	10	N/A	Cytoplasmic Membrane
livH	mrub 1327	8	N/A	Cytoplasmic Membrane	b3457	8	N/A	Cytoplasmic Membrane
livG	mrub 1328	1	SP1	Periplasm	b3458	1	SP1	Periplasm

Figure 5: Comparison between different cellular localization features of both sets of genes from E. *coli* and *M. ruber*. Number of transmembrane helices (TMH) between the orthologous genes in the two species is shown. Possession of signal peptide is compared between orthologous genes. Cellular localization differences show possible differences in functionality.

Figure 6 depicts some of the structural similarities between these. Important to consider here is that there are some significant differences, especially in terms of livG, which is likely again due to the large difference in length, and the lack of a livM protein in the *M. ruber* system.

Protein	*M. ruber*			E. coli		
	Locus Tag	PFam name	PDB best hit	Locus Tag	Pfam Name	PDB best hit
livF	mrub1325	ABC_tran	1JIO	b3454	ABC-tran	1JI0
livG	mrub1326	BPD_transp_2	1G6H	b3455	ABC tran	5L75
livM				b3456	BPD transp 2	N/A
livH	mrub1327	BPD_transp_2	N/A	b3457	BPD transp 2	N/A
livG	mrub1328	Peripla_BP_6	3TD9	b3458	Peripla BP 6	1USG

Figure 6: Summary of structural similarities for the orthologs. Protein assignment made based on Conserved Domains and all proteins were identified as Urea ABC transporters by TigrFam. Important to note is that both of these systems were operons.

Concluding Comments:

Identification of the orthologs for this study relied on several features including shared identity, conserved domains, structural similarities, and cellular localization. While we do believe we have identified the correct orthologous genes, there are some contradicting data we need to consider.

The pair of orthologs we propose are coding for livF are Mrub_1325 and B_3454. This identification was based initially on their shared identify of 52%. It was then supported by the determination that they are both localized to the cytoplasmic membrane. It was further

strengthened by the shared conserved domain coding for livF, the shared PFAM identification, and the shared most likely structure from PDB.

The pair of orthologs we propose are coding for livG are Mrub_1326 and B_3455. This identification was based initially on their shared identity of 49%. There was some contradictory information in the cellular localization as the *M. ruber* was identified as being localized to the cytoplasmic membrane, as the *E. coli* variant was localized to the cytoplasm. Further complicating the analysis they also differed on the number of transmembrane helices going from 9 in *M. ruber* to 0 in *E. coli*. Furthermore, there was a signal peptide sequence in *M. ruber* that dd not exist for this protein in *E. coli*. However, they were both identified as likely coding for this protein based upon their conserved domain. However, they also did have different PFAM names and different "most likely" PDB hits. Our strongest conclusion based upon this mix of conflicting evidence is that livG in *M. ruber* is having to serve the functions of both livG and livM in *E. coli* and this fact can help explain for many of the discrepancies.

The pair of orthologs we propose are coding for livH are Mrub_1327 and B_3457. This identification was based initially on their shared identify of 41%. It was then supported by the determination that they are both localized to the cytoplasmic membrane. It was further strengthened by the shared conserved domain coding for livH, and the shared PFAM identification.

The pair of orthologs we propose are coding for livG are Mrub_1328 and B_3458. This identification was based initially on their shared identify of 35%. It was then supported by the determination that they are both localized to the periplasm. Furthermore, they both have a signal peptide of the type SPI. It was further strengthened by the shared conserved domain coding for

livG, and the shared PFAM identification. While the PDB structures were differently they are somewhat related in that they are both forms of ABC transporters.

Site directed mutagenesis of Mrub_1325: When determining which amino acid should be mutated for a site directed mutagenesis study several things need to be considered. First and foremost it must be a highly conserved amino acid. Highly conserved amino acids are more likely to be important for either structure or function of the receptor. In this case the highly conserved valine located at position 22 will be mutated to alanine. This will require a change at nucleotide 17 from thymine to cysteine. (Betts and Russell 2003; Biolabs) The degree of conservation of the valine suggests that it could be a potential loss-of-function mutation. Changing it to an alanine is a relatively minor mutation that primarily removes a methyl group, potentially increasing the flexibility of the structure in that region. More aggressive mutations could also be tested, perhaps glycine which is another simple amino acid, but in this case would be missing an entire isopropyl group, potentially increasing the degree of change. It is always important to consider in what ways the amino acid substitution is likely to affect the structure, so the substitutions outlined here could be described as structural substitutions meant to elucidate the role of this valine in the structure. There could also be "functional" substitutions in which amino acids with different properties are substituted, for example different polarities or pH in order to determine how those features affect the protein as a whole. Figure 7 depicts the primers that would be used in order to make this site-directed mutagenesis study.

Name (F/R)	Oligo (Uppercase = target-specific primer)	Len	% GC	Tm	Ta *
Q5SDM_2/10/2018_F	GTGCTGGAAGcGCAAAACCTC	21	57	61°C	**62°C**
Q5SDM_2/10/2018_R	GCTCATTCAGACTTCGGTC	19	53	62°C	

Figure 7: Target-specific primers for site-directed mutagenesis of Mrub_1325 (Betts and Russell 2003; Biolabs).

References

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

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/.

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.

Betts, M. J., and Russell, R. B. 2003. Amino-Acid Properties and Consequences of Substitutions. Bioinformatics for Geneticists. 2003: 311-342.

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

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

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.

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

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

Holland IB and Blight MA. 1999. ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. J Mol Biol 293(2):381-99.

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/

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)

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)

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

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

Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology, 305(3):567-580, January 2001. (PDF, 959503 bytes)

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

Loginova LG and Egorova LA. 1975. [Thermus ruber obligate thermophilic bacteria in the thermal springs of kamchatka]. Mikrobiologiia 44(4):661-5.

Loginova LG, Egorova LA, Golovacheva RS, Seregina LM. 1984. Thermus ruber sp. nov., nom. rev. International Journal of Systematic and Evolutionary Microbiology 34(4):498-9.

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

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]

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

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1):205-17.

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

Saurin W, Hofnung M, Dassa E. 1999. Getting in or out: Early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. J Mol Evol 48(1):22-41.

Shilpa Basavanna, Suneeta Khandavilli, Jose Yuste, Jonathan M. Cohen, Arthur H. F. Hosie, Alexander J. Webb, Gavin H. Thomas, Jeremy S. Brown. 2009. Screening of streptococcus pneumoniae ABC transporter mutants demonstrates that LivJHMGF, a branched-chain amino acid ABC transporter, is necessary for disease pathogenesis. Infection and Immunity 77(8):3412-23.

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.

Tindall BJ, Sikorski J, Lucas S, Goltsman E, Copeland A, Glavina Del Rio T, Nolan M, Tice H, Cheng J, Han C, et al. 2010. Complete genome sequence of meiothermus ruber type strain (21). Standards in Genomic Sciences 3(1):26-36.

Tomii K and Kanehisa M. 1998. A comparative analysis of ABC transporters in complete microbial genomes. Genome Research 8(10):1048.

Vasiliou V, Vasiliou K, Nebert DW. 2009. Human ATP-binding cassette (ABC) transporter family. Human Genomics 3(3):281-90.

Wilkens S. 2015. Structure and mechanism of ABC transporters. F1000prime Reports 7:14.

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