J Membrane Biol (2011) 241:77–101
DOI 10.1007/s00232-011-9364-8

Bioinformatic Characterization of the Trimeric Intracellular Cation-Specific Channel Protein Family

Abe L. F. Silverio · Milton H. Saier Jr.

Received: 14 December 2010 / Accepted: 26 March 2011 / Published online: 26 April 2011

Abstract Trimeric intracellular cation-specific (TRIC) channels are integral to muscle excitation–contraction coupling. TRIC channels provide counter-ionic flux when calcium is rapidly transported from intracellular stores to the cell cytoplasm. Until recently, knowledge of the presence of these proteins was limited to animals. We analyzed the TRIC family and identified a profusion of prokaryotic family members with topologies and motifs similar to those of their eukaryotic counterparts. Prokaryotic members far outnumber eukaryotic members, and although none has been functionally characterized, the evidence suggests that they function as secondary carriers. The presence of fused N- or C-terminal domains of known biochemical functions as well as genomic context analyses provide clues about the functions of these prokaryotic homologs. They are proposed to function in metabolite (e.g., amino acid/nucleotide) efflux. Phylogenetic analysis revealed that TRIC channel homologs diverged relatively early during evolutionary history and that horizontal gene transfer was frequent in prokaryotes but not in eukaryotes. Topological analyses of TRIC channels revealed that these proteins possess seven putative transmembrane segments (TMSs), which arose by intragenic duplication of a three-TMS polypeptide-encoding genetic element followed by addition of a seventh TMS at the C terminus to give the precursor of all current TRIC family homologs. We propose that this family arose in prokaryotes.

Electronic supplementary material The online version of this article (doi:10.1007/s00232-011-9364-8) contains supplementary material, which is available to authorized users.

A. L. F. Silverio · M. H. Saier Jr. (✉)
Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093-0116, USA
e-mail: msaier@ucsd.edu

Keywords TRIC channel · Sarcoplasmic reticulum · Prokaryotic homolog · Potassium metabolite transport · Evolutionary origin · Topology

Introduction

Trimeric intracellular cation-specific (TRIC) channels are critical for proper management of intracellular Ca^{2+} stores and successful excitation–contraction (E–C) coupling in animals (Yazawa et al. 2007). The family of these channel proteins (TRIC, TC 1.A.62) is divided into two mammalian TRIC channel subtypes: TRIC-A and TRIC-B. Although similar in structure and biochemical function, TRIC-A and TRIC-B have distinctive properties (Yazawa et al. 2007). As seen in the Conserved Domain Database (CDD), both subtypes feature proteins that contain the TRIC conserved domain. Characteristic of all TRIC subtypes is their permeability to monovalent ions, with a preference for potassium. As putative ion channels, TRIC proteins translocate these monovalent ions across intracellular membranes in an energy-independent process (Gadsby 2009).

TRIC channels are expressed in mammalian cell types, where TRIC-A is found primarily in excitable tissues, particularly in the brain and striated (skeletal and cardiac) muscle, and TRIC-B is globally expressed throughout most mammalian tissues (Yamazaki et al. 2009a). Prior research revealed that the cation-selective TRIC-A channels are scattered throughout the sarcoplasmic reticulum (SR) of muscle cells but absent in cell-surface membranes. Skeletal muscle TRIC-A-negative cells exhibit SR instability and Ca^{2+} overload (Zhao et al. 2010). Similarly, TRIC-B channels were shown to localize to the surface of the endoplasmic reticulum (ER) (Yazawa et al. 2007). Moreover, TRIC-A is regulated strongly by transmembrane voltage, whereas

TRIC-B is activated by different mechanisms, thereby providing maximal flexibility and scope for facilitating monovalent cation flux across the SR membrane (Pitt et al. 2010). The evidence strongly suggests that TRIC channels localize to membranes that house the intracellular Ca^{2+} stores and facilitate Ca^{2+} ion transport across internal membranes of mammals (Yazawa et al. 2007; Zhao et al. 2010).

Controversial findings regarding TRIC channel hydropathy properties have led to the proposal that a single TRIC monomer contains three transmembrane segments (TMSs) (Yazawa et al. 2007). The amino terminus of the TRIC subunit is oriented toward the SR/ER lumen, whereas the carboxy terminus is oriented toward the cytoplasm (Yazawa et al. 2007). It was proposed that a hydrophobic loop between the first and second TMSs contributes to the ion-conducting pore. TRIC subunits associate to form homotrimers with a triangular pyramidal structure (Yazawa et al. 2007) and, therefore, have a quaternary structural resemblance to P2X channels (TC 1.A.7) (Mio et al. 2005) and some bacterial porin channels (TC 1.B.1) (Cowan et al. 1992). TRIC channels have an affinity for potassium over sodium cations, with the permeability ratio being 1.5 (Yazawa et al. 2007).

Both TRIC-A and TRIC-B channels are instrumental in E–C coupling in striated muscle (Pitt et al. 2010). In the SR, the propagation of an action potential triggers the opening of the L-type Ca^{2+} channels on the T-tubule surface. External Ca^{2+} enters the skeletal/cardiac muscle cell cytoplasm until SR membrane-bound ryanodine receptors (RyR, TC 1.A.3) detect the increase in concentration (Meissner 1994). A Ca^{2+}-induced Ca^{2+} release (CICR) event stimulates release from the SR into the cytoplasm, promoting muscle contraction (Weisleder et al. 2008). In the ER, the action potential triggers voltage-induced Ca^{2+} release (Rios et al. 1991, 1992; Schneider 1994). Evidence suggests that membrane-bound K^+ channels specifically counter the Ca^{2+} movement to neutralize the transient negative potential in the SR generated by calcium efflux (Fink and Stephenson 1987; Yamazaki et al. 2009b). Although not yet certain, it is likely that TRIC channels (both A and B subtypes) contribute to the neutralization of this negative potential (Weisleder et al. 2008; Zhao et al. 2010).

Preliminary experiments have identified at least two kinds of K^+ channels, both of which act to neutralize the transient luminal negative charge caused by Ca^{2+} release. With one being the TRIC channel(s) and the other a generic SR K^+ channel, differences between the two are exhibited by the unhindered activity of the former in the presence of high decamethonium concentrations (Coronado and Miller 1980; Weisleder et al. 2008). E–C coupling allows rapid calcium release from intracellular stores into the cytoplasm of cardiomyocytes and skeletal muscle cells in order to facilitate muscular contraction and movement.

It is well known that eukaryotic cell signaling is dependent on the efficient translocation of Ca^{2+} ions from intracellular sources within the lumen of the SR into the cytoplasm (Weisleder et al. 2008). Although current evidence implicates the TRIC-A protein in E–C coupling, primarily for counter-ion movement during Ca^{2+} release into the cytoplasm, the detailed process by which TRIC channels neutralize the transient negative charge has yet to be fully characterized (Pitt et al. 2010).

Single knockout mutations, disrupting either of the TRIC subtypes, revealed contrasting effects on mammalian physiological responses. TRIC-A knockout experiments demonstrated that the genetically altered mice are otherwise healthy and maintain the ability to propagate. Alternatively, TRIC-B knockout mice suffer neonatal lethality (Yazawa et al. 2007). Normally abundant throughout most mammalian tissues, such as alveolar epithelial cells, the lack of TRIC-B expression results in respiratory dysfunction (Yamazaki et al. 2009a). Although TRIC-A and TRIC-B are expressed at similar levels in the adult lung, TRIC-B is the more populous subtype in neonatal lungs (Yamazaki et al. 2009a).

The study of TRIC-A and TRIC-B double knockout mice revealed weaker cardiac activity and a decline in cardiomyocyte development compared with wild-type mice (Pitt et al. 2010). The strength of spontaneous cytoplasmic Ca^{2+} oscillations in these double knockouts is lower compared to wild-type progeny, indicating a compromised CICR response, due to limited RyR activity (Takeshima et al. 1998). Evident from excess accumulation of intracellular Ca^{2+} stores, TRIC-A/TRIC-B double knockout mutants feature swollen SR/ER organelles (Yazawa et al. 2007).

This study focuses on the bioinformatic characterization of the TRIC family (Saier 2003a; Yen et al. 2009). Using TRIC subtype sequences from the TCDB Web site (www.tcdb.org), phylogenetic relationships were defined. Specifically, we established an evolutionary connection between the *Mus musculus* TRIC-A protein (GenBank index (gi) 121957073) and an archaeal protein, the *Sulfolobus solfataricus* P2 hypothetical protein SSO0012 (gi 15896983). Further statistical analyses allowed us to establish that TRIC homologs occur throughout the three domains of life (see Matias et al. 2010 and Wang et al. 2009 for methodology). The relationships of 342 TRIC homologs are demonstrated. Furthermore, our topological analyses of TRIC channels lead us to suggest that, in contrast to a previous suggestion that TRIC monomers possess three TMSs (Yazawa et al. 2007), an intragenic duplication event played an essential role in their evolution where a genetic element encoding a primordial three-TMS precursor duplicated to give rise to a six-TMS sequence followed by the addition of one more TMS at the C

terminus. This last event may have occurred by a gene fusion event, giving rise to the proposed seven-TMS topology, common to all recognized members of this family (Saier 2003b).

Methods

Using the sequences (1) *M. musculus* TRIC-A (gi 121957073), (2) *Chlamydomonas reinhardtii* TRIC homolog (gi 159466938) and (3) *S. solfataricus* TRIC homolog (gi 15896983), PSI-BLAST searches (Altschul et al. 1997) were utilized to screen the nonredundant protein database in the National Center for Biotechnology Information (NCBI). Two iterations with a cutoff of e^{-4} were run. Corresponding search results were converted into TinySeqXML format, and all sequences were combined into a single file for analysis.

The sequences were screened with a 90% identity threshold using the MakeTable5 script (Yen et al. 2009). This program eliminated all but one sequence from each set of redundant and similar proteins that shared >90% identity. Output files were as follows: the FASTA file with the protein sequences that share <90% identity with each other, a FASTA file containing the corresponding 16S and 18S ribosomal RNA sequences of the represented genera and a table containing protein abbreviations, descriptions, taxonomic origins, gi numbers, sizes, organisms, organismal phyla and organismal domains. The sequence descriptions were omitted, and fragments were eliminated. The results are presented in Table 1.

The CLUSTAL X program (Thompson et al. 1997) was used to produce multiple alignments of (1) the prokaryotic protein set, (2) the eukaryotic protein set and (3) all proteins combined. From multiple alignment (3), a neighbor-joining phylogenetic tree was produced and depicted with the TreeView and FigTree applications, which gave comparable results (Zhai et al. 2002; http://tree.bio.ed.ac.uk/software/figtree/). Individual protein sequences that were subjected to topological analyses were examined using the WHAT program (Zhai and Saier 2001a) as well as the TMHMM 2.0 (Krogh et al. 2001) and the HMMTOP programs (Tusnády and Simon 1998, 2001). Predictions of average hydropathy, amphipathicity and similarity among multiply aligned protein sequences were made using the default settings of the AveHAS program (Zhai and Saier 2001b).

To validate prokaryotic–eukaryotic TRIC protein homology, internal similarities and repeats, several statistical methods were used. Among them, the IC(Faa2) program was used to compare sequences (Yen et al. 2009). The GAP program (Devereux et al. 1984) was then used to confirm the highest matching pairs of sequences as identified by the IC(Faa2) program and to display the alignments. According to our statistical criteria, homology was established when the GAP comparison score was 10 standard deviations (SDs) or higher, which corresponds to a probability of 10^{-24} or less that the degree of sequence similarity between the two proteins occurred by chance (Saier 1994; Saier et al. 2009; Yen et al. 2009). Scores were optimized by removing unpaired regions and minimizing gaps with retention of at least 60 contiguous residues.

To further substantiate homology and internal repeats, additional programs were used. The GGSEARCH program of the FASTA package from the University of Virginia (http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml) was used to align the prokaryotic and eukaryotic sequences in order to determine significant similarity between the two sets. Using a threshold of e^{-3}, which gives evidence for homology, this program supported the conclusion of homology between prokaryotic and eukaryotic sequences and between the first three TMSs compared to the second three TMSs.

HMMER 2.0 (http://hmmer.janelia.org) was also used to substantiate homology (Eddy 1998, 2008). The purposes of this study required three applications within the HMMER 2.0 program to facilitate similarity analysis. Both sets of sequences (prokaryotic and eukaryotic homologs) and both sets of halves (internal repeats) were used to generate a profile: the hmmbuild component transformed the input set of sequences into a profile, which was used as a consensus sequence for comparison against an input database comprised of the second set of sequences; the hmmcalibrate component refined the profile to achieve more accurate results; the hmmsearch mode was involved in aligning the profile sequence with the database set, where halves were compared against each other and prokaryotic and eukaryotic homolog sets were compared against each other. The commands for HMMER are as follows:

```
hmmbuild <hmm file> <alignment file>
hmmcalibrate <hmm file>
hmmsearch <hmm file> <sequence file>
```

The output file featured the alignment results for sequences best matching the profile.

MEME analyses were performed to provide further support for prokaryotic vs. eukaryotic TRIC homology relationships (Bailey and Elkan 1995). Default settings for MEME were maintained except for two parameters: the range of motif length (in amino acyl residues) and the maximal number of distinctive motifs. The former was set at six residues minimum and 25 residues maximum, and the latter was set at six motifs. The eukaryotic TRIC protein set comprised the first set of results. Limitations of input size of the MEME program necessitated splitting the prokaryotic protein set, giving two sets of results for the prokaryotic group. In order to compare the eukaryotic and

Table 1 The 342 TRIC family proteins included in this study

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Cluster 1					
Dre1	*Danio rerio*	41053814	Metazoa	Eukaryota	295
Ssa1	*Salmo salar*	213512422	Metazoa	Eukaryota	289
Tni1	*Tetraodon nigroviridis*	47210146	Metazoa	Eukaryota	358
Gga1	*Gallus gallus*	119331148	Metazoa	Eukaryota	296
Tgu1	*Taeniopygia guttata*	224087569	Metazoa	Eukaryota	284
Ocu1	*Oryctolagus cuniculus*	153792068	Metazoa	Eukaryota	295
Ptr1	*Pan troglodytes*	114675952	Metazoa	Eukaryota	299
Mmu3	*Macaca mulatta*	109123817	Metazoa	Eukaryota	295
Mdo1	*Monodelphis domestica*	126324081	Metazoa	Eukaryota	303
Oan2	*Ornithorhynchus anatinus*	149639123	Metazoa	Eukaryota	348
Xla1	*Xenopus laevis*	148237167	Metazoa	Eukaryota	295
Xla2	*Xenopus laevis*	189083794	Metazoa	Eukaryota	284
Orf5	*Xenopus* (*Silurana*) *tropicalis*	62859331	Metazoa	Eukaryota	284
Gga2	*Gallus gallus*	50761922	Metazoa	Eukaryota	287
Tgu2	*Taeniopygia guttata*	224091517	Metazoa	Eukaryota	286
Tni2	*Tetraodon nigroviridis*	47218673	Metazoa	Eukaryota	345
Dre4	*Danio rerio*	41055766	Metazoa	Eukaryota	289
Oan3	*Ornithorhynchus anatinus*	149638747	Metazoa	Eukaryota	356
Eca1	*Equus caballus*	149739836	Metazoa	Eukaryota	367
Bta1	*Bos taurus*	115497006	Metazoa	Eukaryota	291
Mdo2	*Monodelphis domestica*	126335835	Metazoa	Eukaryota	302
Bfl1	*Branchiostoma floridae*	219436268	Metazoa	Eukaryota	266
Bfl2	*Branchiostoma floridae*	219436396	Metazoa	Eukaryota	283
Cbr1	*Caenorhabditis briggsae* AF16	157752930	Metazoa	Eukaryota	346
Cel2	*Caenorhabditis elegans*	17537591	Metazoa	Eukaryota	258
Phu1	*Pediculus humanus corporis*	212518051	Metazoa	Eukaryota	277
Ame1	*Apis mellifera*	66554647	Metazoa	Eukaryota	275
Tca1	*Tribolium castaneum*	91086497	Metazoa	Eukaryota	276
Aae1	*Aedes aegypti*	157167601	Metazoa	Eukaryota	275
Dan2	*Drosophila ananassae*	194770140	Metazoa	Eukaryota	276
Api1	*Acyrthosiphon pisum*	193594177	Metazoa	Eukaryota	274
Isc1	*Ixodes scapularis*	215495339	Metazoa	Eukaryota	284
Nve1	*Nematostella vectensis*	156354448	Metazoa	Eukaryota	263
Hma1	*Hydra magnipapillata*	221108707	Metazoa	Eukaryota	465
Dan1	*Drosophila ananassae*	194749507	Metazoa	Eukaryota	293
Der1	*Drosophila erecta*	194873599	Metazoa	Eukaryota	282
Dps1	*Drosophila pseudoobscura pseudoobscura*	198462809	Metazoa	Eukaryota	285
Dvi1	*Drosophila virilis*	195374868	Metazoa	Eukaryota	274
Dmo1	*Drosophila mojavensis*	195135689	Metazoa	Eukaryota	274
Dgr1	*Drosophila grimshawi*	195011963	Metazoa	Eukaryota	383
Cin1	*Ciona intestinalis*	198437585	Metazoa	Eukaryota	281
Cre1	*Chlamydomonas reinhardtii*	159466938	Viridiplantae	Eukaryota	348
Cin2	*Ciona intestinalis*	198427185	Metazoa	Eukaryota	252
Average protein size (amino acids)					301
SD					40

Table 1 continued

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Cluster 2					
Bvi1	*Burkholderia vietnamiensis* G4	134291272	Betaproteobacteria	Bacteria	214
Bsp3	*Burkholderia* sp. H160	209521230	Betaproteobacteria	Bacteria	214
Asp3	*Acidovorax* sp. JS42	121596018	Betaproteobacteria	Bacteria	208
Rpi2	*Ralstonia pickettii* 12 J	187929597	Betaproteobacteria	Bacteria	209
Hso2	*Haemophilus somnus* 129PT	113460328	Gammaproteobacteria	Bacteria	205
Nha1	*Nitrobacter hamburgensis* X14	92119375	Alphaproteobacteria	Bacteria	222
Nmu1	*Nitrosospira multiformis* ATCC 25196	82701983	Betaproteobacteria	Bacteria	231
Xca1	*Xanthomonas campestris* pv. *campestris* str. ATCC 33913	21233198	Gammaproteobacteria	Bacteria	204
Rpa2	*Rhodopseudomonas palustris* CGA009	39936058	Alphaproteobacteria	Bacteria	217
Rba3	*Rhodobacterales bacterium* HTCC2654	84685072	Alphaproteobacteria	Bacteria	205
Mgi1	*Mycobacterium gilvum* PYR-GCK	145221584	Actinobacteria	Bacteria	212
Msp6	*Mycobacterium* sp. MCS	108797746	Actinobacteria	Bacteria	212
Rxy1	*Rubrobacter xylanophilus* DSM 9941	108805439	Actinobacteria	Bacteria	215
Nsp2	*Nocardioides* sp. JS614	119717139	Actinobacteria	Bacteria	211
Bce2	*Burkholderia cenocepacia* J2315	206564245	Betaproteobacteria	Bacteria	240
Pfl2	*Pseudomonas fluorescens* Pf-5	70729939	Gammaproteobacteria	Bacteria	239
Taq1	*Thermus aquaticus* Y51MC23	218297036	Deinococcus-Thermus	Bacteria	198
Tth1	*Thermus thermophilus* HB27	46199739	Deinococcus-Thermus	Bacteria	199
Dge1	*Deinococcus geothermalis* DSM 11300	94984121	Deinococcus-Thermus	Bacteria	216
Rjo1	*Rhodococcus jostii* RHA1	111022587	Actinobacteria	Bacteria	202
Spr1	*Streptomyces pristinaespiralis* ATCC 25486	197775305	Actinobacteria	Bacteria	188
Sgr1	*Streptomyces griseus* subsp. *griseus* NBRC 13350	182435829	Actinobacteria	Bacteria	223
Scl1	*Streptomyces clavuligerus* ATCC 27064	197769495	Actinobacteria	Bacteria	216
Sco1	*Streptomyces coelicolor* A3(2)	21223838	Actinobacteria	Bacteria	219
Sav1	*Streptomyces avermitilis* MA-4680	29829305	Actinobacteria	Bacteria	218
Ssp5	*Streptomyces* sp. Mg1	197754983	Actinobacteria	Bacteria	218
Ssp2	*Streptomyces* sp. SPB74	197762625	Actinobacteria	Bacteria	220
Ser1	*Saccharopolyspora erythraea* NRRL 2338	134098461	Actinobacteria	Bacteria	218
Mab1	*Mycobacterium abscessus*	169629213	Actinobacteria	Bacteria	230
Ach1	*Arthrobacter chlorophenolicus* A6	220913567	Actinobacteria	Bacteria	218
Mlu1	*Micrococcus luteus* NCTC 2665	177671355	Actinobacteria	Bacteria	303
Average protein size (amino acids)					218
SD					19
Cluster 3					
Gur1	*Geobacter uraniireducens* Rf4	148262346	Deltaproteobacteria	Bacteria	206
Ppr1	*Pelobacter propionicus* DSM 2379	118581321	Deltaproteobacteria	Bacteria	206
Glo1	*Geobacter lovleyi* SZ	189426048	Deltaproteobacteria	Bacteria	206
Gme1	*Geobacter metallireducens* GS-15	78221477	Deltaproteobacteria	Bacteria	205
Gsp2	*Geobacter* sp. FRC-32	222054293	Deltaproteobacteria	Bacteria	206
Dac1	*Desulfuromonas acetoxidans* DSM 684	95929748	Deltaproteobacteria	Bacteria	209
Pca1	*Pelobacter carbinolicus* DSM 2380	77918218	Deltaproteobacteria	Bacteria	201
Gbe1	*Geobacter bemidjiensis* Bem	197119189	Deltaproteobacteria	Bacteria	205
Dre3	*Desulfotomaculum reducens* MI-1	134298491	Firmicutes	Bacteria	206

Table 1 continued

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Average protein size (amino acids)					206
SD					2
Cluster 4					
Slo1	*Shewanella loihica* PV-4	127512086	Gammaproteobacteria	Bacteria	217
Swo1	*Shewanella woodyi* ATCC 51908	170725707	Gammaproteobacteria	Bacteria	213
Ssp1	*Shewanella* sp. MR-4	113969547	Gammaproteobacteria	Bacteria	213
Spi1	*Shewanella piezotolerans* WP3	212634142	Gammaproteobacteria	Bacteria	212
Sfr2	*Shewanella frigidimarina* NCIMB 400	114561781	Gammaproteobacteria	Bacteria	212
Spe2	*Shewanella pealeana* ATCC 700345	157960969	Gammaproteobacteria	Bacteria	212
Sam2	*Shewanella amazonensis* SB2B	119775597	Gammaproteobacteria	Bacteria	203
Msp1	*Moritella* sp. PE36	149909407	Gammaproteobacteria	Bacteria	205
Pat3	*Pectobacterium atrosepticum* SCRI1043	50121262	Gammaproteobacteria	Bacteria	208
Asp1	*Acidovorax* sp. JS42	121594165	Betaproteobacteria	Bacteria	221
Cte1	*Comamonas testosteroni* KF-1	221066150	Betaproteobacteria	Bacteria	221
Asp2	*Azoarcus* sp. BH72	119897407	Betaproteobacteria	Bacteria	204
Dar1	*Dechloromonas aromatica* RCB	71908826	Betaproteobacteria	Bacteria	222
Cvi1	*Chromobacterium violaceum* ATCC 12472	34497467	Betaproteobacteria	Bacteria	204
Bja1	*Bradyrhizobium japonicum* USDA 110	27380451	Alphaproteobacteria	Bacteria	213
Bsp2	*Bradyrhizobium* sp. ORS278	146341243	Alphaproteobacteria	Bacteria	214
Bbr1	*Bordetella bronchiseptica* RB50	33603510	Betaproteobacteria	Bacteria	190
Bpe1	*Bordetella petrii* DSM 12804	163854725	Betaproteobacteria	Bacteria	212
Bav1	*Bordetella avium* 197 N	187479653	Betaproteobacteria	Bacteria	190
Xau1	*Xanthobacter autotrophicus* Py2	154246684	Alphaproteobacteria	Bacteria	209
Bam1	*Burkholderia ambifaria* MEX-5	171322332	Betaproteobacteria	Bacteria	203
Bam2	*Burkholderia ambifaria* AMMD	115359524	Betaproteobacteria	Bacteria	203
Bce3	*Burkholderia cenocepacia* J2315	206564469	Betaproteobacteria	Bacteria	203
Bvi2	*Burkholderia vietnamiensis* G4	134292345	Betaproteobacteria	Bacteria	203
Bma4	*Burkholderia mallei* ATCC 23344	53716241	Betaproteobacteria	Bacteria	202
Pst2	*Pseudomonas stutzeri* A1501	146283012	Gammaproteobacteria	Bacteria	187
Vsp1	*Vibrio splendidus* 12B01	84394401	Gammaproteobacteria	Bacteria	207
Vsp5	*Vibrio* sp. Ex25	194540500	Gammaproteobacteria	Bacteria	216
Vvu1	*Vibrio vulnificus* CMCP6	27365327	Gammaproteobacteria	Bacteria	214
Vch1	*Vibrio cholerae* O1 biovar El Tor str. N16961	15642114	Gammaproteobacteria	Bacteria	207
Vsh2	*Vibrio shilonii* AK1	149188134	Gammaproteobacteria	Bacteria	207
Pha1	*Pseudoalteromonas haloplanktis* TAC125	77360305	Gammaproteobacteria	Bacteria	190
Psp1	*Photobacterium* sp. SKA34	89072754	Gammaproteobacteria	Bacteria	206
Ppr2	*Photobacterium profundum* SS9	54309873	Gammaproteobacteria	Bacteria	196
Vfi1	*Vibrio fischeri* ES114	59712411	Gammaproteobacteria	Bacteria	213
Psp4	*Psychromonas* sp. CNPT3	90408648	Gammaproteobacteria	Bacteria	210
Pfl1	*Pseudomonas fluorescens* Pf0-1	77456744	Gammaproteobacteria	Bacteria	203
Ppu1	*Pseudomonas putida* GB-1	167035942	Gammaproteobacteria	Bacteria	203
Psy1	*Pseudomonas syringae* pv. *syringae* B728a	66043832	Gammaproteobacteria	Bacteria	203
Orf8	Synthetic construct	49079778	none	n/a	207
Asp4	*Acinetobacter* sp. ADP1	50083511	Gammaproteobacteria	Bacteria	214
Sma1	*Stenotrophomonas maltophilia* R551-3	194367381	Gammaproteobacteria	Bacteria	205
Asp5	*Alcanivorax* sp. DG881	223479350	Gammaproteobacteria	Bacteria	204
Yps1	*Yersinia pseudotuberculosis* IP 32953	51594397	Gammaproteobacteria	Bacteria	205

Table 1 continued

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Ype1	*Yersinia pestis* KIM	22124019	Gammaproteobacteria	Bacteria	207
Spr2	*Serratia proteamaculans* 568	157373095	Gammaproteobacteria	Bacteria	205
Pst1	*Providencia stuartii* ATCC 25827	183597042	Gammaproteobacteria	Bacteria	204
Pmi1	*Proteus mirabilis* HI4320	197286690	Gammaproteobacteria	Bacteria	205
Plu1	*Photorhabdus luminescens* subsp. *laumondii* TTO1	37524305	Gammaproteobacteria	Bacteria	205
Eta1	*Erwinia tasmaniensis* Et1/99	188532210	Gammaproteobacteria	Bacteria	205
Sen1	*Salmonella enterica* subsp. *arizonae serovar* 62:z4,z23:–str. RSK2980	161505726	Gammaproteobacteria	Bacteria	205
Kpn1	*Klebsiella pneumoniae* subsp. *pneumoniae* MGH 78578	152972500	Gammaproteobacteria	Bacteria	205
Hso1	*Haemophilus somnus* 129PT	113460555	Gammaproteobacteria	Bacteria	223
Pmu1	*Pasteurella multocida* subsp. *multocida* str. Pm70	15602800	Gammaproteobacteria	Bacteria	226
Hin1	*Haemophilus influenzae* Rd KW20	16273159	Gammaproteobacteria	Bacteria	220
Msu1	*Mannheimia succiniciproducens* MBEL55E	52425660	Gammaproteobacteria	Bacteria	227
Asu1	*Actinobacillus succinogenes* 130Z	152979435	Gammaproteobacteria	Bacteria	226
Ahy1	*Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966	117618232	Gammaproteobacteria	Bacteria	204
Cup1	*Campylobacter upsaliensis* RM3195	57505866	Epsilonproteobacteria	Bacteria	205
Cje1	*Campylobacter jejuni* RM1221	57238211	Epsilonproteobacteria	Bacteria	210
Hhe1	*Helicobacter hepaticus* ATCC 51449	32266578	Epsilonproteobacteria	Bacteria	210
Cla1	*Campylobacter lari* RM2100	222824473	Epsilonproteobacteria	Bacteria	208
Cef1	*Corynebacterium efficiens* YS-314	25027246	Actinobacteria	Bacteria	221
Cgl1	*Corynebacterium glutamicum* ATCC 13032	19551902	Actinobacteria	Bacteria	219
Cam1	*Corynebacterium amycolatum* SK46	213966349	Actinobacteria	Bacteria	205
Cje2	*Corynebacterium jeikeium* K411	68536780	Actinobacteria	Bacteria	285
Cur1	*Corynebacterium urealyticum* DSM 7109	172040108	Actinobacteria	Bacteria	293
Lpn1	*Legionella pneumophila* subsp. *pneumophila* str. Philadelphia 1	52842930	Gammaproteobacteria	Bacteria	209
Average protein size (amino acids)					211
SD					16
Cluster 5					
Rsp1	*Reinekea* sp. MED297	88800179	Gammaproteobacteria	Bacteria	214
Orf2	*gamma proteobacterium* HTCC5015	198261805	Gammaproteobacteria	Bacteria	207
Pla1	*Parvibaculum lavamentivorans* DS-1	154253294	Alphaproteobacteria	Bacteria	209
Average protein size (amino acids)					210
SD					4
Cluster 6					
Bun1	*Bacteroides uniformis* ATCC 8492	160889861	Bacteroidetes	Bacteria	208
Bst1	*Bacteroides stercoris* ATCC 43183	167762683	Bacteroidetes	Bacteria	211
Bth1	*Bacteroides thetaiotaomicron* VPI-5482	29346611	Bacteroidetes	Bacteria	209
Bvu1	*Bacteroides vulgatus* ATCC 8482	150005854	Bacteroidetes	Bacteria	204
Bco2	*Bacteroides coprocola* DSM 17136	189461400	Bacteroidetes	Bacteria	208
Bco3	*Bacteroides coprophilus* DSM 18228	224023376	Bacteroidetes	Bacteria	211
Bpl1	*Bacteroides plebeius* DSM 17135	198275310	Bacteroidetes	Bacteria	211
Bfr1	*Bacteroides fragilis* NCTC 9343	60681438	Bacteroidetes	Bacteria	208
Pme1	*Parabacteroides merdae* ATCC 43184	154492668	Bacteroidetes	Bacteria	203
Apu2	*Alistipes putredinis* DSM 17216	167751972	Bacteroidetes	Bacteria	209
Pco1	*Prevotella copri* DSM 18205	223463272	Bacteroidetes	Bacteria	216
Mha1	*Mannheimia haemolytica* PHL213	197748956	Gammaproteobacteria	Bacteria	211
Apl1	*Actinobacillus pleuropneumoniae serovar* 1 str. 4074	53728860	Gammaproteobacteria	Bacteria	214
Hpa1	*Haemophilus parasuis* 29755	167855727	Gammaproteobacteria	Bacteria	210

Table 1 continued

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Csp1	*Capnocytophaga sputigena* Capno	213962585	Bacteroidetes	Bacteria	209
Ddo1	*Dokdonia donghaensis* MED134	86130165	Bacteroidetes	Bacteria	206
Orf3	*unidentified eubacterium* SCB49	149370587	Bacteroidetes	Bacteria	208
Lbl1	*Leeuwenhoekiella blandensis* MED217	86143049	Bacteroidetes	Bacteria	206
Cat1	*Croceibacter atlanticus* HTCC2559	83856981	Bacteroidetes	Bacteria	209
Pto1	*Psychroflexus torquis* ATCC 700755	91214670	Bacteroidetes	Bacteria	209
Fba2	*Flavobacteriales bacterium* ALC-1	163787946	Bacteroidetes	Bacteria	203
Fba3	*Flavobacteria bacterium* BBFL7	89891564	Bacteroidetes	Bacteria	202
Fba1	*Flavobacteriales bacterium* HTCC2170	88712311	Bacteroidetes	Bacteria	204
Rbi1	*Robiginitalea biformata* HTCC2501	88806387	Bacteroidetes	Bacteria	206
Fjo1	*Flavobacterium johnsoniae* UW101	146299991	Bacteroidetes	Bacteria	200
Fps1	*Flavobacterium psychrophilum* JIP02/86	150025911	Bacteroidetes	Bacteria	201
Kal1	*Kordia algicida* OT-1	163756307	Bacteroidetes	Bacteria	201
Gfo1	*Gramella forsetii* KT0803	120435341	Bacteroidetes	Bacteria	215
Psp8	*Pedobacter* sp. BAL39	149279591	Bacteroidetes	Bacteria	203
Asp6	*Algoriphagus* sp. PR1	126645896	Bacteroidetes	Bacteria	200
Lbi1	*Leptospira biflexa* serovar Patoc strain Patoc 1 (Paris)	183221994	Spirochaetes	Bacteria	209
Mma1	*Microscilla marina* ATCC 23134	124007269	Bacteroidetes	Bacteria	205
Pir1	*Polaribacter irgensii* 23-P	88803173	Bacteroidetes	Bacteria	208
Psp5	*Polaribacter* sp. MED152	85820462	Bacteroidetes	Bacteria	208
Average protein size (amino acids)					207
SD					4
Cluster 7					
Psp2	*Psychrobacter* sp. PRwf-1	148652600	Gammaproteobacteria	Bacteria	211
Pcr1	*Psychrobacter cryohalolentis* K5	93005552	Gammaproteobacteria	Bacteria	211
Par2	*Psychrobacter arcticus* 273-4	71065295	Gammaproteobacteria	Bacteria	220
Average protein size (amino acids)					214
SD					5
Cluster 8					
Mal1	*Marinobacter algicola* DG893	149377123	Gammaproteobacteria	Bacteria	206
Maq1	*Marinobacter aquaeolei* VT8	120555723	Gammaproteobacteria	Bacteria	205
Rso1	*Ralstonia solanacearum* GMI1000	17547063	Betaproteobacteria	Bacteria	207
Rpi1	*Ralstonia pickettii* 12J	187929632	Betaproteobacteria	Bacteria	207
Rme1	*Ralstonia metallidurans* CH34	94309627	Betaproteobacteria	Bacteria	215
Cta1	*Cupriavidus taiwanensis*	194288872	Betaproteobacteria	Bacteria	213
Bce1	*Burkholderia cenocepacia* PC184	194557052	Betaproteobacteria	Bacteria	208
Bce4	*Burkholderia cenocepacia* J2315	206561382	Betaproteobacteria	Bacteria	207
Bma5	*Burkholderia mallei* ATCC 23344	53725968	Betaproteobacteria	Bacteria	206
Bph1	*Burkholderia phymatum* STM815	186475398	Betaproteobacteria	Bacteria	206
Bxe1	*Burkholderia xenovorans* LB400	91782386	Betaproteobacteria	Bacteria	207
Har1	*Herminiimonas arsenicoxydans*	134093950	Betaproteobacteria	Bacteria	202
Jsp1	*Janthinobacterium* sp. Marseille	152981420	Betaproteobacteria	Bacteria	210
Hau1	*Herpetosiphon aurantiacus* ATCC 23779	159899999	Chloroflexi	Bacteria	202
Pna1	*Polaromonas naphthalenivorans* CJ2	121603779	Betaproteobacteria	Bacteria	213
Lsp1	*Limnobacter* sp. MED105	149925313	Betaproteobacteria	Bacteria	214
Average protein size (amino acids)					208
SD					4

Table 1 continued

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Cluster 9					
Ama1	*Alteromonas macleodii* Deep ecotype	196156518	Gammaproteobacteria	Bacteria	216
Pat1	*Pseudoalteromonas atlantica* T6c	109898588	Gammaproteobacteria	Bacteria	208
Average protein size (amino acids)					212
SD					6
Cluster 10					
Bma1	*Bermanella marisrubri*	94500795	Gammaproteobacteria	Bacteria	205
Oan1	*Ochrobactrum anthropi* ATCC 49188	153007524	Alphaproteobacteria	Bacteria	218
Bsu2	*Brucella suis* 1330	23501081	Alphaproteobacteria	Bacteria	218
Avi1	*Agrobacterium vitis* S4	222147159	Alphaproteobacteria	Bacteria	211
Ara1	*Agrobacterium radiobacter* K84	222084313	Alphaproteobacteria	Bacteria	274
Ret1	*Rhizobium etli* IE4771	218661820	Alphaproteobacteria	Bacteria	225
Rle1	*Rhizobium leguminosarum* bv. viciae 3841	116249880	Alphaproteobacteria	Bacteria	211
Atu1	*Agrobacterium tumefaciens* str. C58	159184188	Alphaproteobacteria	Bacteria	212
Sme1	*Sinorhizobium meliloti* 1021	15964003	Alphaproteobacteria	Bacteria	211
Hph1	*Hoeflea phototrophica* DFL-43	163757820	Alphaproteobacteria	Bacteria	211
Msp2	*Mesorhizobium* sp. BNC1	110636344	Alphaproteobacteria	Bacteria	210
Mlo1	*Mesorhizobium loti* MAFF303099	13472807	Alphaproteobacteria	Bacteria	212
Hne1	*Hyphomonas neptunium* ATCC 15444	114798893	Alphaproteobacteria	Bacteria	210
Lal1	*Labrenzia alexandrii* DFL-11	224397513	Alphaproteobacteria	Bacteria	203
Lag1	*Labrenzia aggregata* IAM 12614	118593734	Alphaproteobacteria	Bacteria	206
Esp1	*Erythrobacter* sp. SD-21	149184343	Alphaproteobacteria	Bacteria	211
Esp2	*Erythrobacter* sp. NAP1	85709024	Alphaproteobacteria	Bacteria	231
Nar1	*Novosphingobium aromaticivorans* DSM 12444	87199197	Alphaproteobacteria	Bacteria	212
Swi1	*Sphingomonas wittichii* RW1	148553449	Alphaproteobacteria	Bacteria	208
Ccr1	*Caulobacter crescentus* CB15	16127910	Alphaproteobacteria	Bacteria	229
Csp2	*Caulobacter* sp. K31	167648816	Alphaproteobacteria	Bacteria	210
Rsp4	*Rhodobacter sphaeroides* ATCC 17025	146277879	Alphaproteobacteria	Bacteria	213
Rpa1	*Rhodopseudomonas palustris* HaA2	86748428	Alphaproteobacteria	Bacteria	209
Rru1	*Rhodospirillum rubrum* ATCC 11170	83592327	Alphaproteobacteria	Bacteria	238
Orf6	*alpha proteobacterium* BAL199	163794407	Alphaproteobacteria	Bacteria	210
Rba2	*Rhodobacterales bacterium* Y4I	206686839	Alphaproteobacteria	Bacteria	207
Pga1	*Phaeobacter gallaeciensis* BS107	163738224	Alphaproteobacteria	Bacteria	207
Ssp4	*Silicibacter* sp. TM1040	99082414	Alphaproteobacteria	Bacteria	207
Rsp2	*Roseobacter* sp. MED193	86137260	Alphaproteobacteria	Bacteria	207
Spo1	*Silicibacter pomeroyi* DSS-3	56698045	Alphaproteobacteria	Bacteria	209
Rsp3	*Roseovarius* sp. 217	85705028	Alphaproteobacteria	Bacteria	207
Oin1	*Oceanibulbus indolifex* HEL-45	163744868	Alphaproteobacteria	Bacteria	207
Sst1	*Sagittula stellata* E-37	126728521	Alphaproteobacteria	Bacteria	211
Pde1	*Paracoccus denitrificans* PD1222	69937880	Alphaproteobacteria	Bacteria	210
Orf1	*gamma proteobacterium* HTCC2207	90417417	Gammaproteobacteria	Bacteria	205
Oal1	*Oceanicaulis alexandrii* HTCC2633	83858213	Alphaproteobacteria	Bacteria	210
Ota1	*Ostreococcus tauri*	116057349	Viridiplantae	Eukaryota	311
Average protein size (amino acids)					216
SD					20

Table 1 continued

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Cluster 11					
Ilo1	*Idiomarina loihiensis* L2TR	56461251	Gammaproteobacteria	Bacteria	204
Iba1	*Idiomarina baltica* OS145	85711205	Gammaproteobacteria	Bacteria	207
Pha2	*Pseudoalteromonas haloplanktis* TAC125	77359101	Gammaproteobacteria	Bacteria	204
Aba1	*Alteromonadales bacterium* TW-7	119468295	Gammaproteobacteria	Bacteria	204
Ptu1	*Pseudoalteromonas tunicata* D2	88861038	Gammaproteobacteria	Bacteria	204
Pat2	*Pseudoalteromonas atlantica* T6c	109900549	Gammaproteobacteria	Bacteria	206
Vba1	*Verrucomicrobiae bacterium* DG1235	198258994	Verrucomicrobia	Bacteria	204
Cps1	*Colwellia psychrerythraea* 34H	71279728	Gammaproteobacteria	Bacteria	219
Pin1	*Psychromonas ingrahamii* 37	119944631	Gammaproteobacteria	Bacteria	208
Psp7	*Psychromonas* sp. CNPT3	90407400	Gammaproteobacteria	Bacteria	205
Csa1	*Chromohalobacter salexigens* DSM 3043	92112483	Gammaproteobacteria	Bacteria	246
Spe1	*Shewanella pealeana* ATCC 700345	157960826	Gammaproteobacteria	Bacteria	207
Spi2	*Shewanella piezotolerans* WP3	212636573	Gammaproteobacteria	Bacteria	208
Sbe1	*Shewanella benthica* KT99	163751951	Gammaproteobacteria	Bacteria	206
Swo2	*Shewanella woodyi* ATCC 51908	170725576	Gammaproteobacteria	Bacteria	211
Slo2	*Shewanella loihica* PV-4	127511957	Gammaproteobacteria	Bacteria	206
Ssp3	*Shewanella* sp. MR-4	113971212	Gammaproteobacteria	Bacteria	208
Sde1	*Shewanella denitrificans* OS217	91794155	Gammaproteobacteria	Bacteria	208
Sfr1	*Shewanella frigidimarina* NCIMB 400	114564133	Gammaproteobacteria	Bacteria	208
Sam1	*Shewanella amazonensis* SB2B	119773989	Gammaproteobacteria	Bacteria	207
Apu1	*Aeromonas punctata*	11135907	Gammaproteobacteria	Bacteria	210
Msp5	*Moritella* sp. PE36	149909290	Gammaproteobacteria	Bacteria	207
Vsp2	*Vibrio* sp. MED222	86147371	Gammaproteobacteria	Bacteria	205
Vpa1	*Vibrio parahaemolyticus* 16	219549565	Gammaproteobacteria	Bacteria	206
Vsp3	*Vibrio* sp. Ex25	194539256	Gammaproteobacteria	Bacteria	206
Vvu2	*Vibrio vulnificus* YJ016	37678816	Gammaproteobacteria	Bacteria	221
Vch2	*Vibrio cholerae* O1 biovar El Tor str. N16961	15642379	Gammaproteobacteria	Bacteria	239
Vfi2	*Vibrio fischeri* ES114	59712737	Gammaproteobacteria	Bacteria	203
Vsh1	*Vibrio shilonii* AK1	149190320	Gammaproteobacteria	Bacteria	177
Psp3	*Photobacterium* sp. SKA34	89072505	Gammaproteobacteria	Bacteria	203
Ppr3	*Photobacterium profundum* SS9	54307740	Gammaproteobacteria	Bacteria	203
Eco1	*Escherichia coli* O157:H7 EDL933	15799841	Gammaproteobacteria	Bacteria	207
Sen2	*Salmonella enterica* subsp. *arizonae serovar* 62:z4,z23:–	161504677	Gammaproteobacteria	Bacteria	205
Ype2	*Yersinia pestis* KIM	22124716	Gammaproteobacteria	Bacteria	220
Efe1	*Escherichia fergusonii* ATCC 35469	218547613	Gammaproteobacteria	Bacteria	209
Sgl1	*Sodalis glossinidius* str. morsitans	85058480	Gammaproteobacteria	Bacteria	205
Tau1	*Tolumonas auensis* DSM 9187	223580357	Gammaproteobacteria	Bacteria	212
Hch1	*Hahella chejuensis* KCTC 2396	83648724	Gammaproteobacteria	Bacteria	210
Aeh1	*Alkalilimnicola ehrlichei* MLHE-1	114320536	Gammaproteobacteria	Bacteria	204
Pme2	*Pseudomonas mendocina* ymp	146307375	Gammaproteobacteria	Bacteria	207
Sru1	*Salinibacter ruber* DSM 13855	83816124	Bacteroidetes	Bacteria	216
Tde1	*Thiobacillus denitrificans* ATCC 25259	74317133	Betaproteobacteria	Bacteria	204
Vsp4	*Verrucomicrobium spinosum* DSM 4136	171910474	Verrucomicrobia	Bacteria	203
Nca1	*Neptuniibacter caesariensis*	89094836	Gammaproteobacteria	Bacteria	222
Rba1	*Rhodobacterales bacterium* HTCC2255	114769786	Alphaproteobacteria	Bacteria	206
Amu1	*Akkermansia muciniphila* ATCC BAA-835	187735153	Verrucomicrobia	Bacteria	217

Table 1 continued

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Pne1	*Polynucleobacter necessarius* subsp. *necessarius* STIR1	171464135	Betaproteobacteria	Bacteria	209
Pne2	*Polynucleobacter necessarius* subsp. *asymbioticus* QLW-P1DMWA-1	145590030	Betaproteobacteria	Bacteria	209
Pma1	*Planctomyces maris* DSM 8797	149178809	Planctomycetes	Bacteria	224
Msp3	*Marinomonas* sp. MWYL1	152995441	Gammaproteobacteria	Bacteria	204
Msp4	*Marinomonas* sp. MED121	87120574	Gammaproteobacteria	Bacteria	199
Dsa1	*Desulfovibrio salexigens* DSM 2638	218148950	Deltaproteobacteria	Bacteria	207
Average protein size (amino acids)					209
SD					10
Cluster 12					
Gsp1	*Geobacillus* sp. WCH70	171324432	Firmicutes	Bacteria	206
Afl1	*Anoxybacillus flavithermus* WK1	212639342	Firmicutes	Bacteria	202
Gka1	*Geobacillus kaustophilus* HTA426	56419858	Firmicutes	Bacteria	205
Bsp1	*Bacillus* sp. SG-1	149183191	Firmicutes	Bacteria	202
Bsp5	*Bacillus* sp. NRRL B-14911	89100984	Firmicutes	Bacteria	203
Bco1	*Bacillus coahuilensis* m4-4	205373537	Firmicutes	Bacteria	207
Bsu1	*Bacillus subtilis* subsp. *subtilis* str. 168	16080399	Firmicutes	Bacteria	202
Bli1	*Bacillus licheniformis* ATCC 14580	52081835	Firmicutes	Bacteria	202
Ban2	*Bacillus anthracis* str. Ames	30263712	Firmicutes	Bacteria	207
Bsp4	*Bacillus* sp. B14905	126650178	Firmicutes	Bacteria	223
Gsp3	*Geobacillus* sp. Y412MC10	192808813	Firmicutes	Bacteria	217
Psp6	*Paenibacillus* sp. JDR-2	169191088	Firmicutes	Bacteria	207
Esi1	*Exiguobacterium sibiricum* 255-15	172056257	Firmicutes	Bacteria	203
Sin1	*Streptococcus infantarius* subsp. *infantarius* ATCC BAA-102	171778106	Firmicutes	Bacteria	208
Spy1	*Streptococcus pyogenes* M1 GAS	15675558	Firmicutes	Bacteria	201
Aac1	*Alicyclobacillus acidocaldarius* LAA1	218289417	Firmicutes	Bacteria	204
Csp3	*Carnobacterium* sp. AT7	163791373	Firmicutes	Bacteria	207
Sso1	*Sulfolobus solfataricus* P2	15896983	Crenarchaeota	Archaea	205
Ssu1	*Streptococcus suis* 05ZYH33	146319496	Firmicutes	Bacteria	246
Cce1	*Clostridium cellulolyticum* H10	220928457	Firmicutes	Bacteria	203
Cma1	*Caldivirga maquilingensis* IC-167	159042484	Crenarchaeota	Archaea	209
Sto1	*Sulfolobus tokodaii* str. 7	15921908	Crenarchaeota	Archaea	203
Sac1	*Sulfolobus acidocaldarius* DSM 639	70607007	Crenarchaeota	Archaea	207
Mse1	*Metallosphaera sedula* DSM 5348	146302965	Crenarchaeota	Archaea	204
Tne1	*Thermoproteus neutrophilus* V24Sta	171185944	Crenarchaeota	Archaea	213
Pis1	*Pyrobaculum islandicum* DSM 4184	119872142	Crenarchaeota	Archaea	208
Pae1	*Pyrobaculum aerophilum* str. IM2	18313279	Crenarchaeota	Archaea	211
Par1	*Pyrobaculum arsenaticum* DSM 13514	145591759	Crenarchaeota	Archaea	209
Pca2	*Pyrobaculum calidifontis* JCM 11548	126460288	Crenarchaeota	Archaea	220
Mma2	*Methanococcus maripaludis* C7	150403287	Euryarchaeota	Archaea	223
Nsp1	*Nitratiruptor* sp. SB155-2	152990839	Epsilonproteobacteria	Bacteria	201
Average protein size (amino acids)					209
SD					9
Cluster 13					
Sac2	*Syntrophus aciditrophicus* SB	85859719	Deltaproteobacteria	Bacteria	255
Dvu1	*Desulfovibrio vulgaris* str. Hildenborough	46581076	Deltaproteobacteria	Bacteria	207

Table 1 continued

Abbreviation	Organism	GenBank index	Group	Domain	Protein size
Mfe1	*Mariprofundus ferrooxydans* PV-1	114777007	Proteobacteria	Bacteria	205
Nma1	*Nitrosopumilus maritimus* SCM1	161527883	Crenarchaeota	Archaea	210
Orf4	Uncultured marine crenarchaeote HF4000_APKG3E18	167043871	Crenarchaeota	Archaea	203
Orf7	Uncultured crenarchaeote	42557779	Crenarchaeota	Archaea	210
Average protein size (amino acids)					215
SD					20
Cluster 14					
Cbo1	*Clostridium bolteae* ATCC BAA-613	160938401	Firmicutes	Bacteria	219
Cph1	*Clostridium phytofermentans* ISDg	160879149	Firmicutes	Bacteria	213
Rob1	*Ruminococcus obeum* ATCC 29174	153813403	Firmicutes	Bacteria	217
Dlo1	*Dorea longicatena* DSM 13814	153855137	Firmicutes	Bacteria	235
Mmu5	*Mitsuokella multacida* DSM 20544	218252448	Firmicutes	Bacteria	221
Average protein size (amino acids)					221
SD					8
Cluster 15					
Bca1	*Bifidobacterium catenulatum* DSM 16992	212716008	Actinobacteria	Bacteria	262
Blo1	*Bifidobacterium longum* DJO10A	23335025	Actinobacteria	Bacteria	263
Average protein size (amino acids)					263
SD					1

Proteins are organized by cluster in order of their positions in the phylogenetic tree (Figs. 1, 2). Taxonomic origins, gi numbers, sizes, organismal phyla and organismal domains are provided. The average sizes of the proteins of a cluster ± SD are featured below each cluster

prokaryotic consensus sequences, the matching regions of the two sets of prokaryotic sequences were combined.

SEED analyses were conducted to predict possible functions of prokaryotic homologs. Genome context analyses were performed using the SEED comparative genomics database (Overbeek et al. 2005), which can be found at http://seed-viewer.theseed.org/. Selected proteins from all prokaryotic clusters were used as query sequences to identify the 20 closest homologs in the SEED database to determine the genome context of regions encoding TRIC family homologs. Predicted functions of the prokaryotic sequences were based on the association and coregulation of known proteins in the corresponding operons and surrounding regions. The analyses reported represent those which allowed reasonable prediction of function for the TRIC homologs. Supplementary material can be found at the following web address: http://www.biology.ucsd.edu/~msaier/supmat/TRIC/index.html.

Results

Prokaryotic and Eukaryotic TRIC Homologs

In order to retrieve homologs of the mouse TRIC-A protein (gi 121957073), this protein was used as the query sequence in an NCBI PSI-BLAST search with two iterations with a cutoff of e^{-4}. Two of the proteins obtained were homologs from the alga *C. reinhardtii* (gi 159466938) and the archaeon *S. solfataricus* (gi 15896983). These proteins as well as the mouse TRIC-A protein were used as query sequences in further BLAST searches, as described in Methods.

Proteins obtained were then multiply aligned using the CLUSTAL X program, and the sequences were visually assessed for completeness. Fragmentary sequences were removed, leaving 342 proteins, which were included in our primary studies and multiply aligned (Fig. S1-1, http://www.biology.ucsd.edu/~msaier/supmat/TRIC/index.html). These proteins are arranged in alphabetical order of their abbreviations in Table S1 and a corresponding table, where the proteins are listed according to phylogenetic cluster (Fig. 1), with proteins within each cluster arranged according to position within that cluster (Table 1). Neighbor-joining phylogenetic trees, generated using TreeView or FigTree (Zhai et al. 2002; http://tree.bio.ed.ac.uk/software/figtree/) (Fig. 1) revealed 15 phylogenetic clusters. The corresponding dendogram is shown in Fig. S2.

Multiple Alignments of TRIC Family Homologs

The multiple alignment of all 342 TRIC family homologs is shown in Fig. S1-1, while the multiple alignment for the

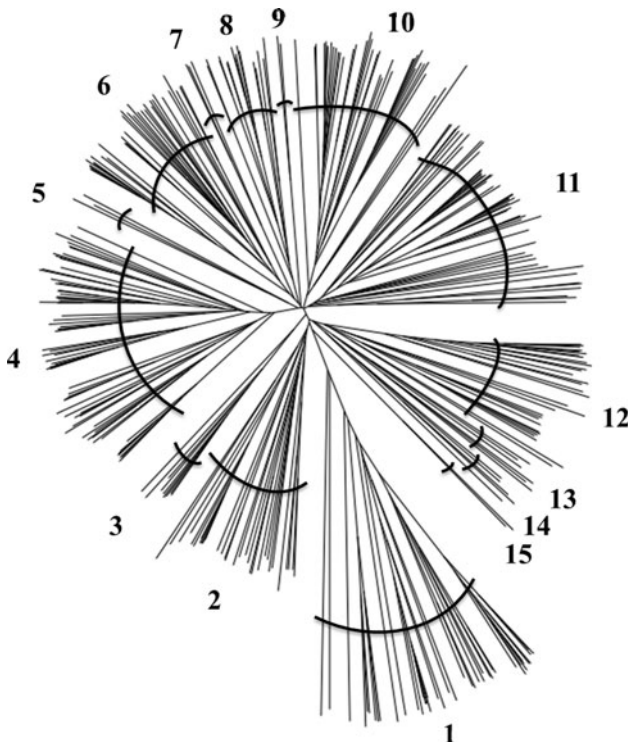

Fig. 1 A radial phylogenetic tree of the 342 TRIC family proteins included in this study depicted with the FigTree program and based on the CLUSTAL X multiple alignment shown in Fig. S1. Protein labels were removed but are presented in Fig. S2. *Bold numbers* refer to the 15 clusters (see Table 1). Cluster 1, shown at the bottom of the tree, features the eukaryotic TRIC protein sequences. Clusters 2–15 feature the prokaryotic TRIC family homologs, with the lone eukaryotic sequence, Ota1, in Cluster 10. Protein abbreviations and characteristics are included in Tables 1 and S1

prokaryotic proteins is shown in Fig. S1-2 and the multiple alignment for the eukaryotic proteins is shown in Fig. S1-3. Figure S1-2 features 299 prokaryotic homologs, one of which may be a mitochondrial protein from the plant *Ostreococcus tauri* since it clusters with α-proteobacterial homologs (Lang et al. 1999). Since the protein is distantly related to the other members of this cluster, it may have been obtained by vertical descent from an α-proteobacterium, the mitochondrial precursor, rather than by horizontal transfer (Lerat et al. 2005; Woese 2000).

The multiple alignment of the 299 prokaryotic sequences reveals four fully conserved residues: G, D, G and Y. The first G is near the C terminus of TMS 2, where three glycine residues are adjacent to each other in most of the homologs. The D appears at the N terminus of TMS 4. The second conserved G occurs in TMS 5, where again we find either two or three adjacent glycines, as observed in TMS 2 (see above). Finally, the fourth fully conserved residue is a Y in the N terminus of TMS 6.

Figure S1-3 features 43 eukaryotic homologs and reveals three fully conserved residues, W, P, and G, where the W and P occur in the most conserved parts of the proteins, in TMS 3, while the fully conserved G appears in TMS 5. Examination of TMSs 2 and 5 reveals that in both regions there exist two or three well conserved consecutive Gs, as observed for the prokaryotic proteins. The first two of the three fully conserved residues are part of a well-conserved

```
                    * :       : *   .
```

motif, which is WY(LIV)$_2$FYCPX(DN), where residues in parentheses represent alternative possibilities at a single position and X indicates any residue. While the W and P are fully conserved (*), the two Ys can be replaced only by related aromatic residues (:) and the (DN) at motif position 10 includes more distantly related similarities (.). The first Y is substituted by W only in the three most divergent proteins, while the second Y is only substituted by F. Interestingly, none of the fully conserved residues in the prokaryotic homologs is fully conserved in the eukaryotic homologs.

Phylogenetic, Organismal and Size Analyses of TRIC Family Homologs

The phylogenetic tree for all TRIC family members analyzed in this study is shown in Fig. 1, while the 342 proteins are tabulated according to cluster in Table 1. These proteins exhibit a surprising degree of size homogeneity, with only two clusters showing appreciable variation. Cluster 1 includes all eukaryotic proteins that average 301 residues and are roughly 90 residues (45%) larger than the prokaryotic proteins with the sole exception of Cluster 15, which has an average size of 263 residues. The larger sizes and size variation of the eukaryotic proteins are due primarily to hydrophilic extensions at both the N and C termini. The larger prokaryotic Cluster 15 proteins, all from Actinobacteria, reflect the presence of strongly hydrophilic C-terminal extensions. In addition to these size variations, the average sizes of all the prokaryotic clusters range between 206 and 221 residues, a most surprising degree of size conservation.

The organismal distributions of these proteins are provided in Table 1. The eukaryotic Cluster 1 proteins are all derived from animals, both vertebrates and invertebrates, with the sole exception of a single algal homolog from *C. reinhardtii*. Cluster 2 includes proteins from the Actinobacteria; the α-, β- and γ-proteobacteria; *Deinococcus*; and *Thermus*. Cluster 3 includes proteins from δ-proteobacteria, with the single exception of one firmicute protein. Cluster 4 includes proteins from α-, β- and ε-proteobacteria as well as Actinobacteria. Cluster 5 features homologs only from the α- and γ-proteobacterial classes. Cluster 6 proteins are from Bacteroidetes, with the exception of three γ-proteobacterial homologs and one spirochete protein. The three Cluster 7 proteins and the two Cluster 9 proteins derive exclusively from γ-proteobacteria.

Cluster 8 features mostly proteins from β-proteobacteria, with two γ-proteobacterial and one Chloroflexi protein. The majority of Cluster 10 homologs derive from α-proteobacteria, with the exception of two γ-proteobacterial proteins and one plant protein, which could be localized to mitochondria. Cluster 11 includes proteins primarily from γ-proteobacteria; but three are from β-proteobacteria, and one each is from an α- and a δ-proteobacterium. Verrucomicrobia, Bacteroidetes and Planctomycetes are also sparsely represented. Cluster 12 displays bacterial homologs from Firmicutes and ε-proteobacteria as well as archaeal homologs from Crenarchaeota and Euryarcheaota. The bacterial and archaeal Cluster 13 contains proteins from δ-proteobacteria and Crenarchaeota. Cluster 14 contains proteins from Firmicutes, whereas Cluster 15 contains proteins from Actinobacteria. In conclusion, the clusters exhibit characteristic features with distinctive size ranges and organismal representations. Many of the proteins present in underrepresented organisms may have been obtained by horizontal gene transfer (see below).

Orthologous Relationships of TRIC Family Proteins

Figure 2 shows the phylogenetic relations of complete 16S and 18S rRNA sequences of the genera explored in this study. This unrooted tree was produced using the neighbor-joining method and the FigTree program (http://tree.bio.ed.ac.uk/software/figtree/). The bulk of these sequences are of bacterial origin and encompass a wide variety of bacterial genera. Opposite to the bacterial genera is an intermediate-sized collection of eukaryotic 18S rRNA sequences. Although most of these sequences are derived from the animal kingdom, there are a couple of sequences that derive from the green algal kingdom (*Chlamydomonas* and *Ostreococcus*). The smallest cluster is comprised of archaeal 16S rRNA sequences. Most genera represented in this study are depicted in this tree with the exception of the few unclassified proteins. Genera excluded from the tree are *Tetraodon*, *Taeniopygia*, *Macaca*, *Pan*, *Tribolium*, *Pedobacter*, *Eubacteria* and *Bermanella*. They were excluded either because of their known close relationships with included genera or because the 16/18S rRNA could not be found.

The 16S/18S rRNA tree reveals that the majority of the genera further segregate according to the specific phylum and class. The bacterial section of the phylogenetic tree is characterized by close clustering of the α-, β-, γ-, δ-, ε- and ζ-proteobacterial sequences. As expected, the β- and γ-proteobacterial sequences are closer together than the rest of the proteobacterial sequences, as shown at the top of the tree. Although there are two distinct γ-proteobacterial clusters and one distinct β-proteobacterial cluster, there exists between them a single mixed cluster comprised of one β- and two γ-proteobacterial rRNA sequences, *Thiobacillus*, *Stenotrophomonas* and *Xanthomonas*, respectively. However, it is not surprising to have intermingling of these rRNA sequences since the β- and γ-proteobacteria diverged most recently in proteobacterial evolutionary history and therefore share more similarities. The α-proteobacterial rRNAs cluster closest with the δ-proteobacterial rRNAs at the left side of Fig. 2. A single 16S rRNA sequence from *Mariprofundus* derives from a ζ-proteobacterium, which diverges from the other clusters, in between the β- and γ-proteobacterial rRNAs and the α- and δ-proteobacterial rRNAs. The ε-proteobacterial sequences also exhibit clear divergence from the other proteobacterial clusters. Seen at the bottom left-hand side of the figure, the bacterial cluster of the 16S/18S rRNA tree includes a cluster of firmicute rRNAs with a single spirochete rRNA sequence, a branch where a single *Planctomyces* rRNA can be found and a cluster including rRNAs from Verrucomicrobia, Deinococcus-Thermus, Chloroflexi and Actinobacteria. A Bacteroidetes cluster contains one sequence showing divergence from the rest of the cluster. The archaeal cluster (most from Crenarchaeota) lies in between the bacterial and eukaryotic clusters. The eukaryotic 18S rRNA sequences cluster together and separately from the I6S rRNAs from prokaryotes.

Orthologous relationships between homologs are best determined by comparing clustering patterns between the protein tree and the rRNA tree. Due to lateral gene transfer events, which occurred frequently in prokaryotes but not eukaryotes, orthology among the former organisms is frequently not observed, as discussed below.

Cluster 1 (Fig. S3-1) consists entirely of eukaryotic homologs, all but one from animals. The one exception is from *C. reinhardtii* (Cre1). At the top of the tree are six probable *Drosophila* orthologs, each derived from a different species. The next set of proteins as we progress clockwise around Fig. S3-1 shows four proteins that branch from points near the center of the tree. Two of these proteins, Cin1 and Cin2, are from the sea squirt; another protein, Cre1, is from an alga; and the last, Nve1, is from a sea anemone. The next subcluster includes potential TRIC-A orthologs, where the three fish proteins cluster together, the amphibian protein clusters separately but next to two bird proteins and, finally, five mammalian proteins cluster together. This arrangement is fully consistent with orthology. Continuing clockwise, the next subcluster includes putative TRIC-B orthologs. Once again, the mammalian proteins cluster together, separately from the amphibian, bird and fish orthologs, each comprising its own subcluster. It should be noted that the TRIC-B subcluster is much less compact than the TRIC-A subcluster, reflective of their greater sequence divergence. Next are two *Branchiostoma* paralogs, followed by the *Caenorhabditis* proteins. These

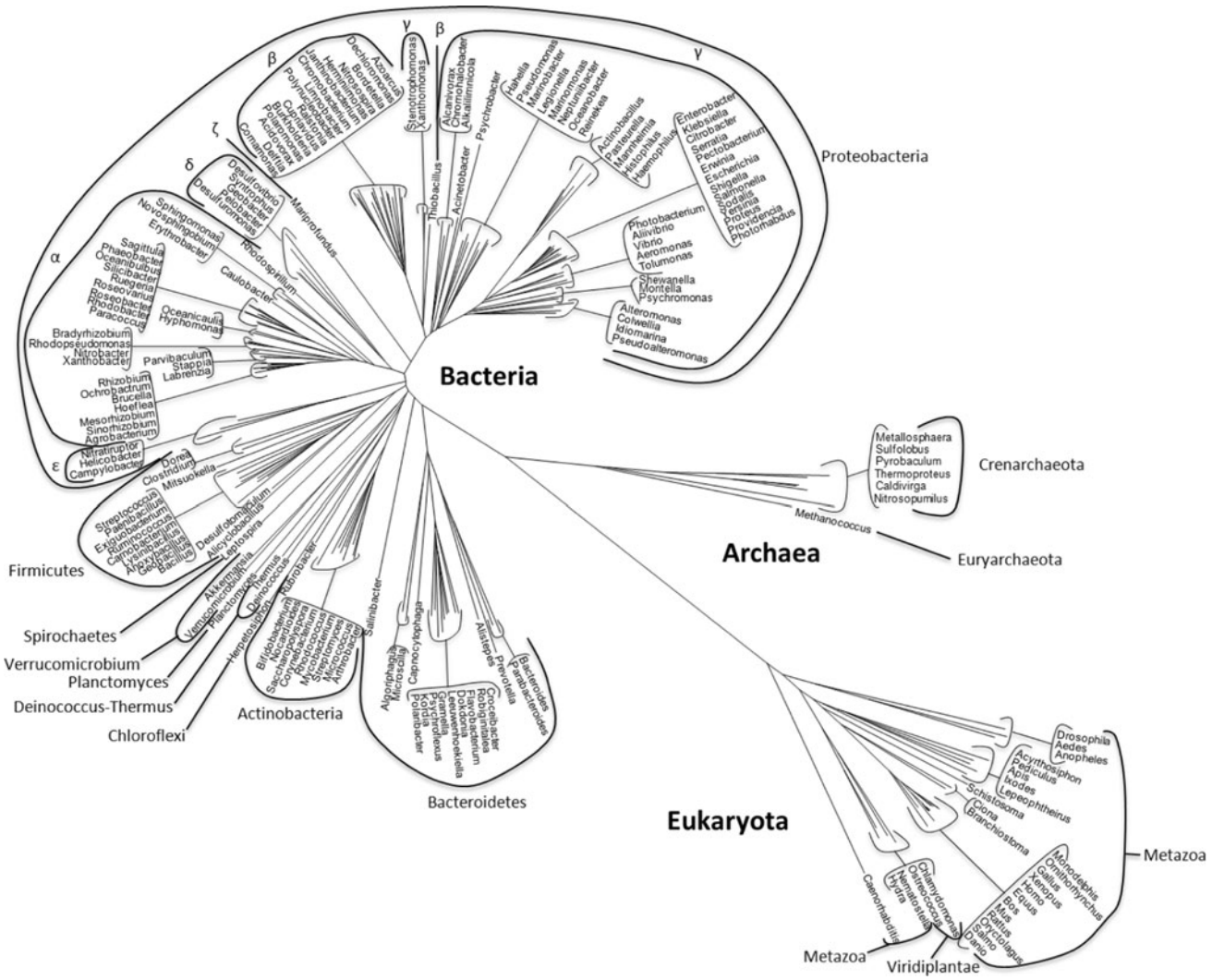

Fig. 2 Phylogenetic tree of 16S/18S rRNA sequences including most of the genera included in this study. *Tetraodon*, *Taeniopygia*, *Macaca*, *Pan*, *Tribolium*, *Pedobacter*, *Eubacteria* and *Bermanella* are not shown. Most sequences are from bacteria; bacterial sequences comprise the largest cluster, eukaryotic sequences comprise a smaller group and archaeal sequences comprise the smallest group. Each cluster is labeled with respect to its domain (*large bold print*), phylum and class (*medium-sized print*). Each branch is labeled by the genera represented

latter proteins represent two sets of paralogs present in two closely related worms, *C. elegans* and *C. briggsae*. The remaining seven proteins that cluster loosely together with the worm proteins are all derived from insects and appear to be orthologous. Finally, Hma1 from *Hydra magnipapillata* branches from a point near the center of the tree, surprisingly distant from the sea anemone and the sea squirt. It seems likely that these distantly related proteins are not orthologs.

Cluster 2 consists of proteins derived from a wide range of organisms including both gram-negative and gram-positive bacteria. The Proteobacterial homologs cluster roughly together (upper left of Fig. S3-2). However, it is clear that these proteins are not orthologous to each other since the α-, β- and γ-proteobacterial proteins are interspersed. The only clusters that exhibit relationships consistent with orthology (indicated in the clockwise direction) are the three proteins Rxy1, Mgi1 and Msp6; the three proteins Dge1, Taq1 and Tth1,; and the large cluster containing Mab1, Ser1 and seven orthologs from seven different *Streptomyces* species.

Cluster 3 (Fig. S3-3) shows a group of δ-proteobacterial proteins, derived from species of *Geobacter* and *Pelobacter*, with the single exception of a firmicute protein, derived from *Desulfotomaculum*. While there are probable paralogous relationships, there is no indication of orthology.

Cluster 4 is large and complex, containing 10 subclusters. These are numbered in Fig. S3-4. The proteins derive from α-, β-, γ- and ε-proteobacteria as well as Actinobacteria.

Subcluster 1 consists of γ- and ε-proteobacterial homologs that segregate as expected according to proteobacterial class. Thus, all γ-proteobacterial proteins group together, and all ε-proteobacterial proteins group together. However, neither of these clusters consists exclusively of orthologs. For example, in the ε-proteobacterial cluster, *Campylobacter* proteins flank the *Helicobacter* homolog. Proteins from subclusters 2, 3 and 4 derive exclusively from γ-proteobacteria. While subclusters 2 and 3 do not appear to consist of orthologs, the relative phylogenetic distances observed in subcluster 4 are consistent for orthology. The same is true for subcluster 5, which consists of five probable orthologs from *Corynebacteria*. Subcluster 6 consists of a single protein from *Alcanivorax*. Subclusters 7–10 all appear to exhibit intermixing. In general, it appears that horizontal transfer has occurred extensively in the Proteobacteria. The appearance of a group of orthologous corynebacterial proteins within Cluster 4 suggests that, while no horizontal transfer has occurred between these proteins, they may have acquired them by horizontal transfer from a γ-proteobacterium in a single step prior to the divergence of the corynebacterial species. The relationships of the three members in Cluster 5 (Fig. S3-5) are similarly inconsistent with orthology.

Cluster 6 (Fig. S3-6) consists primarily of proteins from the Bacteroidetes phylum except for one spirochete (Lbi1) and three tightly clustering γ-proteobacterial proteins. Interestingly, Lbi1 clusters loosely with Asp6 from *Algoriphagus*, while the three γ-proteobacterial proteins cluster with *Capnocytophaga*. Subclusters 1 and 2 consist of proteins from various Bacteroidetes genera, but the order of these proteins differs from that of the rRNAs, suggesting either that there has been horizontal transfer of these protein genes or that there is substantial error in the corresponding parts of one of these trees. Subcluster 3 contains Cps1 from *Capnocytophaga* (within the Bacteroidetes phylum) and three closely related γ-proteobacterial orthologs, Hpa1, Mha1 and Apl1. It seems likely that a gene transfer event from a Bacteroidetes species to the common ancestor of the three γ-proteobacteria occurred just once. Subcluster 4 includes five proteins, four of which are probably orthologs. However, the fifth protein, Lbi1, is from the spirochete *Leptospira*. This is another clear example of horizontal gene transfer from a Bacteroidetes species to another bacterial phylum. Finally, subcluster 5 features 11 proteins derived from closely related members of the Bacteroidetes phylum that may well be orthologous and therefore serve a single function in all of these organisms. Cluster 7 (Fig. S3-7) consists of three proteins, all derived from *Psychrobacter* species, which are likely to be orthologous and serve a single function.

Cluster 8 (Fig. S3-8) includes two subclusters at the top and bottom of the tree. The top subcluster consists of four sub-subclusters, the most distant derived from γ-proteobacteria. With one exception, all remaining proteins in this subcluster are derived from β-proteobacteria. The one exception, Hau1, is derived from a distant bacterial phylum, Chloroflexi. It seems likely that the last mentioned protein was obtained by *Herpetosiphon* by horizontal transfer from a β-proteobacterium. In the lower subcluster, all proteins may be orthologous including Bce1 and Bce4, which are derived from two different strains of *Burkholderia cenocepacia*. However, the closeness of Cta1 and Rme1 suggests that one of these two proteins may have been the product of horizontal gene transfer. Cluster 9 (Fig. S3-9) features only two probable orthologs from two closely related genera, *Alteromonas* and *Pseudoalteromonas*.

Cluster 10 (Fig. S3-10) includes proteins from α-proteobacteria with three exceptions, two from γ-proteobacteria and one from the alga *Ostreococcus*. This last may well be a mitochondrial protein because the ancestor of mitochondria was an α-proteobacterium. Perhaps the two γ-proteobacterial proteins derived from an α-proteobacterium by horizontal transfer. It is interesting to note that the plant and the two γ-proteobacterial proteins together with one α-proteobacterial protein, Oal1, cluster together (subcluster 3), while all proteins in other subclusters are derived from α-proteobacteria. Subclusters 1, 2, 4 and 5 appear to consist of sets of orthologs and this might be true of subcluster 6 as well; however, several discrepancies between the proteins in subcluster 6 and the corresponding genera of the rRNA tree suggest that horizontal gene transfer may have occurred among these closely related species.

Cluster 11 (Fig. S3-11), like Cluster 10, is somewhat complex. This tree can be subdivided into three primary subclusters. The top subcluster 1 is surprisingly diverse with members from four phyla (Proteobacteria, Bacteroidetes, Verrucomicrobia and Planctomyces), and among the Proteobacteria we have representatives from the α, β, γ and δ classes. While most of these proteins are distantly related to each other, we nevertheless note that these orthologs are intermixed, suggestive of trans-phylum genetic exchange. In contrast to subcluster 1, subcluster 2 is derived exclusively from γ-proteobacteria; and the relationships of the proteins are in agreement with orthology. Subcluster 3 is derived from γ-proteobacteria with one exception, Vba1 from a distinct phylum, Verrucomicrobia. This protein is an obvious candidate for horizontal gene transfer. It appears that subcluster 3 consists of a collection of orthologs with some paralogs specifically from species of *Pseudoalteromonas* and related genera.

Cluster 12 (Fig. S3-12) features two subclusters at the top and bottom of the tree. The top subcluster consists exclusively of Firmicutes, and many of the proteins may be orthologs. However, the positions of some of the proteins are indicative of horizontal transfer. For example, the

Paenibacillus and *Geobacillus* proteins cluster more closely in the protein tree than the *Exiguobacterium* homolog, although the opposite relationship is observed in the rRNA tree. Note that the *Geobacillus* and *Anoxybacillus* proteins cluster within the large group of *Bacillus* proteins. Comparison with the rRNA tree leads us to suggest that these latter two genera actually represent a subdivision of the bacilli. Most of the proteins in the bottom subcluster are derived from archaea, but three bacterial proteins are also present. These three proteins are distantly related to each other as well as the *Methanococcus* sequence, Mma2, the only homolog in this subcluster from the Euryarchaeota. All remaining archaeal proteins, which cluster relatively closely together, are members of the Crenarchaeotal phylum. The sub-subcluster of five proteins to the left is derived from *Pyrobaculum* species with the single exception of a *Thermoproteus* protein, Tne1, which is very similar to the Pis1 protein from *Pyrobaculum*. These two genera are closely related to each other, so it is not possible to determine if these proteins are orthologs or arose by horizontal transfer. If the former, *Thermoproteus* may truly belong to the *Pyrobaculum* genus.

Cluster 13 (Fig. S3-13) consists of only six proteins from two δ-proteobacteria (Sac2 and Dvu1), one ζ-proteobacterial protein (Mfe1) and three putative crenarcheaotal proteins (Nma1, Orf4 and Orf7) at the bottom of the tree. In view of the fact that two of the latter proteins were obtained from uncultured and unclassified Crenarchaeota, little can be said about orthologous relationships within this cluster. The five Cluster 14 proteins (Fig. S3-14) derive from Firmicutes and exhibit phylogenetic relationships that are clearly inconsistent with orthology. Finally, only two Actinobacterial proteins comprise Cluster 15 (Fig. S3-15), so nothing can be said about their potentially orthologous relationship.

Establishment of Homology and Motif Analyses of Prokaryotic and Eukaryotic Proteins

Eukaryotic TRIC-A homologs were shown to be homologous to a much larger group of prokaryotic proteins of the same topology. Two proteins in TCDB are the eukaryotic mouse TRIC-A protein (TC 1.A.62.1.1, Acc Q3TMP8) and the prokaryotic (archaeal) *S. solfataricus* protein (TC 1.A.62.3.1, Acc Q981D4). The former protein is closely related to a frog homolog (Acc Q6GN30) where the two proteins gave an e value of e^{-70} with BLAST. The latter protein is closely related to a *Bacteroides* protein (Acc A0M015) with a BLAST e value of e^{-26}. Comparison of the frog protein with the *Bacteroides* protein using the IC program, confirmed with the GAP program, yielded a comparison score of 13.9 SD (Fig. 3). This value is in excess of what is required to establish homology (Saier 1994; Saier et al. 2009). This alignment shows 22.1% identity and 33.8% similarity. Invoking the superfamily principle, these values are sufficient to establish homology between the prokaryotic and eukaryotic proteins (Doolittle 1981, 1986).

Two more alignment comparison programs were used to confirm these results. Using GGSEARCH, which implements the Needleman-Wunsch algorithm and has an e-value threshold of e^{-3} to suggest homology, a comparison of the sole eukaryotic cluster against all the prokaryotic clusters yielded an e value of 0.00013, comparing the *Acyrthosiphon pisum* homolog (gi 193594177) with the *Haemophilus somnus* homolog (gi 133460555). Furthermore, a second test using the homologs of the prokaryotic clusters as the first set and the eukaryotic cluster as the second set generated several e values, the most significant of which was 0.015 from the same pair of homologs. The other standardized comparison program used, HMMER 2.0, requires a profile hidden Markov model input and a profile hidden Markov model database to assess protein relationships and uses an e-value threshold of 0.1 (Eddy 1998). With the eukaryotic cluster as the profile HMM and the prokaryotic cluster as the HMM database, the comparison gave an e value of 0.07 when the profile was paired with a *Syntrophus* homolog (gi 85859719). Similarly, with the prokaryotic cluster as the profile HMM and the eukaryotic cluster as the HMM database, an e value of 0.0011 resulted when the profile was paired with a *Nasonia* homolog (gi 156546697) from Table 1.

In order to corroborate these results, the MEME program (Bailey and Elkan 1995) was used to identify three conserved motifs which shared common features between the eukaryotic and prokaryotic homologs. For this purpose, all of the 43 eukaryotic homologs as well as the 299 prokaryotic homologs listed in Table 1 were used. Figure 4 shows these three motifs where the eukaryotic consensus motif (top) is aligned with the prokaryotic consensus motif (bottom). Limitations of the MEME program disallowed the input of all prokaryotic sequences in a single run, so the sequences were randomly split into two separate runs. The resulting consensus motifs were nearly identical between the two prokaryotic groups and are therefore reported as a single prokaryotic entity for each motif (refer to Fig. 4).

Figure 4 demonstrates the alignments of the corresponding prokaryotic and eukaryotic motifs. Motif 1 occurs in TMS 1 in both prokaryotic and eukaryotic proteins. The alignment shows 54% identity and 62% similarity with no gaps when using the approach shown in Fig. 4. Motif 2 occurs in TMS 2. This motif contains 38% identity and 77% similarity with no gaps. Motif 3 occurs at the end of TMS 3 and the beginning of the loop region between TMSs 3 and 4, where there is 42% identity and 58% similarity with one gap in the prokaryotic motif. These results further

```
Euk      9 VQFSQLSMFPFFDMAHYLASVMSAREQAGALDIASH..SPMA 48
           .   :||:|   |.   :|  .|     |||
Prok     1 MPSMELSLFNILDVLGTIAFAIS.....GALSAMNRRLDLFG 37

Euk     49 SWFSAMLHCFGGGILSSILLAEPPVGILANTTNIMLASAIWYMVYYFPYD 98
            .  | .   ||| .  ||: | ||  : ||  : |   :  :
Prok    38 IFIIAFVTAIGGGTVRDILIGETPVTWMENTVYVYLIGVVTILAIIFRNK 87

Euk     99 LFYNCFFFLPIRLIAAGMKEVTRTWKILSGITHAHSHYKDAWLVMITIGW 148
           : |          |  |.  :|       |:     .  |  :: : :|
Prok    88 INYLKKSLFLFDTIGLGVFTIT.......GVETGIQNDLDP.IISVALGA 129

Euk    149 ARGAGGGLISNFEQLVRGVWKPESNEFLKMSYPVKVTLIGAVLFTLQHGH 198
             |  ||.|       | :   |     :        ||||. :   :
Prok   130 MTGTFGGVI.......RDILCNEIPVIFRKEIYATACLIGALAYVTLYD. 171

Euk    199 YLPISRHNLMFIYTMFLVSIKVTMMLTHSAGSPFLPLETPLHRI 242
            | .    :  : .: .:||:: ..  |     | |      ||
Prok   172 .LGMDSDVIYIVTSLTVISIRIVVVKYHITLPSFYPTSPNSSRI 214
```

Fig. 3 Alignment of a major segment of a eukaryotic TRIC homolog with the corresponding segment of a prokaryotic TRIC homolog. This alignment was used to establish homology among TRIC proteins of eukaryotes and prokaryotes. The eukaryotic group is represented by a *Xenopus laevis* protein (gi 147900352, Acc Q6GN30); the prokaryotic group is represented by a *Gramella forsetii* protein (gi 117577491, Acc A0M015). The IC program was used to identify the most similar pair of proteins. The GAP program was used to produce the alignment and confirm homology with default settings and 500 randomized shuffles, which gave a comparison score of 13.9 SD. The residue positions are denoted by numbers at the beginning and end of each line. The alignment shows identity of 22.1% and similarity of 33.8%. The plot reveals identities (|), conserved substitutions (:), and more distantly conserved substitutions (.)

support the conclusion that the eukaryotic and prokaryotic proteins share a common origin, possibly providing related functions. They therefore belong to a single family.

Topological Analyses of Eukaryotic and Prokaryotic TRIC Family Homologs

The multiple alignments shown in Figs. S1-1 (all 342 proteins), S1-2 (prokaryotes) and S1-3 (eukaryotes) were used to generate average hydropathy and similarity plots using the AveHAS program (Zhai and Saier 2001b). These plots are shown in Fig. 5a–c, respectively. Sequence analyses described below led to the conclusion that the region in the alignment between alignment positions 180 and 240 in Fig. 5a represents the first TMS, which is spread out due to the presence of several gaps in the multiple alignment. Peaks 2–7 follow as labeled. It can be seen that the transmembrane region is flanked by extensive hydrophilic regions, about 170 alignment positions in both the N- and C-terminal regions. In the far N-terminal region of the plot, a strong hydrophobic peak is observed, followed by a long hydrophilic region immediately preceding the first TMS. The hydrophobic peak, found in the single protein, Dgr1 of *Drosophila grimshawi*, could be a targeting sequence for the general secretory apparatus. The following hydrophilic region of 100 residues occurs in numerous *Drosophila* species, and these proteins are presumably orthologs of each other. No conserved domain was recognized by CDD. Hma1, from *Hydra magnipapillata*, had a 94-residue N-terminal extension found in numerous eukaryotic proteins. Proteins containing this region of homology exhibit overlapping PHD/BAH finger domains involved in protein–protein interactions. They include the chimeric MOZ-ASXH2 fusion protein of *Homo sapiens* (Acc BAD00088), many MYST histone acetyltransferases (e.g., Acc CAM14129) and the monocytic leukemia zinc finger protein of *Danio rerio* (Acc AAT11171). Three homologs, Tni1 of the pufferfish, Cbr1 of the roundworm and Cre1 of the alga *C. reinhardtii*, had hydrophilic C-terminal extensions that showed no sequence similarity with each other or any other protein in the NCBI protein database.

The prokaryotic AveHAS plot was much clearer than that of the eukaryotic homologs (Fig. 5b). Seven peaks of hydrophobicity correspond to the seven putative TMSs. Charge analysis, using the positive-inside rule, confirmed the topological orientation of the loops that connect the prokaryotic TMSs (von Heijne 1986, 1992) with the N termini outside and the C termini inside. In fact, there were no discrepancies; all putative cytoplasmic loops bore more Ks and Rs than any of the putative extracytoplasmic loops, as quantified in Fig. 5a–c. This tendency is not as clearly seen in the average hydropathy plot of the eukaryotic cluster because the positive-inside rule is less pronounced for eukaryotic proteins (Gafvelin et al. 1997), as seen in Fig. 5a.

```
Motif 1
Eukaryotic:   [VYF] F D  [IL] A [HY]  Y  [IL] [LV]  S  [AVI] [LM] [YSA]
               |    . |   |                      |       |      |    |
Prokaryotic:  [VI]  L D [LI]  I  G  [TIV]  A  [AV] [FE] A    [IM]  [ST]

Motif 2
Eukaryotic:   [LF]  [SC]  A   M L Y C   F  [GA] [GS] [GY] I   L
                :         |   . :   :   .   |    |    |       |
Prokaryotic:  [VI] [LIV] [AG] F V T A [ILV] G    G    G   T [ILV]

Motif 3
Eukaryotic:   Y L [IV]   F [YF] C   P F  [DN] [LI] [FG] Y
              |     |    |  :              |    .        |
Prokaryotic:  Y P  [VLP] F  W  [VI] K -  [DNH] P    E   Y
```

Fig. 4 MEME analysis and alignment of the eukaryotic vs. prokaryotic conserved motifs. Motif 1 occurs in TMS 1 for both eukaryotic and prokaryotic sets. The alignment exhibits conserved residues of 54% identity and 62% similarity with no gaps. Motif 2 occurs in TMS 2 for both sets. The alignment exhibits 38% identity and 77% similarity with no gaps. Motif 3 occurs at the end of TMS 3 and the beginning of the loop region between TMSs 3 and 4. The alignment exhibits 42% identity and 58% similarity with one gap (−) in the prokaryotic consensus motif. Residues in brackets represent alternative possibilities at any one position, with the dominant residue presented first. Percent identity and similarity are as defined here and for the GAP program

Poorly conserved hydrophilic N and C termini were present in most of these proteins. A domain within the homolog, Ota1 of *Ostreoccocus tauri*, proved to be related to a FlgB domain found also in several basal body proteins (FlgB, FlgC, FlgE, FlgF, FlgG and FlgK) of the bacterial flagellum (Wong et al. 2007) with as much as 46% identity and 59% similarity between Ota1 and various FlgB homologs. In addition, the same region showed similar degrees of identity and similarity with a central hydrophilic region between TMSs 7 and 8 in RND-type multidrug resistance pumps. At the C termini of two actinobacterial homologs, both from *Corynebacteria*, Cur1 and Cje2, C-terminal extensions were present that showed no sequence similarity with other proteins. Figure 5c presents the average hydropathy plot for the 342 eukaryotic and prokaryotic proteins included in this study.

Evolutionary Origin of TRIC Family Proteins

Many, perhaps most, integral membrane transporters have arisen by intragenic duplication events (Saier 2003b). We therefore examined the TRIC family homologs from both prokaryotes and eukaryotes for repeat sequences that would indicate the evolutionary origins of these proteins. All prokaryotic and eukaryotic TRIC family homologs proved to be homologous throughout their lengths, and consequently, the superfamily principle could be applied to look for internal repeats using the IC and GAP programs (Yen et al. 2009; Zhai and Saier 2002). Representative results are presented in Fig. 6, which shows an alignment of TMSs 1–3 of a TRIC family homolog from *Shewanella amazonensis* (gi 119775597) with TMSs 4–6 of a second homolog from the *Nitratiruptor* genus (gi 152990839). This alignment shows 40.0% identity and 53.3% similarity. Using IC and confirming with GAP, this alignment gave a comparison score of 19.4 SD, far in excess of what is required to establish homology (Saier 1994; Yen et al. 2009). TMS 7 showed no sequence similarity with other regions of the homologous proteins.

In order to confirm the presence of the three-TMS repeat sequence, several control experiments were conducted. In order to eliminate the hydrophilic vs. hydrophobic contrast, the hydrophilic loops were removed and the remaining hydrophobic helices were artificially fused. Using the same two proteins as listed above, TMSs 1–3 of the *Shewanella* protein were compared with TMSs 4–6 and 5–7 of the *Nitratiruptor* protein and TMSs 2–4 of the *Shewanella* protein were also compared with TMSs 5–7 of the *Nitratiruptor* protein. In the first example (TMSs 1–3 vs. TMSs 4–6), all three TMSs aligned, and the comparison score was 15.7 SD. In the second comparison (TMSs 1–3 vs. TMSs 5–7), TMSs 2 and 3 aligned with TMSs 5 and 6, and the comparison score was 11.7 SD. Finally, when TMSs 2–4 were aligned with TMSs 5–7, again TMSs 2 and 3 aligned with TMSs 5 and 6 with a comparison score of 11.0 SD. This control served two purposes: (1) to show that elimination of the hydrophilic residues did not prevent retention of good comparison scores and (2) regardless of the three-TMS comparisons, the program always aligned TMSs 2 and 3 with TMSs 5 and 6. With this evidence, we therefore conclude that these proteins arose by duplication of a three-TMS-encoding gene segment giving rise to six TMS proteins to which a seventh TMS of unknown origin was added at the C terminus. Our favored evolutionary pathway is represented in Fig. 7a, along with two other possible pathways (Fig. 7b, c).

Another control that was performed was to compare 100 different repeat 1 sequences (TMSs 1–3) with 100 different

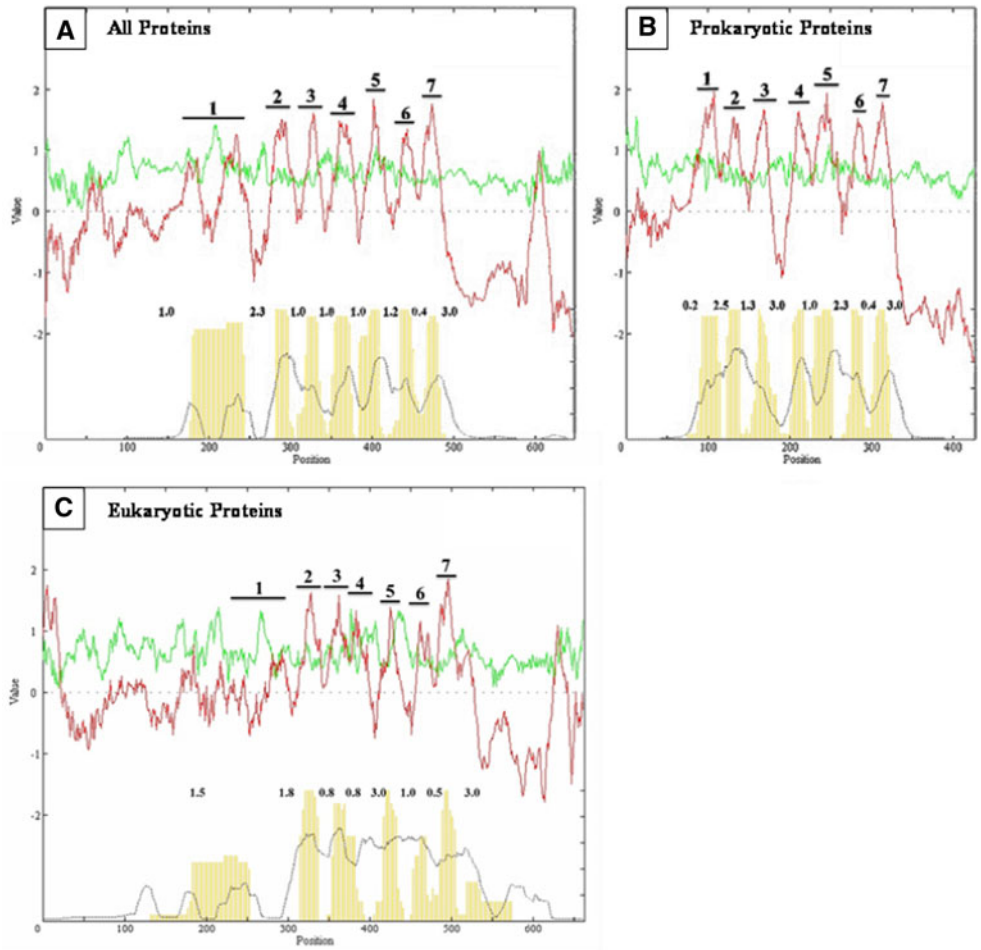

Fig. 5 The average hydropathy (*dark line, top*), amphipathicity (*light line, top*) and similarity (*dotted line, bottom*) plots of the TRIC family sequences included in this study as generated by the AveHAS program. *Light vertical lines* (*bottom*) provide an independent estimation of transmembrane segments (TMSs). **a** Results for the 342 TRIC proteins included in this study. The TRIC domain is located at the center of the plot (TMSs 1–7) for each graph. Charge analyses of each hydropathy plot (*bold print, center*) present the average numbers of Rs and Ks per protein in each loop region before and after each TMS. See text for analysis of the hydrophilic N- and C-terminal domains. **b** The 299 prokaryotic TRIC homologs. The vast majority of these sequences show seven conserved peaks of hydrophobicity, which are believed to correspond to seven TMSs. Putative TMSs 1–7 are labeled. **c** The 43 eukaryotic TRIC homologs with most showing seven TMSs and with TMS 1 being relatively poorly conserved. Putative TMSs 1–7 are labeled on the graphs

repeat 2 sequences (TMSs 4–6) of the TRIC proteins. These were compared with results obtained when either repeats 1 of the TRIC proteins was compared with repeats 2 (TMSs 5–7) of the microbial rhodopsins (TC 3.E.1) or repeats 2 of TRIC proteins were compared with repeats 1 (TMSs 1–3) of microbial rhodopsins (Kuan and Saier 1994; Zhai et al. 2001). Using the IC2 program, which is essentially the same as IC, except that it uses a cutoff to eliminate low comparison scores, 36 high scores were obtained when the two repeat sequences of the TRIC proteins were compared but only one such score was obtained when the three-TMS repeat elements of the TRIC proteins were compared with the three-TMS repeat elements of the microbial rhodopsins. This control also substantiated the significance of the statistical data responsible for concluding that TRIC family proteins contain two adjacent three-TMS repeat elements.

To further confirm the IC and GAP results for this internal duplication, two more alignment comparison programs were utilized: GGSEARCH and HMMER 2.0. Given that e values of e^{-3} or smaller are considered significant and e^{-8} or smaller establishes homology, the GGSEARCH program evaluated the alignment of TMSs 1–3 of the *Psychrobacter arcticus* homolog (gi 71065295) with TMSs 4–7 of a *Nitratiruptor* homolog (gi 152990839) with an e value of $2e^{-20}$. A second test with reversed inputs resulted in the same pair of regions scoring with an e value of $5.8e^{-19}$. Many other organisms gave values in excess of those required to substantiate the claim of homology. Using HMMER 2.0 and a threshold of 0.1, TMSs 1–3 as the

Fig. 6 Alignment of TMSs 1–3 (*Half 1*) with TMSs 4–6 (*Half 2*) of TRIC family homologs using the GAP program. TMSs 1–3 are from a *Shewanella amazonensis* protein (gi 119775597), while TMSs 4–6 are from a *Nitratiruptor* protein (gi 152990839). The IC program was used to identify the most similar pair of proteins. The GAP program was used to produce the alignment and confirm homology with default settings and 500 randomized shuffles, giving a score of 19.4 SD. Residue positions are denoted by numbers at the beginning and end of each line. TMHMM and HMMTOP were used to determine the positions of TMSs 1 (1–16), 2 (25–42), 3 (52–74), 4 (96–113), 5 (117–134) and 6 (143–161). These programs generally, but not always, agree. The alignment shows identity of 40.0% and similarity of 53.3%

profile HMM input compared with TMSs 4–7 as the HMM database attained an e value of $2.1e^{-5}$ when aligning the profile against the *Nitratiruptor* homolog (gi 152990839). An additional study with TMSs 4–7 as the profile HMM and TMSs 1–3 as the HMM database resulted in a maximal e value of $2.1e^{-7}$, where the profile matched best with the *Geobacter lovleyi* homolog (gi 189426048). As expected, TMSs 1–3 aligned with TMSs 4–6. TMS 7 did not align.

These results and values establish that the two halves of these proteins share a common origin. With this evidence, we conclude that TRIC proteins arose by duplication of a three-TMS-encoding gene segment giving rise to six-TMS proteins to which a seventh TMS of unknown origin was added at the C terminus. Our favored evolutionary pathway is represented in Fig. 7a.

Functional Analyses on Prokaryotic Homologs

The SEED database (Overbeek et al. 2005) was used in order to determine the genome context of the various TRIC homologs found in prokaryotes. The numbering system used below corresponds to the numbers provided in the SEED database for the proteins itself (always 1), with 2, 3, 4, etc. being the proteins that most frequently occur with protein 1.

In Cluster 2, the TRIC homolog Sco1 (gi 21223838) from *Streptomyces coelicolor* A3(2), appears to be a peptide uptake or an amino acid efflux transporter for the following reasons:

- In *Streptomyces coelicolor*, the TRIC homolog (designated 1 below) is in an operon with a complete ABC oligopeptide transporter, as well as a TesB acyl-CoA esterase type 3.
- In *Streptosporangium roseum*, in the same operon with 1, there is a di-/tripeptide permease DtpT and a GntR regulator is divergently transcribed.
- In *Nitrobacter hamburgensis*, 1 is in an operon with a peptidase.
- In *Nitrosospira multiformis*, 1 is in an operon with an *N*-acetyl-L,L-diaminopimelate deacetylase (amino acid metabolism).
- In *Burkholderia cepacia*, it is in an operon with a metal-dependent amidase/aminoacetylase/carboxypeptidase.
- In *Mycobacterium* sp. MCS, 1 is in an operon with a DedA putative permease (function unknown).
- In *Stackebrandita nassauensis*, divergently transcribed from 1 is a peptidase, M48.
- In *Pelobacter carbinolicus*, the TRIC homolog is referred to as YadS (YadA is present in *Escherichia coli*). It is divergently transcribed from a rhomboid family protein, a peptidase. Furthermore, present in the same operon is a fumarylacetoacetate hydrolase, which is necessary for tyrosine catabolism.

Using the protein Gme 1 (gi 78221477) from *Geobacter metallireducens* GS-15, a Cluster 3 homolog as the query, the following hits were obtained:

- In *Flavobacterium* sp. BBFL7, 1 is divergently transcribed from two genes encoding a rhomboid family peptidase and an RDD putative transport protein (TC 9.B.45).
- In *Geobacter uraniireducens*, 2 is a hypothetical protein that is divergently transcribed from 1 and is involved in lipopolysaccharide synthesis.
- In *Desulfuromonas acetoxidans*, there is a rhomboid serine protease that is divergently transcribed in multiple organisms.
- In *Desulfotomaculum reducens* MI, the gene cluster includes a di-/tripeptide uptake permease (DtpA).
- In *Dyadobacter fermentans* D., within the same operon as 1, there is a thymidylate synthase, in the pyrimidine conversion pathway.

We suggest that the TRIC family proteins of Cluster 3 are involved in export of amino acids or their breakdown products following peptide hydrolysis. A role for some of these proteins in nucleotide export is possible.

Using the protein Vch1 (gi15642114) from *Vibrio cholera* O1 biovar E1 Tor str. N16961 in Cluster 4, most evidence suggests that the TRIC homolog may be involved in nucleotide or nucleoside export, but some members of Cluster 4 bring up operons that may be involved in amino acid metabolism. Thus, we found the following: 2 (DNA

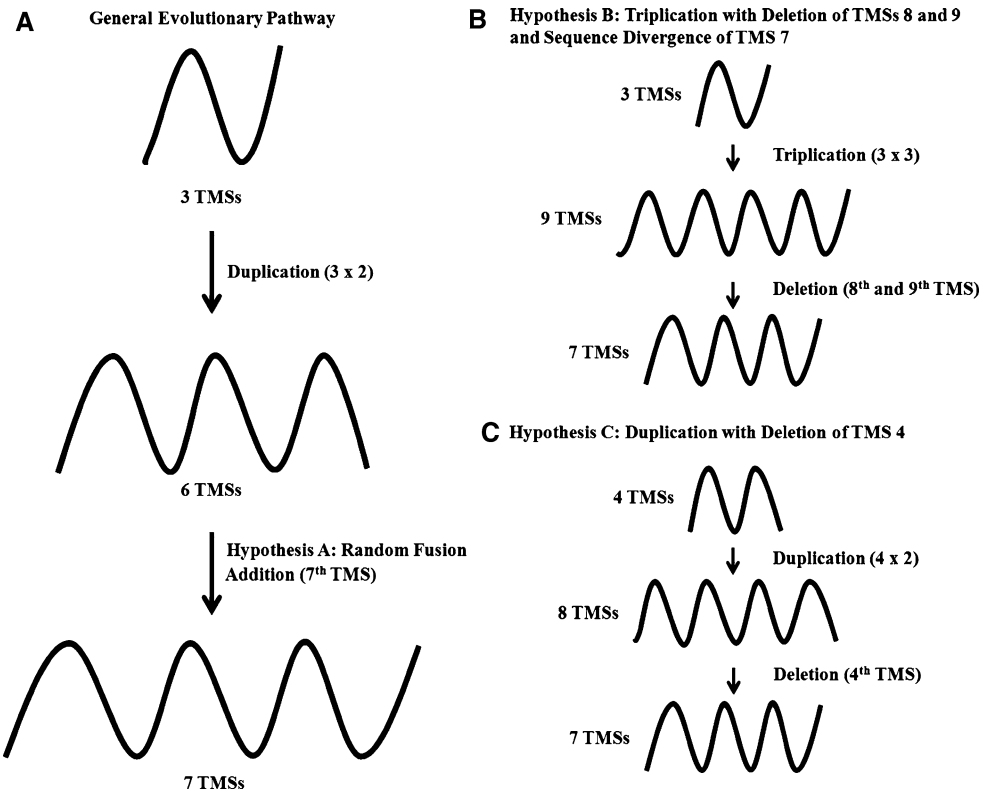

Fig. 7 Three possible evolutionary pathways for the appearance of all members of the TRIC family. **a** Our favored pathway: an initial three-TMS-membered primordial protein duplicated at the gene level to give a six-TMS protein, with a seventh TMS added by fusion. **b** Triplication of the three-TMS progenitor gave nine TMSs, followed by deletion of TMSs 8 and 9 and sequence divergence of TMS 7. Preliminary GAP alignments failed to find significant similarity of TMS 7 with TMS 1 or 4. **c** A primordial four-TMS protein-encoding genetic element duplicated to give an eight-TMS-membered structure or was only partially duplicated to give a seven-TMS protein. In the former case, the eight-TMS protein gave rise to the seven-TMS protein by deletion of the fourth TMS

ligase), 3 (guanylate kinase for purine conversions), 4 (RNA polymerase omega subunit regulatory protein), 5 (ppGpp synthase), 6 (tRNA methyltransferase), 7 (YicC of unknown function), 8 (ATP-dependent DNA helicase), 9 (ribonuclease PH), 12 (NTP hydrolase). Proteins 3–6 and 8 are together in a single operon. These operons are the ones that house the largest numbers of TRIC family homologs in Cluster 4. Therefore, protein 1 could be involved in nucleotide export.

The *Reinekea* sp. MED297 homolog (gi 88800179) from Cluster 5 was used as the query sequence. In *Shewanella putrefaciens* CN-32, genes near the TRIC-A homolog-encoding elements include 11 (glutamate synthase, large chain), 10 (glutamate synthase, small chain), 2 (5′-methylthioadenosine nucleosidase) and 5 (adenosylcoinamide-phosphate synthase), which are all convergently transcribed. Also, 8 is a peptidase, transcribed convergently and 13 is a tyrosyl-tRNA synthetase. All of these enzymes function in amino acid metabolism, suggesting that this set of TRIC homologs are concerned with the transport of amino acids and their derivatives.

Using the Cluster 6 member *Polaribacter irgensii* 23-P (gi 88803173) as the query sequence, we found homologs of all subunits of the RNF (H^+ or Na^+)-translocating NADH-ferredoxin oxidoreductase encoded within an operon that also encodes the TRIC homolog. Possibly in these organisms, the TRIC homologs play a role in electron transfer, possibly transporting a substrate or product of the electron transfer chain. Some of the other operons obtained from this search did not encode RNF systems. In operons lacking RNF subunit-encoding genes, a diverse group of functional proteins could be identified. These included ATP-dependent RNA helicases, thioesterases, MDR efflux pumps, peptidylprolyl *cis–trans* isomerases, adenosylcobinamide-phosphate synthase and 5′-methylthioadenosine nucleosidase. These last two genes appear to be within an operon that also includes YadS, the TRIC family member.

The query sequence for Cluster 7 was the *Psychrobacter cryohaloentis* K-5 homolog (gi 93005552). For all operons, YadS homologs are designated as 1. In *Polaribacter irgensii* 23-P, there is an ABC transporter permease protein (2), a membrane fusion efflux protein (36), a multi-antimicrobial extrusion protein (Na^+/drug antiporter) of the MATE family of MDR efflux pumps (25), a thioeseterase (18), an ATP-dependent RNA helicase (14), two putative Tricorn-like proteases and an acyl dehydratase (22). *Tenecibaculum* sp. MED152 has a very similar operon composition.

Cluster 11 features proteins from *E. coli*, Eco1 (gi 15799841). Protein 1 (YadS) is in the same operon with genes encoding 3 (ABC transporter BtuF) and 5 (*S*-adenosylhomocysteine nucleosidase). Other gene products encoded within this gene cluster include 2 (ErpA, an iron binding protein), 4 (EriC, an H^+/Cl^- symporter) and 6 (deoxyguanosinetriphosphate triphosphohydrolase). We suggest that the TRIC family homologs function to export one of the products of *S*-adenosylhomocysteine hydrolysis.

Using the Cluster 12 homolog from *Geobacillus kaustrophilus* (gi 56419858) as protein 1, all gene clusters depicted have the TRIC-A homologs in its own operon. In several operons examined, 1 is divergently transcribed from a large operon containing enzymes involved in fatty acid metabolism. In another set of operons, we find 1 encoded together with a gamma-glutamyl phosphate reductase and aspartokinase. In one organism, there exists an exodeoxyribonuclease, possibly in the same operon with YadS. Still other operons encoding the TRIC homolog include fumarylacetoacetate hydrolases. In view of these results, we suggest that the TRIC homologs of Cluster 12 transport a variety of different substrates dealing with fatty acids, amino acids and intermediates of the Krebs cycle.

Discussion

TRIC channels are essential for normal muscle function in mammals. The two channel proteins, TRIC-A and TRIC-B, have different tissue distributions that have led to the suggestion that they are important in many aspects of mammalian physiology. Indeed, TRIC-B, distributed throughout many tissues, is essential for life after birth (see "Introduction" section). However, the study of these channels has so far been restricted to mammals.

The analyses reported here clearly show for the first time that homologs of the mammalian TRIC proteins can be found in all domains of life, bacteria, archaea and eukaryotes. A value of 13.9 SD for the eukaryotic–prokaryotic comparison is substantially in excess of what is required to establish homology (Saier 1994; Saier et al. 2009). The conclusion of homology was further substantiated using three additional independently derived programs based on different assumptions. This crucial conclusion of homology was confirmed by motif analysis, showing that the three best-conserved motifs share substantial sequence similarity between prokaryotic and eukaryotic homologs.

In all three domains of living organisms, these proteins have the same seven-TMS topology in spite of the appreciable size differences observed for prokaryotic vs. eukaryotic members of this family. Similar observations of size differences between homologs of transport proteins within ubiquitous protein families have been documented previously (Chung et al. 2001). Overall, we noted an approximately 30% decrease in size for prokaryotic proteins compared with the eukaryotic homologs, although one cluster of actinobacterial proteins showed an intermediate average size. The discovery of these proteins in prokaryotes leads to a number of questions as to their functions. It is possible that they serve as monovalent cation channels, as in the case of mammals, and that their cellular function could also be countermovements against other ions such as calcium and magnesium. Retention of conserved sequence motifs clearly suggests that at least some structural and functional features are shared by the family members from the three domains of organisms. However, the genome context studies clearly suggested otherwise.

A surprising observation was that one eukaryotic TRIC family homolog appeared in one of the bacterial phylogenetic clusters (Cluster 10). This protein is from *Ostreococcus tauri*, a green alga with the smallest cell size of any eukaryote yet described and with a genome of 12.5 Mbp (Courties et al. 1994; Derelle et al. 2006). It proved to be the most distant member of Cluster 10. Based on these observations, we suggest that this protein may have been obtained by *O. tauri* from the α-proteobacterial precursor of the endosymbiont that gave rise to mitochondria. This suggestion is supported by the fact that Cluster 10 proteins are almost all derived from α-proteobacteria. However, the possibility of lateral gene transfer cannot be eliminated.

Examination of potential orthologous relationships among TRIC family homologs by comparison with 16S/18S rRNAs revealed that horizontal gene transfer between bacteria and eukaryotes was exceptionally rare and that gene exchange between archaea and bacteria was very much more frequent. However, within each of these three domains of life, we found that horizontal gene transfer within the bacterial domain occurred with highest frequency, that within the archaeal domain it seemed to occur with substantially lower frequency and that horizontal gene transfer within the eukaryotic domain occurred with very low frequency. This tendency has been observed for other families of transport proteins (Chan et al. 2010; Smets and Barkay 2005; Gophna et al. 2006).

The topologies of TRIC family proteins revealed a consistent pattern of seven hydrophobic peaks in hydropathy plots, which could well correspond to TMSs. This conclusion was supported by the distribution of positively charged residues (R and K) in these proteins, which further suggested that the N termini are on the outside while the C termini are localized on the inside. We are aware of the proposed TRIC-A three-TMS topology as determined by epitope-tagging analyses (Yazawa et al. 2007). However, our hydropathy and charge analyses suggest that TRIC homologs have seven TMSs (see Fig. 5a–c) (Gafvelin et al. 1997; von Heijne 1986). The presence of a zinc finger domain in one such homolog (Hma1) and of a FlgB domain in another homolog (Ota1), both at the *N* termini of these proteins, suggests that these domains function in protein–protein interactions in the extracytoplasmic space, in this case, in the SR/ER lumen. Such interactions could be important for formation of homo- and hetero-oligomeric proteins.

We were able to demonstrate the presence of an internal repeat in TRIC family homologs. Thus, TMSs 1–3 proved to be homologous to TMSs 4–6, although these duplicate

three-TMS elements are of opposite orientation in the membrane. We could not detect significant sequence similarity between TMS 7 and other parts of these proteins. This led to the possibility of three distinct pathways for the evolution of these proteins. First, a three-TMS-encoding genetic element could have duplicated internally to form a six-TMS protein, and then a genetic element coding for the seventh TMS fused to the six TMS encoding element. Second, the three-TMS element may have triplicated to give a nine-TMS protein, and this protein may have lost its last two TMSs while the remaining C-terminal TMS (TMS 7) underwent extensive sequence divergence. Third, a four-TMS element could have duplicated to give eight TMSs followed by internal deletion of TMS 4 with inversion of the C-terminal region within the membrane. This possibilty seems remote, but a similar scenario has been established for another family of transmembrane proteins (Au et al. 2006). If the primordial sequence giving rise to these seven TMS proteins was a simple three-TMS channel-forming peptide, then it would have formed oligomeric (possibly hexameric) transmembrane structures. It will be interesting to learn if TMS 7 actually plays a role in channel formation. High-resolution 3-D studies using X-ray crystallography of prokaryotic or eukaryotic TRIC family proteins are likely to confirm these findings.

An examination of the biochemical and physiological functions of TRIC family proteins in prokaryotes based on operon structure and organization produced several likely possibilities. Using SEED analytical techniques, it seemed likely that many of the analyzed prokaryotic members are coregulated with other structural genes involved in amino acid, nucleoside or nucleotide transport. These results can be further extrapolated and applied to the function of their respective homologs in neighboring clusters and subclusters. We suggest a role in active metabolite efflux. Further research regarding the functional specifics of each prokaryotic cluster will prove interesting and significant.

Acknowledgment We acknowledge the computational expertise and instructive efforts of Dorjee G. Tamang, Dr. Joshua Kohn for useful advice on phylogenetic tree construction and analysis and the NIH (GM077402) for financial support.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Au KM, Barabote RD, Hu KY, Saier MH Jr (2006) Evolutionary appearance of H^+-translocating pyrophosphatases. Microbiology 152:1243–1247

Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Syst Mol Biol 3:21–29

Chan H, Babayan V, Blyumin E, Gandhi C, Hak K, Harake D, Kumar K, Lee P, Li TT, Liu HY, Lo TCT, Meyer CJ, Stanford S, Zamora KS, Saier MH Jr (2010) The P-type ATPase superfamily. J Mol Microbiol Biotechnol 19:5–104

Chung YJ, Krueger C, Metzgar D, Saier MH Jr (2001) Size comparisons among integral membrane transport protein homologues in Bacteria, Archaea, and Eucarya. J Bacteriol 183:1012–1021

Coronado R, Miller C (1980) Decamethonium and hexamethonium block K^+ channels of sarcoplasmic reticulum. Nature 288:495–497

Courties C, Vaquer A, Troussellier M, Lautier J, Chrétiennot-Dinet MJ, Neveux J, Machado C, Claustre H (1994) Smallest eukaryotic organism. Nature 370:255

Cowan SW, Schirmer T, Rummer G, Steiert M, Ghosh R, Pauptit RA, Jansonius JN, Rosenbusch JP (1992) Crystal structures explain functional properties of two *E. coli* porins. Nature 358:727–733

Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piégu B, Ball SG, Ral JP, Bouget FY, Piganeau G, De Baets B, Picard A, Delseny M, Demaille J, Van de Peer Y, Moreau H (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci USA 103:11647–11652

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–395

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? Science 214:149–159

Doolittle RF (1986) Of urfs and orfs: a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, CA

Eddy SR (1998) Multiple alignment and multiple sequence based searches. http://selab.janelia.org/publications/Eddy98b/Eddy98b-preprint.pdf

Eddy SR (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PloS Comput Biol 4:e1000069

Fink RH, Stephenson DG (1987) Ca^{2+} movements in muscle modulated by the state of K^+-channels in the sarcoplasmic reticulum membranes. Pflugers Arch 409:374–380

Gadsby DC (2009) Ion channels versus ion pumps: the principal difference, in principle. Nat Rev Mol Cell Biol 10:344–352

Gafvelin G, Sakaguchi M, Andersson H, von Heijne G (1997) Topological rules for membrane protein assembly in eukaryotic cells. J Biol Chem 272:6119–6127

Gophna U, Thompson JR, Boucher Y, Doolittle WF (2006) Complex histories of genes encoding 3-hydroxy-3-methylglutaryl-coenzymeA reductase. Mol Biol Evol 23:168–178

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Kuan G, Saier MH Jr (1994) Phylogenetic relationships among bacteriorhodopsins. Res Microbiol 145:273–285

Lang BF, Gray MW, Burger G (1999) Mitochondrial gene evolution and the origin of eukaryotes. Annu Rev Genet 33:351–397

Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3:e130

Matias MG, Gomolplitinant KM, Tamang DG, Saier MH Jr (2010) Animal Ca^{2+} release-activated Ca^{2+} (CRAC) channels appear to

be homologous to and derived from the ubiquitous cation diffusion facilitators. BMC Res Notes 3:158

Meissner G (1994) Ryanodine receptor/Ca^{2+} release channels and their regulation by endogenous effectors. Annu Rev Physiol 56:485–508

Mio K, Kubo Y, Ogura T, Yamamoto T, Sato C (2005) Visualization of the trimeric $P2X_2$ receptor with a crown-capped extracellular domain. Biochem Biophys Res Commun 337:998–1005

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691–5702

Pitt SJ, Park KH, Nishi M, Urashima T, Aoki S, Yamazaki D, Ma J, Takeshima H, Sitsapesan R (2010) Charade of the SR K^+-channel: two ion-channels, TRIC-A and TRIC-B, masquerade as a single K^+-channel. Biophys J 99:417–426

Rios E, Ma JJ, Gonzalez A (1991) The mechanical hypothesis of excitation–contraction (EC) coupling in skeletal muscle. J Muscle Res Cell Motil 12:127–135

Rios E, Pizarro G, Stefani E (1992) Charge movement and the nature of signal transduction in skeletal muscle excitation–contraction coupling. Annu Rev Physiol 54:109–133

Saier MH Jr (1994) Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. Microbiol Rev 58:71–93

Saier MH Jr (2003a) Answering fundamental questions in biology with bioinformatics. ASM News 69:175–180

Saier MH Jr (2003b) Tracing pathways of transport protein evolution. Mol Microbiol 48:1145–1156

Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C (2009) The transporter classification database: recent advances. Nucleic Acids Res 37:D274–D278

Schneider MF (1994) Control of calcium release in functioning skeletal muscle fibers. Annu Rev Physiol 54:463–484

Smets BF, Barkay R (2005) Horizontal gene transfer: perspectives at a crossroads of scientific disciplines. Nat Rev Microbiol 3:675–678

Takeshima H, Komazaki S, Hirose K, Nishi M, Noda T, Iino M (1998) Embryonic lethality and abnormal cardiac myocytes in mice lacking ryanodine receptor type 2. EMBO J 17:3309–3316

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The Clustal_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Tusnády GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283:489–506

Tusnády GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850

von Heijne G (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J 5:3021–3027

von Heijne G (1992) Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule. J Mol Biol 225:487–494

Wang B, Dukarevich M, Sun EI, Yen MR, Saier MH Jr (2009) Membrane porters of ATP-binding cassette transport systems are polyhyletic. J Membr Biol 1:1–10

Weisleder N, Takeshima H, Ma J (2008) Immuno-proteomic approach to excitation–contraction coupling in skeletal and cardiac muscle: molecular insights revealed by the mitsugumins. Cell Calcium 43:1–8

Woese CR (2000) Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 97:8392–8396

Wong T, Amidi A, Dodds A, Siddiqi S, Wang J, Yep T, Tamang DG, Saier MH Jr (2007) Evolution of the bacterial flagellum. Microbe 2:335–340

Yamazaki D, Komazaki S, Nakanishi H, Mishima A, Nishi M, Yazawa M, Yamazaki T, Taguchi R, Takeshima H (2009a) Essential role of the TRIC-B channel in Ca^{2+} handling of alveolar epithelial cells and in perinatal lung maturation. Development 136:2355–2361

Yamazaki D, Yamazaki T, Takeshima H (2009b) Physiological functions of TRIC channels. Seikagaku 81:1004–1008

Yazawa M, Ferrante C, Feng J, Mio K, Ogura T, Zhang M, Lin P-H, Pan Z, Komazaki S, Kato K, Nishi M, Zhao X, Weisleder N, Sato C, Ma J, Takeshima H (2007) TRIC channels are essential for Ca^{2+} handling in intracellular stores. Nature 448:78–83

Yen MR, Choi J, Saier MH Jr (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. J Mol Microbiol Biotechnol 17:163–176

Zhai Y, Saier MH Jr (2001a) A Web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. J Mol Microbiol Biotechnol 3:501–502

Zhai Y, Saier MH Jr (2001b) A Web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. J Mol Microbiol Biotechnol 3:285–286

Zhai Y, Saier MH Jr (2002) A simple sensitive program for detecting internal repeats in sets of multiply aligned homologous proteins. J Mol Microbiol Biotechnol 4:375–377

Zhai Y, Heijne WH, Smith DW, Saier MH Jr (2001) Homologues of archaeal rhodopsins in plants, animals and fungi: structural and functional predictions for a putative fungal chaperone protein. Biochim Biophys Acta 1511:206–223

Zhai Y, Tchieu J, Saier MH Jr (2002) A Web-based Tree View (TV) program for the visualization of phylogenetic trees. J Mol Microbiol Biotechnol 4:69–70

Zhao X, Yamazaki D, Park KH, Komazaki S, Tjondrokoesoemo A, Nishi M, Lin P, Hirata Y, Brotto M, Takeshima H, Ma J (2010) Ca^{2+} overload and sarcoplasmi reticulum instability in TRIC—a null skeletal muscle. J Biol Chem 285:37370–37376