Title

The LysE Superfamily of Transport Proteins Involved in Cell Physiology and Pathogenesis

Author

Tsu, Brian Vay

Publication Date

2015

UNIVERSITY OF CALIFORNIA, SAN DIEGO

The LysE Superfamily of Transport Proteins Involved in Cell Physiology and Pathogenesis

A thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Biology

by

Brian Vay Tsu

Committee in charge:

Professor Milton Saier, Chair
Professor Eric Allen
Professor James Golden

2015

The Thesis of Brian Vay Tsu is approved, and it is acceptable

in quality and form for publication on microfilm and electronically:

____

____

____

Chair

University of California, San Diego

2015

DEDICATION

I dedicate this thesis to my family:

To my father, Ronald, and my mother, Kathy, for providing me with love and support.

To my brother, Richard, and my sisters, Sally and Amy, for all the laughter and excitement we have experienced together.

EPIGRAPH

“Insanity: Doing the same thing over and over again and expecting different results.”

Albert Einstein

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FIGURES

ABSTRACT OF THE THESIS

The LysE Superfamily of Transport Proteins Involved in Cell Physiology and Pathogenesis

by

Brian Vay Tsu

Master of Science in Biology

University of California, San Diego 2015

Professor Milton Saier, Chair

The LysE superfamily consists of transmembrane transport proteins that catalyze export of amino acids, lipids and heavy metal ions. Statistical means were used to show that it includes newly identified families including transporters specific for (1) tellurium, (2) iron/lead, (3) manganese, (4) calcium, (5) nickel/cobalt, (6) amino acids, and (7) peptidoglycolipids as well as (8) one family of transmembrane electron carriers. Internal repeats and conserved motifs were identified, and multiple alignments, phylogenetic trees and average hydropathy, amphipathicity and similarity plots provided evidence that all members of the superfamily derived from a single common 3-TMS precursor peptide via intragenic duplication. Their common origin implies that they share common structural, mechanistic and functional attributes. The transporters of this superfamily play important roles in ionic homeostasis, cell envelope assembly, and protection from excessive cytoplasmic heavy metal/metabolite concentrations. They thus influence the physiology and pathogenesis of numerous microbes, being potential targets of drug action.

INTRODUCTION

1.1 The LysE Superfamily

Members of the LysE superfamily have long been known to catalyze solute export [1]. Three families had been shown to comprise this superfamily: (i) L-lysine and L-arginine exporters (LysE); (ii) homoserine/threonine resistance proteins (RhtB); and (iii) cadmium ion resistance proteins (CadD) [1]. While LysE and RhtB proteins catalyze export of amino acids, the more distantly related CadD proteins are involved in efflux of the heavy metal ion, cadmium (Cd^{2+}) [1,2,3]. Most members of these families share similar sizes, around 200 amino acyl residues (aas), similar hydrophobicity plots suggestive of 6 transmembrane α-helical segments (TMSs), high degrees of sequence similarity within but not between families, and prokaryotic origins [1].

1.2 L-Lysine and L-Arginine Exporters (LysE; TC# 2.A.75)

Widely distributed across the domain, Bacteria, these exporters range from 195-280 aas in size. Efflux is driven by proton antiport [4,5,6,7,8]. Two proteins in this family, LysE of *Corynebacterium glutamicum* and ArgO of *Escherichia coli* have been characterized. Through mutational studies, the wild-type LysE exporter of *C. glutamicum* has been shown to actively export both L-Lysine and D-Lysine [9]. Transcription of the *lysE* gene is activated by LysG, a LysR-type transcriptional regulator, in the presence of arginine, lysine, or histidine [10]. A mutant lacking *lysEG* of *C. glutamicum* accumulates both enantiomers of lysine to reach intracellular concentrations exceeding 230mM. These *lysEG* mutants undergo bacteriostasis when intracellular accumulation of L-Lysine exceeds concentrations of 250mM [9]. Similar mutational studies demonstrate that wild-type ArgO of *E. coli* actively exports L-arginine and canavanine, a plant-derived arginine analog and antimetabolite that competes with arginine for incorporation into nascent polypeptides during translation[14]. Null mutations in *argO* or the gene encoding its transcriptional activator, *argP*, diminish arginine efflux [11]. Lysine, however, suppresses expression of *argO*.

These previous studies suggested that these amino acid exporters may play roles in mediating the secretion of signaling molecules or avoiding cytoplasmic accumulation of the substrate compounds to toxic levels [2,11,12,13]. Both ArgO and LysE may act as "safety valves" to prevent the bacteriostatic accumulation of compounds following their uptake into the cells and the hydrolysis of nutrient Arg-containing and Lys-containing peptides [4,8,11]. ArgO can also export canavanine, a compound known to inhibit bacterial growth after competitive misincorporation into polypeptides in place of arginine. Because intracellular lysine also reduces expression of *argO*, ArgO may serve to maintain a balance of intracellular levels of these two basic amino acids Arg and Lys, for optimal growth [11].

1.3 Homoserine/Threonine Resistance Proteins (RhtB; TC# 2.A.76)

RhtB exporters are found in the domain, Bacteria, and vary between 180-250 aas in size. Characterized proteins in this family have been shown to actively export a wide variety of neutral amino acids and amino acid derivatives, including cysteine, O-acetylserine, azaserine, alanine, leucine, threonine, homoserine and homoserine lactone [15,16,17,18]. An RhtB member, PimT of *Streptomyces natalensis*, can export a quorum-sensing pimaricin-inducer PI-factor (2,3-diamino-2,3-bis(hydroxymethyl)-1,4-butanediol) [15]. One other RhtB member, MrsC of epiphytic strain *Pseudomonas syringae* pv. syringae 22d/93, has been shown to export 3-methylarginine and plays a role in growth inhibition of an antagonist strain, *Pseudomonas syringae* pv. glycinea [19].

Expression of the RhtB protein, EamB of *E. coli*, is thought to be controlled by a LysR-type regulator, YfiE [18]. Similar to the gene orientation of *lysE* and *lysG*, *eamB* and *yfiE* are adjacent, inversely oriented and separated by a short spacer segment. Efflux of cysteine and O-acetylserine via the EamB system requires an upregulated biosynthetic pathway and has also been suggested to act as a “safety valve,” pumping out cysteine and its precursor (and activator) when they attain toxic levels [18]. Also observed to export L-homoserine, L-serine, and L-homoserine lactone, PimT serves as an example of an amino acid exporter involved in secretion of a signaling molecule, PI-factor. High levels of extracellular PI-factor are sensed at the cell surface of *S. natalensis* and induce expression of biosynthetic genes for pimaricin, an antifungal agent [16]. Pimaricin, also known as natamycin, interacts with membrane sterols in fungal cells to inhibit vacuole fusion and cause leakage of cytoplasmic material [20]. Studies on the EamB, PimT and MrsC proteins demonstrate that RhtB proteins are amino acid exporters that are transcriptionally regulated in a manner to that of the LysE proteins structural genes, but they also function to promote growth inhibition of antagonistic organisms.

1.4 Cadmium Ion Resistance Proteins (CadD; TC# 2.A.77)

CadD proteins are limited to the domain, Bacteria, and range from 180-250 aas in size. Analyses of minimal inhibitory concentration assays suggest that the CadD protein of the *Staphylococcus aureus* plasmid pRW001 functions in the export of cadmium ions [21]. Additional studies using RT–PCR have demonstrated that *cadD* expression in *Streptococcus pneumoniae* increased by ~3.6-fold in the presence of 30 μM Cd^{2+} [22]. In *Staphylococcus lugdunensis*, *cadD* expression is activated by the transcriptional regulator, CadX, but in *Streptococcus salivarius,* CadX has been proposed to repress both *cadD* and *cadX* expression and lose binding affinity in the presence of Cd^{2+} to allow for expression [23].

Genes conferring resistance to Cd^{2+} and other heavy metal ions can be found co-localized on staphylococcal plasmids with multidrug resistance genes. Sewage sludge and phosphate fertilizers with high levels of cadmium have been used in agricultural soils and may have accumulated through feedstuffs for livestock. Cadmium intake in humans has been linked to the ingestion of animal-based foods and water [24]. These elevated cadmium levels may play a role in selecting for heavy metal resistance in bacteria including *S. aureus* commonly found on humans and other animals [24].

1.5 Ca^{2+}/H^+ antiporters-2 (CaCA2; TC# 2.A.106)

Members of the Ca^{2+}/H^+ antiporter Family, CaCA2, contain around 200-350 aas with 6 TMSs, typically in a 3+3 TMS arrangement, and are found in all three domains of life. Two functionally characterized members of this family, TMEM165 of *Homo sapiens* and Gdt1p of *Saccharomyces cerevisiae*, are localized in the Golgi apparatus and play roles in Ca^{2+} export driven by coupled H^+ influx [25,26]. One such member, TMEM165, is a gene involved in human congenital disorders of glycosylation (CDG), a family of inborn metabolic diseases affecting glycosylation pathways [27,28].

TMEM165 knock down using siRNA demonstrated a general decrease in the pH in acidic compartments in siRNA-targeted cells, confirming that TMEM165-deficiency affects late endosomal/lysosomal pH homeostasis. Mutational studies with Gdt1p of *S. cerevisiae* demonstrated that growth of the *gdt1* null mutants was not sensitive to the presence of a moderate Ca^{2+} concentration (50mM $CaCl_2$), but in a high Ca^{2+} concentration (750mM), growth of the *gdt1* null mutants was reduced compared with that of the isogenic wild type. A truncated version of the human ortholog, TMEM165, was expressed in *S. cerevisiae* and partially overcame Ca^{2+} sensitivity in *gdtp1* null mutants. These studies suggest that TMEM165 and Gdt1p function similarly in the antiport of Ca^{2+} for H^+ [28].

Maintenance of cytoplasmic Ca^{2+} and pH homeostasis in human cells is essential for organellar function. A mutation in the nonhomologous human Ca^{2+} transporter, Ca^{2+}-ATPase isoform 1 (SPCA1), causes Hailey-Hailey disease, with symptoms that include the increase in cytoplasmic Ca^{2+} levels and the decrease in luminal Golgi Ca^{2+} levels. Similar to deficiencies in TMEM165, this loss of Ca^{2+} homeostasis results in glycosylation (CDG) [28].

1.6 Mn^{2+} exporters (MntP; TC# 2.A.107)

Similar to previously established members of the LysE superfamily, members of the MntP family are characterized by a size of around 200 aas with 6 TMSs in a 3+3 TMS arrangement. So far, they are exclusively found in bacteria and archaea. A member of this family, MntP of *E. coli*, is known to export manganese ions [29,30]. Microarray analyses suggested that MntP is positively regulated by MntR and thus, is part of the

MntR regulon. YebN has been suggested to share significant sequence similarity with members of the LysE family efflux pumps [29].

Although manganese plays roles in enzymatic catalysis and protection against oxidative stress, excess manganese inhibits bacterial growth. Elevated manganese levels could affect the activities of enzymes dependence on iron and other metals [31,32].

1.7 Iron/Lead Transporters (ILT; TC# 2.A.108)

Radiolabeled iron transport assays and mutant complementation studies demonstrate that ILT family members are heavy metal ion uptake transporters specific for iron and/or lead. Topological analyses confirmed that most members of the ILT family have 7 conserved TMSs arranged in a 3+3+1 arrangement [33]. ILT protein sizes vary substantially due to the inclusion of large hydrophilic domains near the N-termini in many of these proteins. A majority of family members are found in bacteria and archaea, but some are also found in eukaryotes such as fungi. In *S. cerevisiae*, the high affinity iron permease Ftr1p and the ferroxidase Fet3p are required for assembly into a functional iron uptake complex. Protein interaction studies showed that Ftr1p and Fet3p act as a minimal heterodimer complex, where both proteins must be present in order to localize to the plasma membrane [34]. Iron permease, EfeU of *E. coli*, forms a high affinity iron uptake complex with EfeO and EfeB, all of which are repressed by transcriptional regulator CpxAR under high pH conditions. However, under acidic aerobic conditions, the Fur regulator derepresses CpxAR to promote transcriptional expression.

ILT proteins take up iron independently of siderophore transporters. In the plant pathogen, *Burkholderia cenocepacia*, lack of siderophore synthesis does not result in iron-limited growth inhibition. The iron uptake complex involving iron permease FtrC of *B. cenocepacia* compensates for the lack of siderophores. Mutants deficient in both FtrC and siderophore synthesis are unable to grow under conditions of iron starvation [35]. As described previously, expression of ILT proteins is derepressed by the Fur regulator in the presence of iron. The Fur regulator has been observed to repress siderophore synthesis and promote expression of pathogenic genes involved in the defense of reactive oxygen species produced by human immune cells. Thus, Fur-activated ILT proteins may be involved in an alternative pathogenic strategy to acquire iron [35,36].

1.8 Tellurium Ion Resistance Proteins (TerC; TC# 2.A.109)

Members of the TerC family are believed to function in tellurium ion resistance and response to cellular stress [37]. These proteins share a 7-TMS core with a 3+3+1 TMS arrangement and are typically found in bacteria and archaea, but are also found in some eukaryotes [38]. Sizes for these proteins range from 180 to 350 aas with as many as 9 TMSs.

The *ter* genes in *Clostridium acetobutylicum* promote resistance to methyl methanesulfonate (MMS), mitomycin C (MC), and UV when expressed in *recA* mutant strains of *E. coli* [39]. In *Yersinia pestis*, expression of TerD, a protein observed to complex with TerC, increases during intracellular growth, along with several other stress response-related genes, including superoxide dismutase-A [40]. In *Streptomyces coelicolor*, loss of TerD resulted in altered differentiation and spore morphology and reduced tellurite resistance [41]. In *Arabidopsis thaliana* cells, the TerC protein (AtTerC) is essential for the maturation of thylakoid stacks in the chloroplast. AtTerC mutants lack thylakoid and display globular structures of varying sizes [42]. In increasing concentrations of potassium tellurite, tellurium resistance determinants promote formation of crystalline tellurium structures in outer membrane vesicles of *Pseudomonas putida* BS228 and *Pseudomonas aeruginosa* ML4262. These crystalline tellurium structures are implicated in resistance to pore-forming colicins [43]. These cases highlight the physiological roles of TerC in membrane morphology, tellurite resistance and other general stress responses.

1.9 The Neutral Amino Acid Transporter Family (NAAT; TC# 2.A.95)

NAAT family proteins are exclusively found in bacteria and archaea. The majority of these proteins have sizes between 190-280 aas with 6 predicted TMSs in a 3+3 TMS arrangement. The best characterized member of the NAAT family, SnatA of hyperthermophilic archaeon *Thermococcus* sp. KS-1, is involved in the uptake of neutral amino acids, glycine and alanine [44]. Several homologues have been annotated as multiple drug resistance proteins. However, a recent study provided evidence that disagrees with this functional assignment [45].

1.10 The Nickel/Cobalt Transporter Family (NicO; TC# 2.A.113)

NicO proteins range from 270-430 aas in size. Through transposon mutagenesis, NicO protein, RcnA of *E. coli*, has been shown to play a role in Ni^{2+} and Co^{2+} efflux [46]. The expression of *rcnA* is expressed when these two metal ions are present. Members of this family are found across all three domains of life. NicO exporters are not related to the nickel cobalt transporter family (NiCoT), which is a family of nickel uptake permeases. RcnA lacks the NiCoT signature present in the second transmembrane helix of these eight-helix permeases [47].

In gammaproteobacteria, nickel and cobalt are essential nutrients but are toxic at high cytoplasmic concentrations. Ni^{2+} and Co^{2+} toxicity in *Pseudomonas putida* results in the accumulation of oxidative stress response proteins [48]. The physiological role of NicO proteins is likely cellular detoxification of nickel or cobalt.

1.11 The Peptidoglycolipid Addressing Protein Family (GAP; TC# 2.A.116)

GAP family proteins are typically found in bacteria and are prominent in members of the mycobacterial genus. The majority of these proteins have sizes between 180-290 aas with 6 predicted TMSs in a 3+3 TMS orientation. The best characterized member of the GAP family, Gap (Q3L890) of *Mycobacterium smegmatis*, has been reported to play a role in biogenesis of the mycobacterial cell envelope via the transport of peptidoglycolipids (glycopeptidolipids; GPLs) to the surface of the cell [49]. This protein is not, however, required to synthesize GPLs. The GPLs produced by a mutant *gap* strain of *M. smegmatis* were retained in the cytoplasmic compartment of the cell. In the complemented strain, the surface location of the GPLs was restored to resemble the wild-type strain. Mass spectrometry demonstrated that the GPLs produced by the mutant, complemented mutant, and wild-type strains were chemically identically, suggesting that Gap does not play any role in GPL modification or biosynthesis [50]. Little is known about the mode of action and energy source for transport. Interestingly, lack of *gap* expression in *M. smegmatis* abolishes sliding motility, suggesting a role of proper cell envelope assembly or motility. Complemented and wild-type strains were able to slide.

GPLs are functionally important due to their roles in inhibition of the blastogenic response of splenic lymphocytes to non-specific mitogens [51], decreasing the oxidative phosphorylation efficiency of mitochondria without modifying active respiration (8003470), alteration of biological membranes via lipid-lipid interactions [52], inhibition of phagocytosis by human macrophages [53] or modulation of TNF-α synthesis in murine macrophages [54]. Surface-localized GPLs are crucial for sliding motility in *M. smegmatis* as noted above, but are also associated with phenotypes biofilm development in *M. smegmatis* and drug resistance in *Mycobacterium avium* [54]. As a result, Gap proteins could represent novel drug targets.

1.12 The Disulfide Bond Oxidoreductase D Family (DsbD; TC# 5.A.1)

The DsbD Family is a large family of transmembrane electron carriers that is represented in all domains of life. Several functional roles have been reported for these proteins: (i) thiol-disulfide exchange, (ii) cytochrome c biogenesis, (iii) methylamine utilization, (iv) mercury resistance, (v) copper resistance, and (vi) various additional reductase functions. Previous studies demonstrated that DsbD of *E. coli* arose from intragenic gene duplication of a 3-TMS element [55].

In this paper, we report investigations allowing expansion of the LysE superfamily to include members from all three domains of life. Using computational methods, we demonstrate that the previously established members of this superfamily are homologous to members of the eight additional families described above: (i) tellurium ion resistance proteins (TerC); (ii) iron/lead transporters (ILT); (iii) Mn^{2+} exporters (MntP); (iv) Ca^{2+}/H^{+} antiporters-2 (CaCA2); (v) Ni^{2+}/Co^{2+} transporters (NicO); (vi) neutral amino acid transporters (NAAT); (vii) peptidoglycolipid addressing proteins (GAP); and (viii) disulfide bond oxidoreductase D proteins (DsbD). We confirm this expansion and provide superfamily descriptions with thorough analyses of identified internal repeats and conserved motifs, multiple alignments of identified homologues, phylogenetic trees and average hydropathy, amphipathicity and similarity plots. The superfamily phylogenetic tree shows the relationships of these eleven families to each other [54].

MATERIALS AND METHODS

2.1 Potential New Families

Previously established members of the LysE superfamily were initially examined in the Transporter Classification Database (TCDB; www.tcdb.org) [56]. PSI-BLAST searches with iterations against TCDB (TC-BLAST) were conducted to locate distant homologues with overlapping TMSs. The Web-based Hydropathy, Amphipathicity & Topology (WHAT) program was used to generate hydropathy plots for preliminary topological predictions of individual proteins [57]. Established families within the LysE superfamily are listed in Table 1 with previously assigned transporter classification numbers (TC#) from TCDB.

Table 1. Characteristics of all families in the LysE superfamily included in this study

Family Name	Family Abbreviation	Transporter Classification No. (TC) #	Relative Family size[a]	Average Protein Size[b]	# TMSs[c]	# Subfamilies in TCDB[d]	Established Substrates (S)	Polarity of transport	Taxonomic Distribution
L-Lysine Exporter	LysE	2.A.75	1799	204 ± 20	6	1	D- and L-lysine, histidine and arginine	in --> out	Bacteria
Resistance to Homoserine/Threonine	RhtB	2.A.76	2711	207 ± 14	5, 6	2	O-aetylserine/cysteine/azaserine, threonine, serine, homoserine, homoserine lactones, leucine, alanine, 3-methyarginine and pimaricin-inducer PI-factor	in --> out	Bacteria
Cadmium Resistance	CadD	2.A.77	578	210 ± 68	4, 5, 6, 7	1	cadmium ions	in --> out	Bacteria
Neutral Amino Acid Transporter	NAAT	2.A.95	588	207 ± 17	6	1	glycine, L-alanine, L-serine, L-threonine and a variety of neutral L-amino acids	in --> out	Bacteria, Archaea
Ca^{2+}:H^+ Antiporter-2	CaCA2	2.A.106	1852	252 ± 106	5, 6, 7	4	calcium ions	cytoplasm --> golgi lumen	Bacteria, Archaea, Eukaryota
Mn^{2+} exporter	MntP	2.A.107	298	188 ± 14	6	3	manganese ions	in --> out	Bacteria, Archaea
Iron/Lead Transporter	ILT	2.A.108	1063	350 ± 128	6, 7, 8	3	iron and lead ions	out --> in	Bacteria, Archaea
Tellurium Ion Resistance	TerC	2.A.109	2592	328 ± 41	6, 7, 8, 9	3	tellurium ions	in --> out	Bacteria, Archaea, Eukaryota
Nickel/cobalt Transporter	NicO	2.A.113	539	345 ± 111	5, 6, 7	2	nickel and cobalt ions	in --> out	Bacteria, Archaea, Eukaryota
Peptidoglycolipid Addressing Protein	GAP	2.A.116	113	233 ± 41	6	3	peptidoglycolipids	in --> out	Bacteria, Archaea
Disulfide Bond Oxidoreductase D	DsbD	5.A.1	1981	533 ± 189	6, 8, 9	6	electrons	cytoplasm --> periplasm	Bacteria, Archaea, Eukaryota

[a] A single search with the first protein in TCDB (x.x.x.1.1) was used as the query sequence to BLAST the NCBI protein database with a 95% cutoff. The BLAST searches were run on July 22, 2013.

[b] Average number of amino acyl residues in the proteins retrieved by Protocol1 for column 4.

[c] Dominant numbers of predicted TMSs for the proteins retrieved by Protocol1 for column 4.

[d] Number of subfamilies currently included in TCDB.

2.2 Obtaining Homologues

A single FASTA-formatted protein sequence was selected from TCDB and used as the input for Protocol1, a program available through the BioV Suite software [58]. With Protocol1, we utilize NCBI PSI-BLAST with a threshold of 0.80 to generate a list of non-redundant homologues. This setting ensured that only one of any set of proteins with greater than 80% identity would be retained [59]. Protocol1 was applied to proteins of each family in the study.

2.3 Establishing Homology between Families

The FASTA-formatted homologue sequences generated with Protocol1 were used as input into another BioV Suite program, Protocol2. Protocol2 requires two such input files and generates a graphical report, displaying sequence alignments between homologous members of two different protein families [58]. Two sequences with strong TMS alignment and z-scores above the value of 13.0 standard deviations (S.D.) are considered sufficient to provide strong evidence of homology. The higher the z-score, the greater the sequence similarity [58]. The z-scores obtained with Protocol2 were then verified through the use of a TCDB web program, Global Sequence Alignment Tool (GSAT) [58]. Good scoring pairs of sequences identified with Protocol2 were then tested using 20,000 random shuffles (GSAT) for more accurate results. Once verified, the GSAT results were analyzed for TMS overlap through use of the TMS prediction program, HMMTOP [60]. The top comparison scores and number of aligned TMSs between each family are shown in Table 2. Finally, a GSAT comparison score, based on 2,000 random shuffles, was generated between sequences of query proteins and respective proteins obtained from Protocol1 to manually check for homology of A versus B and C versus D (Table 3) [61,62]. Specific proteins identified in this paper are reported with UniProt accession numbers (www.uniprot.org). Proteins lacking UniProt accession numbers are assigned NCBI (GenBank) accession numbers.

Table 2: Comparison scores between LysE superfamily members**.** Scores equal to or greater than 13.0 Standard Deviations (S.D.) are bolded. The number of aligned TMSs is included below each score. Comparisons with the negative control, the Mitochondrial Carrier (MC) family, are provided to the right of the bolded border.

	LysE	RhtB	CadD	CaCA2	MntP	ILT	TerC	NAAT	NicO	GAP	DsbD	MC
LysE		**20.1** **(5TMSs)**	12.1 S.D. (4TMSs)	**13.5 S.D.** **(3TMSs)**	11.8 S.D. (3TMSs)	12.5 S.D. (2TMSs)	**14.6 S.D.** **(3TMSs)**	**14.0 S.D.** **(5TMSs)**	10.8 S.D. (6TMSs)	12.7 S.D. (3TMSs)	12.3 S.D. (5TMSs)	4.1 S.D. (0TMSs)
RhtB			11.9 S.D. (3TMSs)	**13.0 S.D.** **(4TMSs)**	**13.7 S.D.** **(3TMSs)**	**13.7 S.D.** **(3TMSs)**	**13.5 S.D.** **(3TMSs)**	**15.0 S.D.** **(5TMSs)**	**13.8 S.D.** **(6TMSs)**	**14.5 S.D.** **(5TMSs)**	**14.0 S.D.** **(5TMSs)**	8.8 S.D. (2TMSs)
CadD				**14.2 S.D.** **(3TMSs)**	**15.7 S.D.** **(4TMSs)**	**13.5 S.D.** **(6TMSs)**	**13.6 S.D.** **(4TMSs)**	**14.4 S.D.** **(5TMSs)**	**15.1 S.D.** **(6TMSs)**	12.3 S.D. (5TMSs)	11.5 S.D. (6TMSs)	8.5 S.D. (2TMSs)
CaCA2					**15.1 S.D.** **(3TMSs)**	**15.3 S.D.** **(3TMSs)**	**16.2 S.D.** **(3TMSs)**	12.0 S.D. (5TMSs)	12.5 S.D. (5TMSs)	11.6 S.D. (5TMSs)	**13.2 S.D.** **(5TMSs)**	10.5 S.D. (1TMS)
MntP						12.5 S.D. (6TMSs)	**13.5 S.D.** **(5TMSs)**	**15.1 S.D.** **(4TMSs)**	12.3 S.D. (5TMSs)	11.3 S.D. (4TMSs)	**16.0 S.D.** **(5TMSs)**	9.1 S.D. (2TMSs)
ILT							**13.1 S.D.** **(5TMSs)**	11.8 S.D. (6TMSs)	12.8 S.D. (6TMSs)	12.8 S.D. (6TMSs)	10.9 S.D. (4TMSs)	9.1 S.D. (1TMS)
TerC								**15.2 S.D.** **(3TMSs)**	**13.9 S.D.** **(5TMSs)**	12.1 S.D. (5TMSs)	12.9 S.D. (5TMSs)	4.4 S.D. (0TMSs)
NAAT									**13.5 S.D.** **(3TMSs)**	12.8 S.D. (4TMSs)	**15.3 S.D.** **(6TMSs)**	10.0 S.D. (1TMS)
NicO										12.7 S.D. (5TMSs)	**14.8 S.D.** **(5TMSs)**	9.3 S.D. (1TMS)
GAP											**13.1 S.D.** **(5TMSs)**	5.8 S.D. (2TMS)
DsbD												9.9 S.D. (1TMS)

Table 3: Use of the Superfamily Principle (transitivity rule) to establish homology: If A and B are homologous, B and C are homologous, and C and D are homologous, then A is homologous to D. Families being compared are presented in column 1. Uniprot IDs are provided in columns 2-5. When a Uniprot accession number is unavailable, an NCBI accession number is provided. Comparison scores, expressed in standard deviations (S.D.), are provided in columns 6-9. Columns 6-8 allow establishment of homology. Column 9 gives the value determined when A is compared to D directly. For example, in a comparison between LysE and RhtB, Protein A and Protein D are query proteins from each respective family. Protein B is a homologue of Protein A. Protein C is a homologue of Protein D. Comparisons with the negative control, the Mitochondrial Carrier (MC) family, are provided below the double-lined border.

	Proteins Compared (Accession numbers provided)				Score for each comparison (S.D.)			
Families Compared	**Protein A**	**Protein B**	**Protein C**	**Protein D**	**A v B**	**B v C**	**C v D**	**A v D[a]**
LysEvRhtB	P94633	H3RH39	Q2SUV5	P76249	32.5	20.1	52.0	9.0
LysEvCadD	P64711	K0HW07	K9TWQ5	Q45153	37.0	12.1[a]	36.1	0.7
RhtBvCadD	P76249	G9Y0F1	G9WHF3	O05469	72.0	11.9[a]	36.0	1.1
LysE v CaCA2	P94633	E0MXD6	C1MR94	P52876	63.0	13.5	31.7	1.6
RhtB v CaCA2	P76249	G9Y0F1	K9ULS7	P52876	73.0	13.0	62.4	1.3
CadD v CaCA2	O05469	L2SR21	B7FUM2	P52876	50.7	14.2	57.2	2.0
RhtB v MntP	P76249	C4GM93	D9SW99	O27840	45.9	13.7	37.5	1.9
CadD v MntP	O05469	H3NKZ1	Q727E5	O27840	48.0	15.7	34.3	1.0
CaCA2 v MntP	P52876	E0UDP4	C0DV56	P76264	74.5	15.1	57.3	1.3
RhtB v ILT	P0AG34	A1RAR9	Q2NBF8	Q58AJ4	50.5	13.7	125.9	0.4
CadD v ILT	O05469	C2D135	G5JVH6	Q5HSD5	43.1	13.5	41.0	4.2
CaCA2 v ILT	P52876	F0Y333	Q97V64	Q4J7V8	52.7	15.3	67.2	5.3
LysE v TerC	P94633	D7GFT1	Q20ZD5	I3XAB3	40.8	14.6	72.7	-0.2
RhtB v TerC	P76249	K8W4X6	WP_010022951	B5UIP4	63.3	13.5	54.9	1.4
CadD v TerC	O05469	WP_010652183	G8LRD3	B5UIP4	46.0	13.6	38.5	3.9
CaCA2 v TerC	P52876	B7FUM2	D7V5X7	B5UIP4	57.2	16.2	62.9	1.3
MntP v TerC	P76264	E7S0L5	A2TWJ9	Q7UHX7	43.9	13.5	40.3	2.6
ILT v TerC	Q58AJ4	G6EJJ4	Q8KAT3	B5UIP4	125.3	13.1	37.6	0.7
LysE v NAAT	P11667	G8QX72	Q2C9W5	O32244	35.1	14.0	40.6	3.9
RhtB v NAAT	P0AG38	L7BNM7	H1S8A2	Q8J305	95.4	15.0	39.2	5.2
CadD v NAAT	Q45153	K6U069	E3T754	Q8J305	27.1	14.4	40.4	-0.1
MntP v NAAT	O27840	A6VQU4	WP_018748573	P67143	20.7	15.1	46.8	2.6
TerC v NAAT	I3XAB3	Q5L1S7	T2GCR6	P67143	26.2	15.2	45.5	3.0
RhtB v NicO	P0AG38	N9DHM2	G2TLK3	F8C138	68.9	13.8	34.5	1.2
CadD v NicO	Q45153	K9ZC80	K6XDF4	F8C138	24.8	15.1	22.4	0.2
TerC v NicO	I3XAB3	F4QZA6	M1YUV4	F8C138	55.7	13.9	32.8	1.4
NAAT v NicO	Q8J305	H1L1H6	WP_022692950	P76425	38.4	13.5	34.9	0.8
RhtB v GAP	P76249	F3KVR3	WP_019358971	K6W6C5	45.2	14.5	16.6	1.7
RhtB v DsbD	P0AG38	M4RA58	R1CD96	P45706	35.6	14.0	43.5	-0.2
CaCA2 v DsbD	B9MIH1	D1JG69	F9DXY9	P45706	23.2	13.2	77.7	-0.5
MntP v DsbD	E4RIT5	F7ZP38	F5SD76	P45706	28.2	16.0	70.7	0.6
NAAT v DsbD	Q8J305	Q8U2T5	K0NNX9	P45706	82.4	15.3	41.9	2.5
NicO v DsbD	B2JAZ6	K9Z039	M1ZHA3	P45706	34.2	14.8	43.2	0.2
GAP v DsbD	K6W6C5	WP_018161757	C6D6Q6	Q939U6	31.7	13.1	41.8	1.0
LysE v MC	P94633	G8QX72	XP_395934	P12235	35.7	4.1[a]	162.4	0.7
RhtB v MC	P76249	F3KVR3	I3WBB4	P12235	43.0	8.8[a]	157.0	1.0
CadD v MC	O05469	D2AZ49	XP_003796317	P12235	30.8	8.5[a]	200.7	1.6
CaCA2 v MC	G0PPC8	L7L942	Q4PMB2	P12235	17.5	10.5[a]	158.1	0.7
MntP v MC	O27840	L7VM13	S7NPK9	P12235	35.2	9.1[a]	153.6	-1.0
ILT v MC	Q5HSD5	L0W8N6	V9KQ68	P12235	48.2	9.1[a]	149.5	-1.4
TerC v MC	I3XAB3	K9CUK2	Q91336	P12235	48.9	4.4[a]	172.4	0.4
NAAT v MC	Q8J305	F9RL32	Q91336	P12235	42.7	10.0[a]	176.1	0.6
NicO v MC	F8C138	G9QNI4	S9XZZ3	P12235	33.1	9.3[a]	171.4	-0.3
GAP v MC	K6W6C5	WP_019971730	V9KQ68	P12235	10.1	5.8[a]	155.4	-0.6
DsbD v MC	P45706	B3E4Q5	XP_007059219	P12235	48.4	9.9[a]	159.0	-0.8

[a] These comparison scores are insufficient to establish homology.

2.4 Viewing Average Hydropathy, Amphipathicity and Similarity Plots

Multiple alignments for each family in the study were generated using the ClustalX, Mafft and ProbCons programs [63,64,65]. The topologies of these sequences were then examined using AveHAS, a web-based program that displays the average hydropathy, amphipathicity and similarity plots for a set of homologues [66].

2.5 Identifying Internal Repeats

The multiple alignment file produced from ClustalX was used as the input for IntraCompare, a program for the detection of internal repeats. Generated AveHAS plots for respective multiple alignment files were referenced to locate comparable regions of interest. IntraCompare generates comparison scores expressed in S.D. for non-overlapping regions of the same homologous proteins [67].

2.6 Motif Analyses

Motif analyses were carried out using the MEME program (The MEME Suite; http://meme.nbcr.net/meme/) [68]. Default settings were used to search for ungapped, conserved residues within a given set of homologues. Results from HMMTOP were used to predict relationships between conserved regions relative to the TMSs. Motifs identified for each family were then paired to different families to observe similar residue conservation.

2.7 Construction of Phylogenetic Trees

Phylogenetic trees were derived using multiple programs. RAxML and FastTree methods have been explored using raxmlgui [69]. Phylip-formatted multiple alignments generated using ClustalX, Mafft and Probcons were used as inputs to generate FastTree trees for each protein family in this study. In addition, a Phylip-formatted multiple alignment of members from all eleven families was generated from Mafft and used to create a set of 100 trees using the RAxML method of analysis [70]. The Mafft alignment used for the RAxML tree analysis was generated using the Mafft-homologs function with 200 homologs retrieved per input sequence at a threshold of $1e^{-20}$ [64]. All FastTree trees and the best tree indicated by the RAxML method were viewed using FigTree. SuperfamilyTree (SFT) [71,72,73,74,75,76,77,78] and TreeView [79] were also utilized. Agreement between 100 trees was evaluated. FASTA-formatted sequences corresponding to the TC families were inputted and used to compile tens of thousands of NCBI BLAST bit-scores upon which SFT trees were based. SFT and Fitch programs then generated a default of 100 superfamily trees based on the results. These 100 trees were used to create a consensus tree [71,72,73,74,75,76,77,78]. The parameters for these programs are described in Supplemental Figure 1.

Results

In addition to the three previously established LysE superfamily members (Table 1), eight families were analyzed in this study: (i) CaCA2 (TC# 2.A.106); (ii) MntP (TC# 2.A.107); (iii) ILT (TC# 2.A.108); (iv) TerC (TC# 2.A.109); (v) NAAT (TC# 2.A.95); (vi) NicO (TC# 2.A.113); (vii) GAP (TC# 2.A.116) and (viii) DsbD (TC# 5.A.1) (Table 1). Mitochondrial carriers (TC# 2.A.29) were used as a negative control when generating comparison scores expressed in standard deviations (S.D.) using the GSAT program [58]. Like most members of the LysE superfamily, MC proteins have 6 TMSs but evolved via a different pathway [80]. They arose by triplication of a 2TMS-encoding genetic element, while LysE superfamily proteins arose by intragenic duplication of a 3TMS-encoding genetic element. Of the eight novel families, seven are included in the 2.A subclass of TCDB, secondary carrier-type facilitators known to catalyze symport, uniport and antiport. The exception, DsbD, is a family of transmembrane 2-electron transfer carriers with TC #5.A.1 [55,56,81].

Statistical evidence (Table 2) argued that the TerC, ILT, MntP, CaCA2, NAAT, NicO, GAP and DsbD families are related to the LysE, RhtB and CadD families. Multiple alignments additionally revealed that six TMSs align across all families included in this study. Statistical evidence for homology, multiple alignments of homologues, AveHAS plots, identified internal repeats, MEME/MAST diagrams of conserved motifs, and a proposed evolutionary pathway (evolutionary history) for this expanded superfamily are presented (Figures 1-4, Supplemental Figures 2-27, Tables 1-5). In addition, our results confirm topological findings reported in previous studies regarding LysE, RhtB, CadD, MntP, ILT, CaCA2, NAAT and DsbD homologues [1,25,26,29,33,44,55].

3.1 Controls

3.1.1 The Mitochondrial Carrier Family and the LysE superfamily

Members of the MC family have been shown to transport keto acids, amino acids, nucleotides, inorganic ions and co-factors across the membranes of mitochondria and other eukaryotic organelles [82,83]. Crystal structures for MC proteins have been elucidated, and these 6-TMS proteins were shown to have arisen via a 2-TMS triplication [80,84,85]. Members of the LysE superfamily, however, are predicted to have arisen via a 3-TMS duplication. Because of the differences in these two evolutionary pathways, MC proteins have been selected as a negative control to establish the highest possible comparison score that can be obtained by chance using non-homologous members of two unrelated superfamilies (Tables 2-3).

The best comparison score between 3-TMS segments of the MC and LysE superfamily members was 10.5 S.D. This score was obtained between proteins of the MC family and the CaCA2 family. The average score for the five best comparisons between LysE superfamily members and the MC family was 9.8 S.D. Although at least 3 TMSs of members of these two superfamilies were included in each alignment, the TMS alignments were poor (Supplemental Figures S16J and S16K). TMS overlap in the alignments is present in Table 2. In contrast, the average score for all of the best comparisons for the eleven LysE superfamily families with each other (Table 3) is 13.5

S.D, and corresponding TMSs were strongly aligned. Based on these results, we suggest that three conditions are sufficient to provide strong evidence for homology: (1) a standard comparison score of at least 13.0 S.D.; (2) proper alignment of at least 3 TMSs and (3) a unified evolutionary pathway for all superfamily members (Figure 1). These criteria were satisfied for all eleven members of the LysE superfamily.

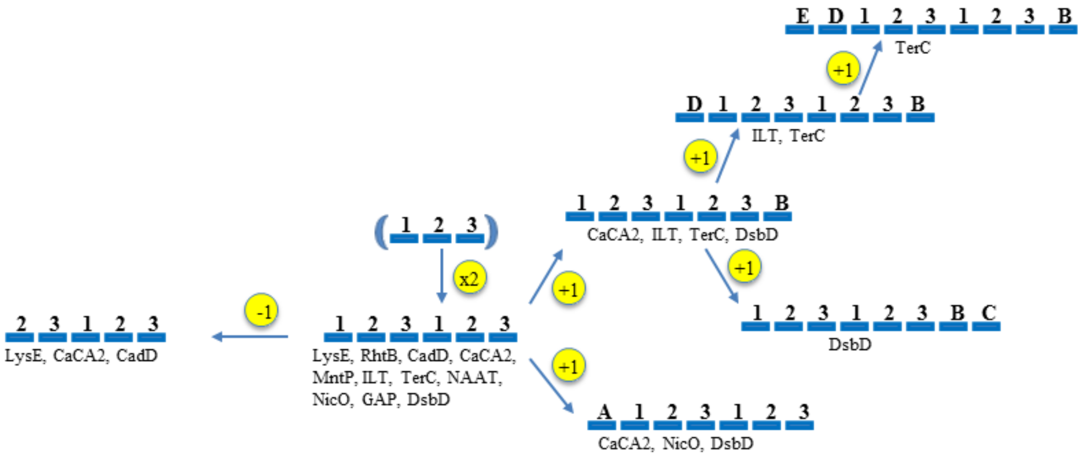


Figure 1. Proposed evolutionary history for the appearance of the eleven recognized families in the LysE superfamily. Protein topologies are indicated with bars representing TMSs and numbers indicating the positions of the TMSs in the proposed TMS primordial protein (in parentheses). Families are indicated by their standard abbreviations while numbers indicate "extra" TMSs outside of their basic 6-TMS unit, resulting from intragenic duplication of the primordial 3TMS precursor. A family abbreviation with a particular topology indicates that at least some members of the family are believed to have this topology.

3.2 Establishing Homology

3.2.1 The L-Lysine and L-Arginine Exporters (LysE; TC# 2.A.75); Homoserine/Threonine Resistance Proteins (RhtB; TC# 2.A.76); Cadmium Ion Resistance Proteins (CadD; TC# 2.A.77)

Previously published studies have shown that LysE, RhtB and CadD are distantly related [1]. We support this conclusion with additional statistical analyses (Supplemental Figures 2A-2C). Six TMSs are predicted for each of the homologues analyzed in this section. The top pair-wise analysis of RhtB and LysE homologues, Pst1 (H3RH39) v Bth1 (Q2SUV5)**,** demonstrated a comparison score of 20.1 S.D. The first five of six TMSs for each of these two proteins aligned (Supplemental Figure 2A). A score of 32.5 S.D. resulted when comparing the full sequences of Pst1 with the LysE protein, TC# 2.A.75.1.1 (P94633). In addition, a score of 52.0 S.D. was obtained when comparing the full sequences of Bth1 with RhtB protein, TC# 2.A.76.1.5 (P76249). These comparison scores satisfy our statistical standards for homology, and thus, we apply the superfamily principle to confirm that these two families are related (Table 3).

TMSs 2-4 of Oki1 (G9WHF3), a CadD homologue, aligned with TMSs 2-4 of the RhtB homologue Hal1 (G9Y0F1) with a comparison score of 11.9 S.D (Supplemental Figure 2B). A comparison score of 12.1 S.D. (Supplemental Figure 2C.) resulted from alignment of TMSs 2-5 of the CadD homologue Cth1 (K9TWQ5) with TMSs 2-5 of the LysE homologue Asp2 (K0HW07). The relationships between CadD proteins and LysE and RhtB proteins are not apparent based on our statistical standards for sequence similarity. Additional evidence will be discussed to expand upon these relationships and establish homology.

3.2.2 Ca^{2+}/H^+ antiporters-2 (CaCA2; TC# 2.A.106)

CaCA2 proteins display significant sequence similarity with 6-TMS CadD, LysE, and RhtB homologues (Supplemental Figure 3A-3C). TMSs 1-3 of the CaCA2 homologue Mpu4 (C1MR94) and the LysE homologue Cac2 (E0MXD6) were compared, yielding a score of 13.5 S.D. A score of 31.7 S.D. occurred when comparing the full sequences of Mpu4 and the CaCA2 protein, TC# 2.A.106.1.1 (P52876). In addition, a score of 63.0 S.D. resulted when comparing the full sequences of Cac2 with LysE, TC# 2.A.75.1.1 (P94633). Therefore these two families are homologous.

Particularly strong evidence was obtained from a comparison between CaCA2 and CadD proteins. TMSs 1-3 of the cadmium resistance protein Efa1 (L2SR21) aligned with TMSs 1-3 of the CaCA2 homologue Ptr2 (B7FUM2) to give a comparison score of 14.2 S.D (Supplemental Figure 3A). A score of 57.2 S.D. resulted when comparing the full sequence of Ptr2 with that of the CaCA2 protein, TC# 2.A.106.1.1 (P52876). In addition, a comparison of the full-length sequences of Efa1 and CadD TC# 2.A.77.1.1 (O05469) yielded a score of 50.7 S.D. Because the CaCA2 family is homologous to CadD, LysE and RhtB family members, we conclude that CaCA2 and CadD are members of the LysE superfamily. Comparison scores between the CaCA2 family and the MntP, ILT, TerC and DsbD families were also 13.0 S.D or greater (Tables 2 and 3).

3.2.3 Mn^{2+} exporters (MntP; TC# 2.A.107)

6-TMS MntP proteins share sufficient sequence similarity with RhtB, CadD and CaCA2 family members to establish homology (Tables 2 and 3, Supplemental Figures 4A-4C). A comparison between the MntP homologue Dvu1 (Q727E5) and the cadmium resistance protein Hku1 (H3NKZ1) displayed an alignment of TMSs 3-6 in both proteins with a score of 15.7 S.D (Supplemental Figure 4B). A score of 34.3 S.D. was obtained when comparing the full sequences of Dvu1 with MntP protein, TC# 2.A.107.1.2 (O27840), and a score of 48.0 S.D. resulted when comparing the full sequences of Hku1 with the CadD protein, TC# 2.A.77.1.1 (O05469). Although significant scores were not observed with LysE homologues, relationships between RhtB, CadD and CaCA2 families have been established, providing sufficient evidence for the inclusion of MntP as a member of the LysE superfamily. Comparison scores between MntP and TerC, NAAT and DsbD family members were also 13.0 S.D or greater (Tables 2 and 3).

3.2.4 Iron/Lead Transporters (ILT; TC# 2.A.108)

ILT proteins demonstrate significant sequence similarity with proteins of CadD, RhtB and CaCA2 families (Supplemental Figures 5A-5C). The 6-TMS cadmium resistance homologue Lbr1 (C2D135) and the 8-TMS ILT homologue Sma2 (G5JVH6) were compared. All of the six TMSs in Lbr1 aligned with TMSs 2-7 of Sma2 with a comparison score of 13.5 S.D (Supplemental Figure 5A). Investigating further with HMMTOP and a WHAT hydropathy plot, we observed that the 8-TMS Sma2 contains the core 3+3+1 arrangement near its C-terminus with a lone TMS at the N-terminus. From these depictions, we note that the 6-TMS Lbr1 protein aligns within the 3+3 region of the 8-TMS Sma2 protein. A score of 41.0 S.D. was obtained when comparing the full sequences of Sma2 with ILT protein, TC# 2.A.108.2.4 (Q5HSD5). In addition, comparing the full length sequences of Lbr1 and CadD TC# 2.A.77.1.1 (O05469), yielded a score of 43.1 S.D., establishing homology between these two families. Additional studies comparing TMSs 1-3 of the 6-TMS RhtB homologue Aau1 (A1RAR9) and TMSs 2-4 of the ILT homologue Eli1 (Q2NBF8) demonstrated a 3-TMS alignment with a score of 13.7 S.D (Supplemental Figure 5B). Eli1 is predicted to have 7 TMSs, but HMMTOP and WHAT did not recognize a strongly hydrophobic region between predicted TMS#1 and TMS#2 as a transmembrane segment, thus suggesting that this protein has 8 TMSs. Finally, we compared TMSs 1-3 of the ILT homologue Sso1 (Q97V64) with TMSs 1-3 of the CaCA2 homologue Aan1 (F0Y333). This comparison yielded a score of 15.3 S.D (Supplemental Figure 5C). A score of 67.2 S.D. resulted when comparing the full sequences of Sso1 and ILT protein, TC# 2.A.108.3.3 (Q4J7V8). In addition, a score of 52.7 S.D. was obtained when comparing the full sequences of Aan1 and CaCA2 protein, TC# 2.A.106.1.1 (P52876). With this statistical evidence, we conclude that ILT is an additional member to the LysE superfamily. A comparison between ILT and TerC proteins also yielded high comparison scores (Tables 1 and 2).

3.2.5 Tellurium Ion Resistance Proteins (TerC; TC# 2.A.109)

TerC members show significant sequence similarities with homologues from a large number of the different families (Supplemental Figures 6A-6F). Of the TerC comparisons, the highest score was observed between TerC and CaCA2 family members (Supplemental Figures 6F). TMSs 1-3 of the 7-TMS TerC protein Lga1 (D7V5X7) and TMSs 1-3 of the 6-TMS CaCA2 protein Ptr2 (B7FUM2) aligned and yielded a score of

16.2 S.D. A score of 62.9 S.D. resulted when comparing the full sequences of Lga1 and TerC protein, TC# 2.A.109.1.3 (B5UIP4). Furthermore, a score of 57.2 S.D. was obtained when comparing the full sequences of Ptr2 and CaCA2 protein, TC# 2.A.106.1.1 (P52876). In addition, TerC proteins yielded significant comparison scores with 8 of the 10 other families shown in Table 2. These relationships provide further evidence for the inclusion of the TerC families in the LysE superfamily.

3.2.6 Neutral Amino Acid Transporter Family (NAAT; TC# 2.A.95)

Significant comparison scores with NAAT proteins were seen between LysE, RhtB, CadD, MntP, and TerC family proteins (Supplemental Figures 7A-7E). The best example of homology is seen with the comparison of TMSs 1-5 of the RhtB homologue Pag1 (L7BNM7) and the NAAT homologue Cba1 (H1S8A2), which yielded a score of 15.0 S.D (Supplemental Figure 7B). When comparing the full length sequences of Cba1 and NAAT protein, TC# 2.A.95.1.4 (Q8J305), a score of 39.2 S.D. was obtained. Comparing the full sequences of Pag1 and RhtB protein, TC# 2.A.76.1.2 (P0AG38), gave a score of 95.4 S.D., thus establishing homology between these two families. In addition to the relationships with members of the LysE, RhtB, CadD, MntP and TerC families, relationships with NicO and DsbD family members were apparent, providing sufficient evidence for the inclusion of NAAT as a member of the LysE superfamily.

3.2.7 Nickel/Cobalt Transporter Family (NicO; TC# 2.A.113)

Here we report significant comparison scores with RhtB, CadD, TerC and NAAT family proteins (Supplemental Figures 8A-8D). Comparing TMSs 1-6 of the CadD homologue Acy3 (K9ZC80) with the NicO homologue Gar1 (K6XDF4) yielded a score of 15.1 S.D (Supplemental Figure 8B). In this comparison, every TMS aligned correspondingly in the two sequences. A score of 22.4 S.D. resulted when the full sequence of Gar1 was compared with that of the NicO protein, TC# 2.A.113.1.9 (F8C138), and a score of 24.8 S.D. was obtained when comparing the full sequence of Acy3 with an established CadD protein, TC# 2.A.77.1.2 (Q45153). These results provided strong evidence that NicO is homologous to the previously discussed families and support further expansion of the LysE superfamily. A significant comparison score between NicO and DsbD was also noted.

3.2.8 Peptidoglycolipid Addressing Protein Family (GAP; TC# 2.A.116)

Although the mechanism by which transport by GAP proteins occurs is largely unknown, statistical relationships between GAP proteins and members of RhtB and DsbD families were determined (Supplemental Figures 9A and 10E). A comparison between sequences containing TMSs 1-5 of the RhtB homologue Hgr1 (F3KVR3) and the GAP homologue Ssp3 (NCBI: WP_019358971.1) yielded a comparison score of 14.5 S.D., demonstrating homology between the two families. A score of 16.6 S.D. was found when comparing the full length sequence of Ssp3 with that of the GAP protein, TC# 2.A.116.1.7 (K6W6C5), and a score of 45.2 S.D. resulted when comparing the full sequences of Hgr1 and RhtB protein, TC# 2.A.76.1.5 (P76249). This relationship with the LysE superfamily allows predictions and guided exploration into the mechanistic features of GAP proteins.

3.2.9 Disulfide Bond Oxidoreductase D Family (DsbD; TC# 5.A.1)

Homology was established between DsbD and the RhtB, CaCA2, MntP, NAAT and GAP family proteins (Supplemental Figures 10A-10E). In exploring these relationships, 6 TMSs of the NAAT homologue Pfu1 (Q8U2T5) were found to align with 6 TMSs of the DsbD homologue Dto1 (K0NNX9), yielding a score of 15.3 S.D (Supplemental Figure 10D). A score of 41.9 S.D. resulted when comparing the full length sequences of Dto1 with DsbD protein, TC# 5.A.1.2.1 (P45706), and comparing the full length sequences of Pfu1 and NAAT protein, TC# 2.A.95.1.4 (Q8J305) yielded a score of 82.4 S.D. These alignments establish membership within the LysE superfamily.

3.3 Topological Analyses

Using ClustalX, Mafft and Probcons, we created multiple alignments for homologues within each family included in our study [63]. The alignments generated with each program showed a high degree of agreement. Because Mafft alignments were able to produce comparable residue patterns to ClustalX without excessive expansion of the residue position axis (Supplemental Figure 11), Mafft alignments were selected to represent the data. With these Mafft alignments, we generated AveHAS plots to examine the relative average hydropathy, amphipathicity and similarity plots for the homologues (Supplemental Figure 11). Additionally, AveHAS plots were generated from multiple alignments of homologues for all families with established statistical relationships (Figure 2).

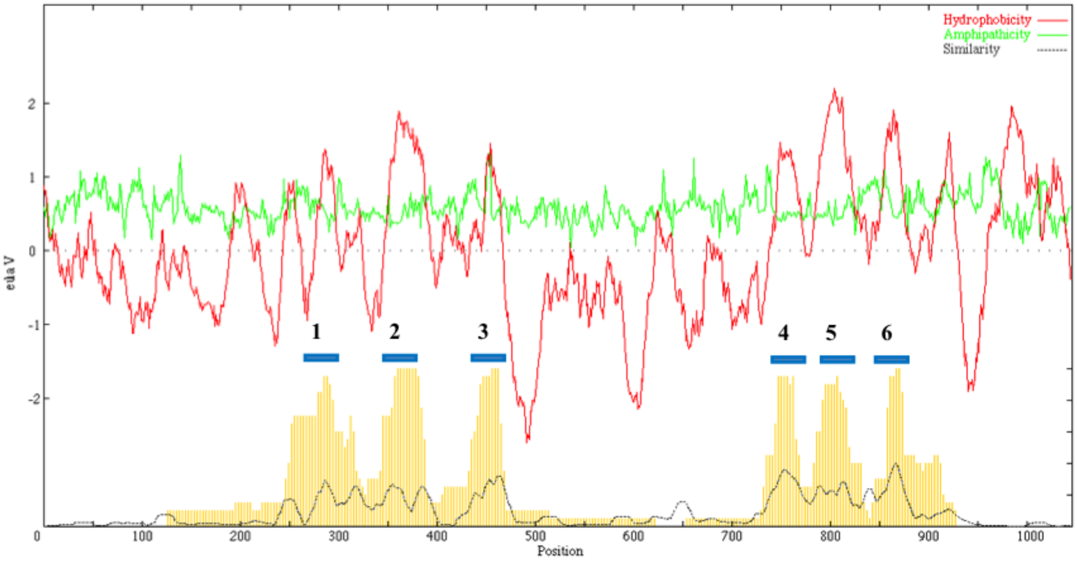


Figure 2. Combined AveHAS plot of proteins in the eleven recognized families in the LysE superfamily. Upper plot: The dark line shows average hydropathy while the light line shows average amphipathicity. Lower plot: The dotted line presents average similarity while the vertical lines indicate average hydropathy, determined by a second method. Numbers above the six bars indicate their TMSs in the basic transport protein unit.

Examining the plots for Supplemental Figures 11A-11K, we observe that the homologues for the LysE, RhtB, CadD, CaCA2, MntP, NAAT, NicO, GAP and DsbD families are most similar in regions corresponding to predicted TMS#1 and TMS#6. Furthermore, these figures show that the largest hydrophilic region separates TMSs #3 and 4, corresponding to regions that are highly dissimilar. These analyses support a 3+3 topological arrangement for all LysE superfamily proteins. Homologues of TerC and ILT display a 7-TMS core (Supplemental Figures 11J-11K) but share the previous characteristics with LysE, RhtB, CadD, CaCA2 and MntP. With respect to the TerC and ILT proteins, we observe a predicted 3+3+1 topological arrangement (Figure 1), but many ILT family homologues have 8 predicted TMSs, where an additional hydrophobic peak occurs at the N-termini. TerC proteins, on the other hand, can vary between 6 to 9 TMSs, and additions may occur either in the C-terminal or N-terminal regions of the sequences.

Finally, we examined a combined AveHAS plot of all eleven families with established statistical relationships. The plot (Figure 2) reveals a core of 6 TMSs among the different families with a large hydrophilic region separating the aligned core TMS#3 and TMS#4. These results further support a 3+3 TMS arrangement for members of the LysE superfamily.

3.4 Identifying Internal Repeats

Previous work on the LysE superfamily suggested that members derived from a 3-TMS internal duplication to result in a 3+3 TMS arrangement [1]. A recent examination of ILT transporters suggested a 3+3+1 arrangement with two 3-TMS repeat elements followed by a single extra TMS [33]. In addition, CaCA2 and DsbD proteins have been suggested to contain 3-TMS repeat elements [25,55]. Using IntraCompare and GSAT, we report evidence for internal 3-TMS repeats in several members of the LysE superfamily (Table 4, Supplemental Figures 12-15). This evidence supports the proposed hypothesis that all of these proteins arose via a common intragenic duplication event.

Table 4: Protein families with Demonstrated Internal Repeat Elements**.** UniProt accession numbers are provided in Column 2. The TMSs aligned refers to the positions of the TMSs from the N-terminus. For 6-TMS proteins, we find the 3-TMS internal repeat elements occur as two tandem 3-TMS elements for all families examined. For 7-TMS proteins, we find the 3-TMS internal repeat elements in the first 6 TMSs, suggesting these 7-TMS proteins have a 3+3+1 topology. The GSAT alignments generated using 20,000 shuffles for these comparisons are presented in Column 6.

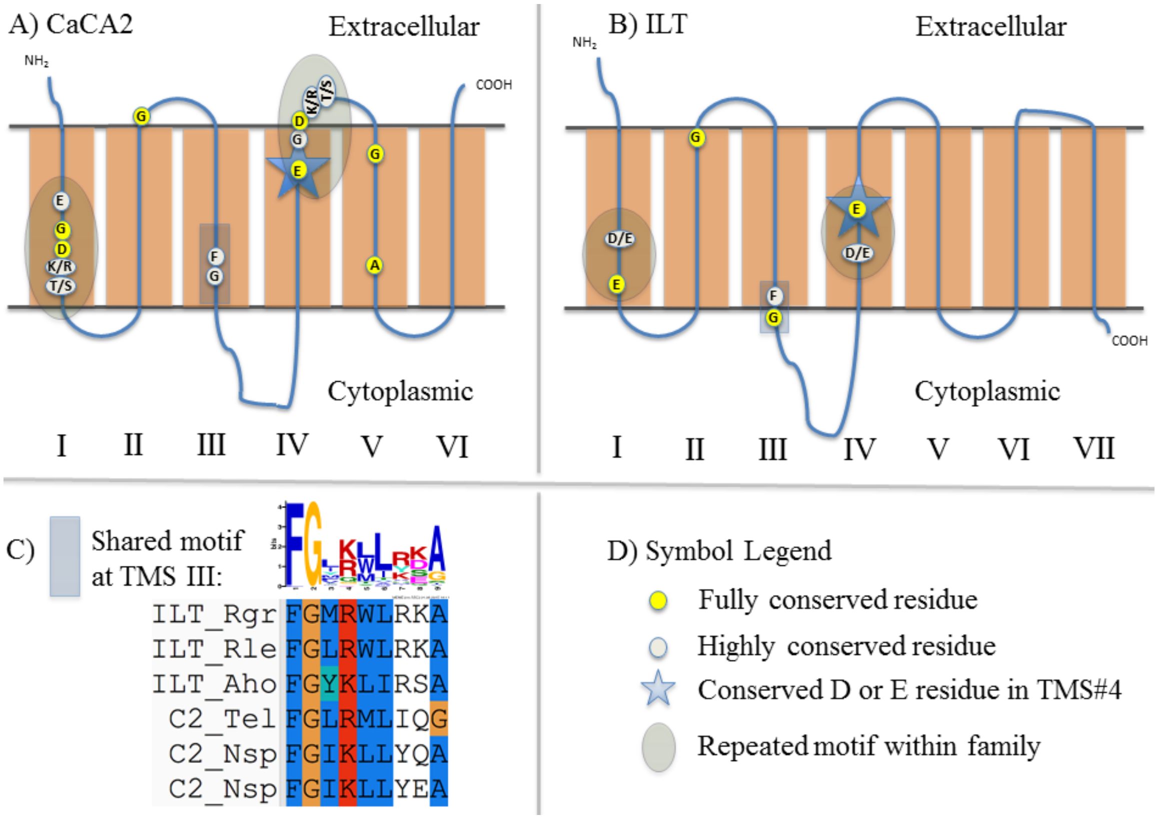
Family	Protein Accession #	# of TMSs in Protein	TMSs aligned	Score (S.D.)	Figure #
CaCA2	Q2JWH3	6	1-3 and 4-6	13.5	S11A
	I7M883	6	1-3 and 4-6	11.3	S11B
	K4DX00	6	1-3 and 4-6	5.7	S11C
ILT	Q8YX33	7	1-3 and 4-6	10.7	S12A
	K9Q6B8	7	1-3 and 4-6	9.4	S12B
	J2KV33	7	1-3 and 4-6	8.0	S12C
MntP	A8SU47	6	1-3 and 4-6	8.1	S13A
	R9SLI6	6	1-3 and 4-6	7.4	S13B
	C6JCY1	6	1-3 and 4-6	6.9	S13C
TerC	A4IKQ1	7	1-3 and 4-6	9.4	S14A
	G8M4S7	7	1-3 and 4-6	9.1	S14B
	R9LI44	7	1-3 and 4-6	7.8	S14C

Strong evidence is seen in the 6-TMS CaCA2 Ssp2 protein (Supplemental Figure 12). Comparing the first and second halves of the Ssp2 protein (Q2JWH3), TMSs 1-3 and TMSs 4-6 were found to align. The comparison yielded a score of 13.5 S.D., which is sufficient to establish the existence of two homologous internal repeats. The existence of this internal repeat element confirms previous reports regarding the repeating ExGD(KR)(TS) motif in TMS#1 and TMS#4 of the CaCA2 family [25]. Since we have demonstrated that CaCA2 is a member of the LysE superfamily, the other LysE superfamily proteins are presumed to share the same evolutionary pathway.

3.5 Motif Analyses

Previous mutation studies on the LysE protein in *Corynebacterium glutamicum* demonstrated the importance of highly conserved residues in the second and fourth hydrophobic segments of the protein [86]. A highly conserved aspartic acid (D) is present in the second hydrophobic segment of LysE, and its negative charge is essential for translocation of L-lysine. In addition, mutations to the fully conserved asparaginyl (N) and prolyl (P) in the fourth hydrophobic segment reduce export function dramatically. The prolyl residue in particular holds importance for three-dimensional structures of the carrier, and any changes in the neighboring asparaginyl residue would introduce steric hindrance. A fully conserved aspartic acid (D) is also present in the fourth hydrophobic segment, and has been proposed to bind the L-lysine substrate. Change of this aspartic acid (D) to a lysyl (K) residue resulted in an inactive protein. In the present study, motifs identified using the MEME/MAST Suite (www.meme.nbcr.net/meme/) for the different families were compared with one another

(Figures 3-6, Table 5) [68]. Here we report strongly conserved residues within and between families.



```
ILT_Rgr FGMRWLRKA
ILT_Rle FGLRWLRKA
ILT_Aho FGYKLIRSA
 C2_Tel FGLRMLIQG
 C2_Nsp FGIKLLYQA
 C2_Nsp FGIKLLYEA
```

Figure 3. Schematic diagrams depicting motifs and highly conserved residues within and between the CaCA2 and ILT families. Highly conserved residues were identified using alignments generated from Mafft. In Part C, the MEME/MAST Suite was used to generate the graphical logo, and the alignment was presented using the ClustalX2 user interface with the associated Mafft multiple sequence alignment (MSA). **A)** Schematic diagram of CaCA2 proteins. **B)** Schematic diagram of ILT proteins. **C)** Graphical representation of the shared motifs depicted in Part A and Part B. **D)** Symbol Legend.

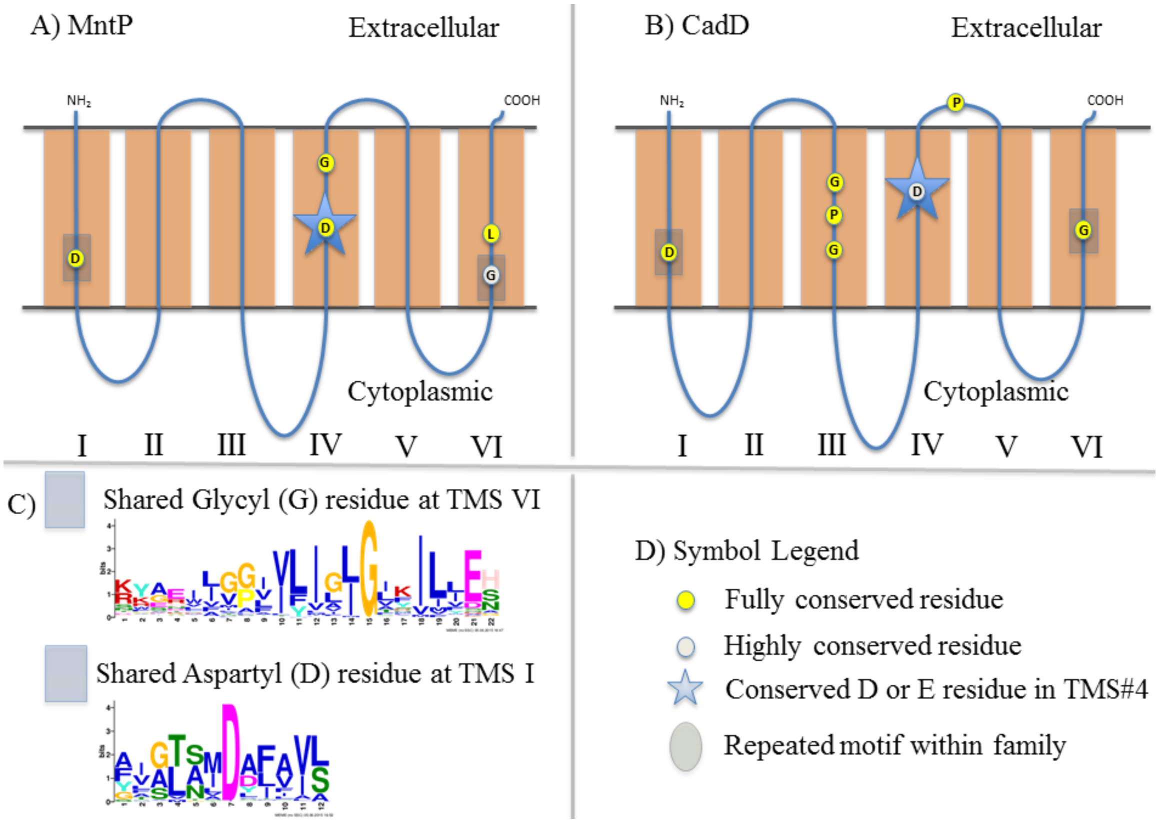
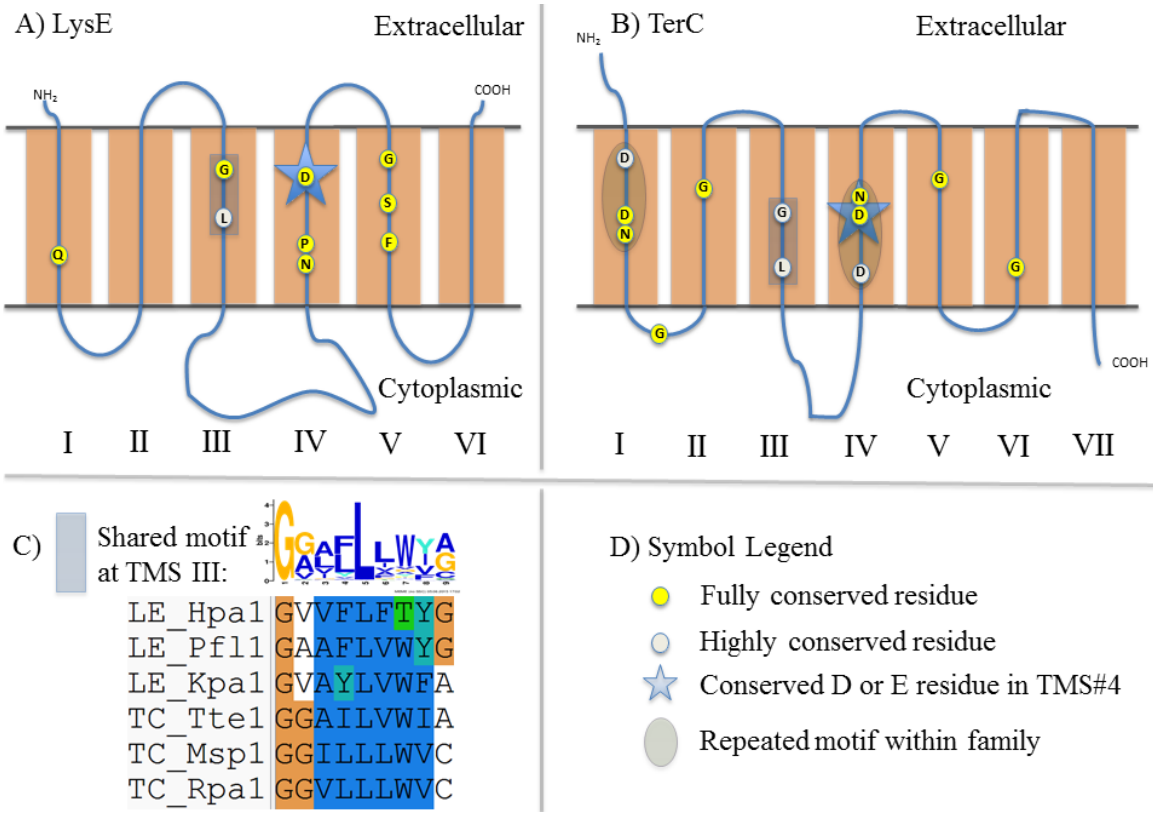


Figure 4. Schematic diagrams depicting motifs and highly conserved residues within and between the MntP and CadD families. **A)** Schematic diagram of MntP proteins. **B)** Schematic diagram of CadD proteins. **C)** Graphical representation of the shared motifs depicted in Part A and Part B. **D)** Symbol Legend.



```
LE_Hpa1 GVVFLFTYG
LE_Pfl1 GAAFLVWYG
LE_Kpa1 GVAYLVWFA
TC_Tte1 GGAILVWIA
TC_Msp1 GGILLLWVC
TC_Rpa1 GGVLLLWVC
```

Figure 5. Schematic diagrams depicting motifs and highly conserved residues within and between the LysE and TerC families. **A)** Schematic diagram of LysE proteins. **B)** Schematic diagram of TerC proteins. **C)** Graphical representation of the shared motifs depicted in Part A and Part B. **D)** Symbol Legend.

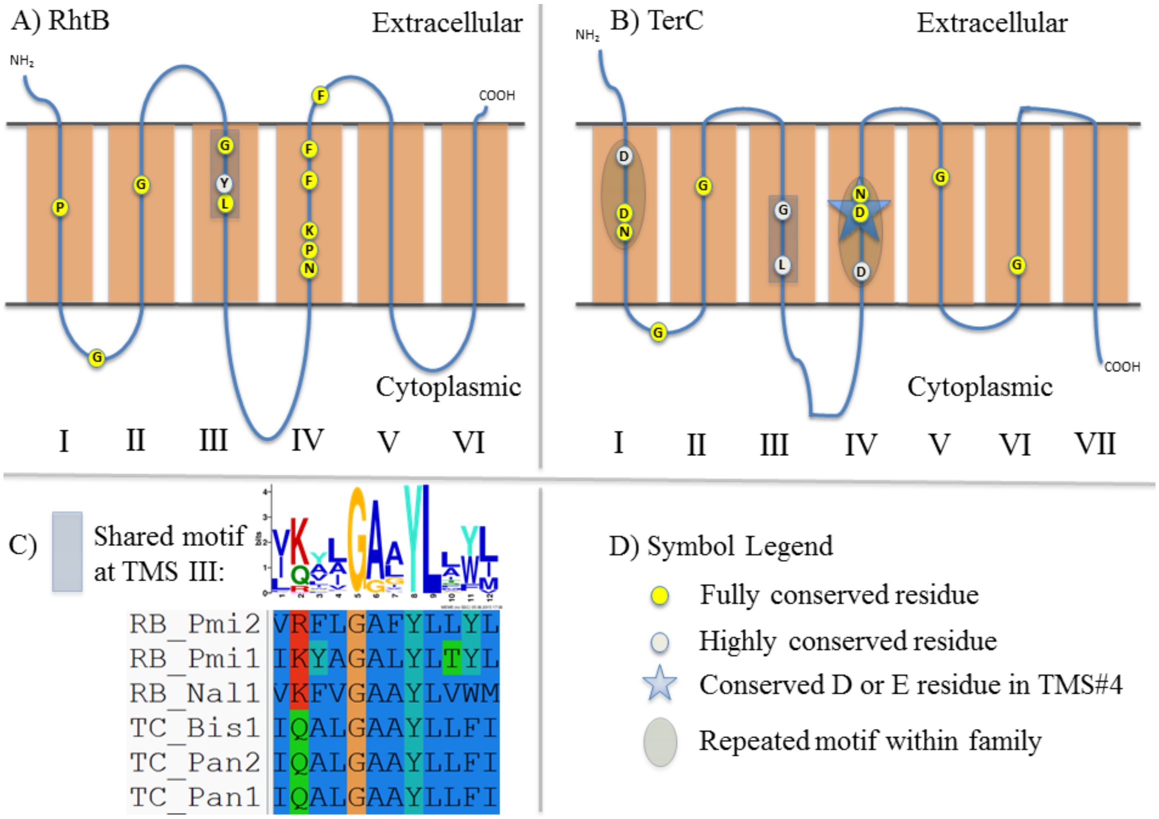


Figure 6. Schematic diagrams depicting motifs and highly conserved residues within and between the RhtB and TerC families. **A)** Schematic diagram of RhtB proteins. **B)** Schematic diagram of TerC proteins. **C)** Graphical representation of the shared motifs depicted in Part A and Part B. **D)** Symbol Legend.

Table 5: Protein families with Identified Motifs using MEME/MAST. Protein families demonstrating shared, conserved residues are shown below. HMMTOP was used to predict the TMS location for each motif. Schematic diagrams showing the motif locations and other highly conserved residues are found in Figures 3-6.

Families	Predicted TMS region	# Proteins displaying motif/# of Total proteins	Motif
CaCA2 & ILT	#3 of both	80/80 (40 ILT, 40 CaCA2)	FGX(K/R)XL
CadD & MntP	#4 of both	170/170 (85 CadD, 85 MntP)	Fully Conserved D
CadD & MntP	#6 of both	170/170 (85 CadD, 85 MntP)	Conserved G
CadD & MntP	#1 of both	170/170 (85 CadD, 85 MntP)	Fully Conserved D
TerC & LysE	#3	248/248 (124 LysE, 124 TerC)	GXXXL
TerC & RhtB	#3	176/176 (88 RhtB, 88 TerC)	GXXYL

3.5.1 CaCA2 vs. ILT

80 proteins of CaCA2 and ILT homologues were combined and found to exhibit a shared motif in TMS#3 in these 6-TMS proteins (Figures 3A-3B, Table 5). Not only do the two motifs align in the MEME/MAST Suite, all tested proteins share many strongly conserved residues. Positions 1-2 of this motif correspond to the second half of TMS#3 that is shared in proteins of the two families. Of the 9 positions, amino acids in positions 1, 3, 5, 6 and 9 consist largely of hydrophobic residues. In positions 1 and 2, both families contain fully conserved phenylalanine (F) and glycine (G) residues, respectively.

At TMS#1 and TMS#4, both families contain two strongly conserved negatively charged amino acyl residues (D/E). Similar to proteins in the CaCA2 and ILT families, conserved negatively charged residues have been found in MntP, CadD and TerC proteins (Figures 3-6). With the exception of the CadD proteins, the conserved, negatively charged residues in TMS#1 and TMS#4 within each protein align (Supplemental Figures 12-15). The D/E residue in these 5 families could have functional significance similar to the D residue in the fourth hydrophobic segment of LysE described previously. However, the biological significance of the conserved, negatively charged residues in TMS#1 is not yet understood. These findings imply an evolutionary relationship between these five families and a closer relationship between CaCA2 and ILT.

3.5.2 MntP vs. CadD

Sequences of 85 MntP and 85 CadD proteins, all containing 6 TMSs, were combined into a single file shown to share motifs (Figures 4A-4B, Table 5). The best shared motif in TMS#4 of MntP and CadD proteins was found in all of 170 selected proteins. Positions 1-13 in this motif correspond to the second half of TMS#4 that is shared in proteins of these two families. A highly conserved aspartic acid (D) is contained in this shared motif. Differing within the TMS#4 motif are positions 5, 8, 12 and 14. Position 5 is a fully conserved serine (S) in MntP homologues, but is a strongly conserved glycine (G) in CadD homologues. Position 8 is a strongly conserved asparagine residue in CadD homologues, but a strongly conserved alanine in MntP homologues. Additionally, position 12 corresponds to a well-conserved tyrosine in CadD proteins, but a fully conserved glycine in MntP proteins. Finally, we note well-conserved polar amino acids in position 14 for MntP homologues, but a conserved proline residue in CadD homologues.

A shared motif corresponding to the entire TMS#6 in 85 MntP and 85 CadD proteins was identified (Figures 4A-4B, Table 5). A completely conserved glycine was shared at position 15, and strongly conserved acidic residues occurred at position 21. Finally, well-conserved hydrophobic amino acids were present in positions 6, 9, 10, 12, 14, 16, 18, 19 and 20, providing additional support for a close evolutionary relationship between MntP and CadD proteins.

The strongly conserved residues of the two sets of homologues differ at positions 4, 7, 8, 11, 13 and 22. In position 4, negatively charged amino acids are largely conserved only in MntP homologues. Position 11 differs where a completely conserved leucine residue in MntP homologues but either a phenylalanine or a tyrosine in CadD homologues is found. A glycine is well-conserved at position 13 of CadD homologues, but it is weakly conserved in MntP homologues. Position 22 of CadD homologues shows well-conserved polar amino acids (S, N), while this position in MntP homologues contains a conserved histidyl residue. Finally, we note two unique residues at positions 7 and 8: proline and glycine. Conserved proline residues can be found in CadD only (position 8), while two almost fully conserved glycines are present in MntP homologues (positions 7 and 8). These unique differences may provide insight into the divergence of these proteins and possibly correlate with their differing specificities.

3.5.3 LysE, RhtB and TerC

More distantly related are the motifs within members of the LysE, RhtB and TerC families. Among these three families, two residues in TMS#3 are shared (Figures 5-6, Table 5). In the middle of TMS#3, all three families show a fully conserved glycine. Additionally, a fully conserved leucine, three residues (one helical turn) away from the glycine, can be found. Strongly conserved hydrophobic residues between the fully conserved glycyl and leucyl residues are present. A tyrosine (Y) is also conserved between 88 RhtB and 88 TerC proteins (GxxYL) but is not observed in LysE proteins (GxxxL).

3.6 Phylogenetic Tree

Proteins listed in TCDB for each family were used to generate a phylogenetic tree based on tens of thousands of BLAST bit-scores using the SFT1 program (Figure 7) [72].

RhtB, LysE and TerC localize to a single branch. Similarly, CaCA2 clusters with ILT, and CadD clusters with MntP. Based on these branching patterns, members in each of these groupings must be more strongly related to each other than to other families as had been suggested from motif analyses. A tree including all eleven families generated using a Mafft multiple alignment and RAxML was included for comparison (Supplemental Figure 17). The SFT and Mafft trees show remarkable agreement, particularly with respect to family relationships. However, the branches sometimes differ between the two trees (compare Figure 7 with Supplemental Figure 17), but all of the proteins cluster with their respective families, with the exception 2.A.109.3.1 (TerC.3.1), 2.A.108.2.6 (ILT.2.6) and 2.A.108.3.2 (ILT.3.2). A significant difference deals with the proteins of the CaCA2 family in the two trees. Based on our previous experience [71,72,73,74,75,76,77,78], this and other differences suggest that the phylogenetic distances between the eleven families are too great to allow the generation of accurate multiple sequence alignments. Trees representing each individual family have been constructed using multiple alignments generated by ClustalX, Mafft and ProbCons (Supplemental Figures 18-28).

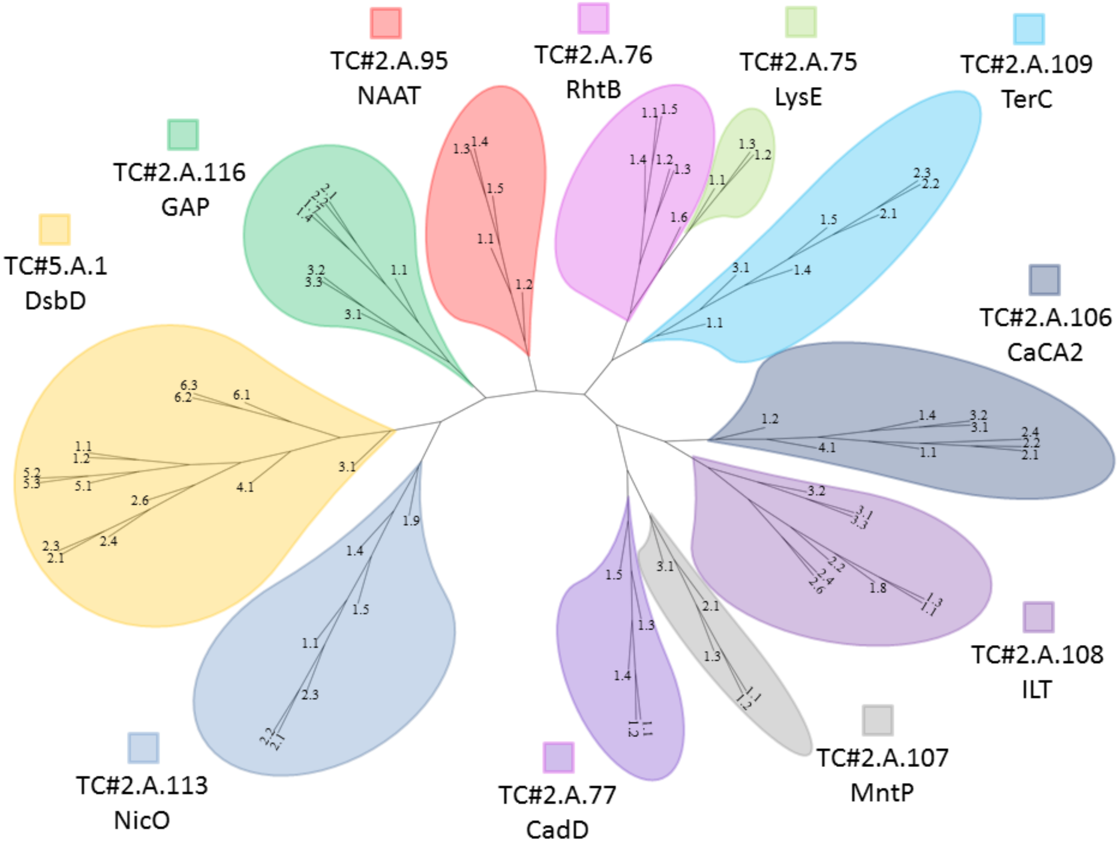


Figure 7. Phylogenetic Tree of the LysE Superfamily. The tree was generated using the SuperFamilyTree program and viewed using FigTree. It depicts the evolutionary relationship between the 11 different families in this study. Clustering indicates closer phylogenetic relationships. The tree is based on tens of thousands of BLAST bit scores generated with the SFT1 program where every protein was compared with every other protein included in the analysis. The SFT2 program was used to integrate all of the information to show the relationships of the eleven families to each other.

Discussion

Using rigorous statistical criteria, we have expanded the LysE superfamily nearly four-fold. In addition to the LysE, RhtB and CadD families identified previously, this superfamily now includes the following families: NAAT, CaCA2, MntP, ILT, TerC, NicO, GAP and DsbD. Members of each of these families have been characterized and shown to play roles in transport of amino acids and resistance of heavy metal ions, along with cell surface maintenance. Most families include secondary carrier type transporters catalyzing heavy metal or amino acid efflux, but one family catalyzes amino acid uptake, another catalyzes heavy metal ion uptake, and a third catalyzes transmembrane electron transfer. GAP proteins have not been mechanistically characterized, but based on their inclusion in the LysE superfamily, we tentatively propose that GAP proteins operate as secondary carriers, where the energy source for lipid export is the proton motive force.

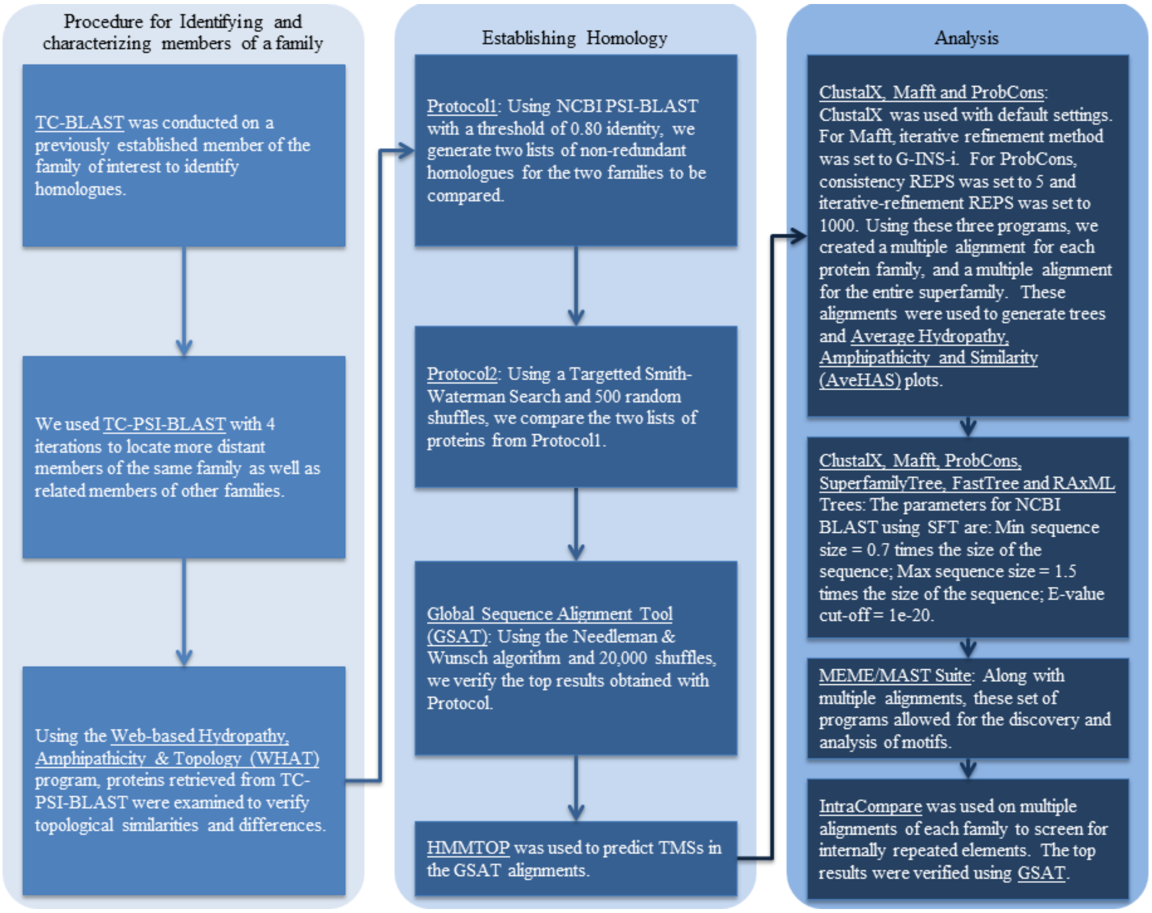
Through sequence analyses, we were able to recognize a distinct pattern of homology. That is, LysE, RhtB, NAAT, CaCA2, MntP, ILT, TerC, NicO, GAP and DsbD proved to be homologous in 3 or more TMSs. The 3 TMSs that aligned are usually between the first 3 TMSs, the second 3 TMSs or both. This observation fits the predicted evolutionary pathway presented in Figure 1. The presence of 3-TMS internal repeats supports the conclusion that all members of the LysE superfamily arose from a 3-TMS precursor via the same pathway in which the proposed duplication gave rise to 6 TMSs in a 3+3 TMS arrangement. In some TerC and ILT proteins, the topologies differ from the 3+3 TMS arrangement with the addition of one or two TMSs at the C- or N-terminal end, resulting in a 3+3+1, 3+3+2, or 1+3+3 arrangement.

According to the phylogenetic tree, amino acid exporter families RhtB and LysE branch close to each other, as suggested from previous studies [1]. In contrast to these two amino acid exporter families, TerC, which branches near RhtB and LysE in the tree, has been observed to play roles in tellurium ion resistance. MntP and CadD cluster together, and both are involved in divalent metal cation transport. Likewise, divalent cation transporters of the CaCA2 and ILT families branch in close proximity.

This study suggests that members of the LysE Superfamily are involved in ionic homeostasis, protection from excessive cytoplasmic heavy metal/metabolite concentrations, cell envelope assembly and transmembrane electron flow. Many of the family members, however, are still poorly understood from functional and physiological standpoints. In continuing this project, genome context analyses will be conducted on members of each family. This will allow functional predictions, further promoting an understanding of the significance of these proteins. To date, no crystal structures exist for a member of this superfamily, and such studies will be crucial for understanding their mechanistic details. Thus, studies on the LysE superfamily remain in their infancy.

Appendix

Supplemental Figures



Procedure for Identifying and characterizing members of a family

TC-BLAST was conducted on a previously established member of the family of interest to identify homologues.

We used TC-PSI-BLAST with 4 iterations to locate more distant members of the same family as well as related members of other families.

Using the Web-based Hydropathy, Amphipathicity & Topology (WHAT) program, proteins retrieved from TC-PSI-BLAST were examined to verify topological similarities and differences.

Establishing Homology

Protocol1: Using NCBI PSI-BLAST with a threshold of 0.80 identity, we generate two lists of non-redundant homologues for the two families to be compared.

Protocol2: Using a Targetted Smith-Waterman Search and 500 random shuffles, we compare the two lists of proteins from Protocol1.

Global Sequence Alignment Tool (GSAT): Using the Needleman & Wunsch algorithm and 20,000 shuffles, we verify the top results obtained with Protocol.

HMMTOP was used to predict TMSs in the GSAT alignments.

Analysis

ClustalX, Mafft and ProbCons: ClustalX was used with default settings. For Mafft, iterative refinement method was set to G-INS-i. For ProbCons, consistency REPS was set to 5 and iterative-refinement REPS was set to 1000. Using these three programs, we created a multiple alignment for each protein family, and a multiple alignment for the entire superfamily. These alignments were used to generate trees and Average Hydropathy, Amphipathicity and Similarity (AveHAS) plots.

ClustalX, Mafft, ProbCons, SuperfamilyTree, FastTree and RAxML Trees: The parameters for NCBI BLAST using SFT are: Min sequence size = 0.7 times the size of the sequence; Max sequence size = 1.5 times the size of the sequence; E-value cut-off = 1e-20.

MEME/MAST Suite: Along with multiple alignments, these set of programs allowed for the discovery and analysis of motifs.

IntraCompare was used on multiple alignments of each family to screen for internally repeated elements. The top results were verified using GSAT.

Supplemental Figure 1 Flowchart of the materials and methods. Along with a step-wise description of the methods, the parameters for the programs used in major analyses are summarized.

S2A

```
# 1: A_Sequence: Pst1 (2.A.75.1.1 homologue)
# 2: B_Sequence: Bth1 (2.A.76.1.5 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 171
# Identity:      57/171 (33.3%)
# Similarity:    87/171 (50.9%)
# Gaps:          20/171 (11.7%)
# Score: 159.0
#========================================
                         1                         2
A_Sequence         1 LILPLGPQNAFVLN----QGVKRHYHLMTATLCSLSDVVLICAGIFGGSA     46
                     ::|| || : :||:    :|||  |         : | ||:     | ::
B_Sequence         1 ILLP-GPNSMYVLSLAAQRGVKAGYRAACGVF--VGDTVLMVLSAAGVAS     47
                         1                         2

                              3
A_Sequence        47 LLQQSPLLLTVITWAGVAFLLWYGWGALRTAFRRELALA-SGLDIRQS-R     94
                     ||: :||| :|: : | |:||: | | || |:|:    | :| |:|::
B_Sequence        48 LLKANPLLFSVVKYGGAAYLLYIGSGMLRGAWRKLARPADAGADVRRAVD     97
                              3
                                4                                     5
A_Sequence        95 G-RIIATLLAVTWLNPHVYLDTFVVLGSLGSQFPD---TH-ARQWFALGT    139
                     | |     | |: |||   |  | :  |   || |    | |  :  ||
B_Sequence        98 GERPFRKALVVSLLNPKAIL--FFI--SFFIQFVDPSYAHPALSFVVLGA    143
                                4                                     5

A_Sequence       140 VS--ASVLWFFGLALLAAWLA    158
                     ::  || ::   |    | ||
B_Sequence       144 IAQFASFVYLSTLIFTGARLA    164
#----------------------------------------
============ FINISHED =============
Average Quality (AQ)      18.75 +/- 6.96
Standard score (Z):       20.0
Precise score (Z):  20.1
```

Supplemental Figure 2. GSAT comparisons between previously established LysE superfamily members. (A) LysE vs. RhtB. (B) RhtB vs. CadD. (C) LysE vs. CadD.

S2B

```
# 1: A_Sequence: Hal1 (2.A.76.1.5 homologue)
# 2: B_Sequence: Oki1 (2.A.77.1.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 124
# Identity:      33/124 (26.6%)
# Similarity:    63/124 (50.8%)
# Gaps:          12/124 ( 9.7%)
# Score: 107.0
#=======================================
                          2                      3
A_Sequence         1 IGDAVLIFCAYIGIASLIRSSPFLFSLVKMLGALYLLYLGLKILYSTLAK      50
                     ||: :||  : : :| |::  |  : :: :|| |: : :|||  :|
B_Sequence         1 IGNGILIVMSLL-LAYLLKFIPESW-ILGLLG-LFPITVGLKTFFS----      43
                          2                      3

                                           4
A_Sequence        51 KGQEQSAAKEEPEHTFRKALTLSLTNPKA--ILFYVSFFVQFIDMDYAHT      98
                     |  | : ||    |  |  : ::||   |  :  |: ||     :|:::
B_Sequence        44 KEDETAKAKASDAHLIRDVVLMTLTTCSADNLAIYIPFFA---SVDFSYL      90
                                           4

                          5
A_Sequence        99 GVSFAILAVILEMISFCYMTLLIF    122
                      |   :  :||  :||  : :  |
B_Sequence        91 PVILIVFLLILSAVSFTALKITKF    114
                          5
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)      23.76 +/- 7.02
Standard score (Z):       12.0
Precise score (Z):  11.9
```

Supplemental Figure 2. GSAT comparisons between previously established LysE superfamily members. (A) LysE vs. RhtB. (B) RhtB vs. CadD. (C) LysE vs. CadD, cont.

S2C

```
# 1: A_Sequence: Asp2 (2.A.75.1.3 homologue)
# 2: B_Sequence: Cth1 (2.A.77.1.2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 144
# Identity:      38/144 (26.4%)
# Similarity:    65/144 (45.1%)
# Gaps:          14/144 ( 9.7%)
# Score: 92.0
#=======================================
                                 2                             3
A_Sequence         1 LRQGLRREHVMPVVLVCALSDAVLLQVGVWGMGGVLLARPEWAQFMRWAG     50
                     :    || |:: :      :  ||  :  :  ||::: | ||   :   |
B_Sequence         1 INANFRRRHIV-IGQYLGFTTIVLASLPGF-FGGLIVPR-EWIGLL---G     44
                                 2                             3
                                                            4
A_Sequence        51 ALFLLMYAAQTAARALRPGQLLVATSGPGTSLRTTLATVVALTWLNPHVY    100
                      | :::   |   | :   |:   ||   :| | :  : : |: |||| |
B_Sequence        45 LLPIIIGFKQLVNRKIETVQVQTVTSFENSSYRNSTFSFL-LSLLNPHTY     93
                                                            4
                                                  5
A_Sequence       100 LDTVVLLGTMATPYPAWGRALFAAGGSLAS-----ALWFLLIGL    139
                         | |         :   || ||  |||     |::||::|:
B_Sequence        94 KVAAVTLANGGDNISIY-IPLF-AGSQLASLSIILAVFFLMVGV    135
                                                  5
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     16.11 +/- 6.25
Standard score (Z):      12.0
Precise score (Z):  12.1
```

Supplemental Figure 2. GSAT comparisons between previously established LysE superfamily members. (A) LysE vs. RhtB. (B) RhtB vs. CadD. (C) LysE vs. CadD, cont.

S3A

```
# 1: A_Sequence: Efa1 (2.A.77.1.1 homologue)
# 2: B_Sequence: Ptr2 (2.A.106.1.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 115
# Identity:      32/115 (27.8%)
# Similarity:    57/115 (49.6%)
# Gaps:          11/115 ( 9.6%)
# Score: 108.0
#=======================================
                            1                                     2
A_Sequence         1 LQNILSALAVYISTSI-DYLFILLIIFSQNHTKKGLRQIFFGQYLGTGIL     49
                          |::|: |:| | |  | :  : |  |::     :|||  |   ::
B_Sequence         1 WNAFTSSVAMIIATEIGDKTFFIAAVLSMKHSRSA---VFFGAILALIVM     47
                            1                                     2
                                         3
A_Sequence        50 VAISLFAAYVL-NFIPQDWIIGLLGLIPIYLGIRVAF------VGEEEEE     92
                       :|     :| ||||:::   | ||: :| | :: :       |:  ||
B_Sequence        48 TVLSTAMGMMLPNFIPKEYTHLLGGLLFLYFGCKLIYDSRQMEAGKTSEE     97
                                         3

A_Sequence        93 EGEVVEKLGSRGTNR    107
                       || |:|  :|   :
B_Sequence        98 LEEVEEELLQQGKKK    112
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     15.13 +/- 6.54
Standard score (Z):      14.0
Precise score (Z):  14.2
```

Supplemental Figure 3. GSAT comparisons with CaCA2. (A) CadD vs. CaCA2. (B) LysE vs. CaCA2. (C) RhtB vs. CaCA2.

S3B

```
# 1: A_Sequence: Cac2 (2.A.75.1.1 homologue)
# 2: B_Sequence: Mpu4 (2.A.106.1.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 139
# Identity:      38/139 (27.3%)
# Similarity:    62/139 (44.6%)
# Gaps:          16/139 (11.5%)
# Score: 91.0
#=======================================
                            1                               2
A_Sequence         1 LSLIVAVGPQNAMLLKYGIRRDHIGLIIVVCALSDVILITSGTAGVGYLV     50
                     | |:  :| :   |      | |  | : : ::| :  :| |:|  ||||
B_Sequence         1 LILLTEIGDKTFFLAMMLAARHGKLQVFLASISALFFMTLGSALAGYLV     49
                            1                              2
                                          3
A_Sequence        51 ----ERFPNALEALKYIGAAYLAFFTFTCFRDAFKTKGEAIDVESTSPNS     96
                         |   :::: : :: |     |      || |   |  | :
B_Sequence        50 STSAEMLHSSVKIMDWVAAVLFVLFGAQMLWDARKLHKE--DAKD-----     92
                                          3
A_Sequence        97 TEEVATFDGDGDSTGGVGTEHGSVATATATQRQEIKRSP    135
                      ||||   | |:  |   : ||  | |  | |:: ::||
B_Sequence        93 -EEVAALLG-GE--GARSSSHGERADAEETLREKDEKSP    127

#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     12.36 +/- 5.83
Standard score (Z):      13.0
Precise score (Z):  13.5
```

Supplemental Figure 3. GSAT comparisons with CaCA2. (A) CadD vs. CaCA2. (B) LysE vs. CaCA2. (C) RhtB vs. CaCA2, cont.

S3C

```
# 1: A_Sequence: Hal1 (2.A.76.1.5 homologue)
# 2: B_Sequence: Cmi1 (2.A.106.1.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 166
# Identity:      42/166 (25.3%)
# Similarity:    81/166 (48.8%)
# Gaps:          19/166 (11.4%)
# Score: 103.0
#=======================================
                         2                                 3
A_Sequence         1 LAVFIGDAVLIFCAYIGIASLIRSSPFLFSL-VKMLGALYLLYLGLKILY     49
                     |:|:||  :::    :|   |  |  ||  : :: :||:   : |:|:||
B_Sequence         1 LSVWIGQLLMLLPKLVG-QYLPPSLGFLTHISIEYVGAVLFFFFGIKLLY     49
                         2                                 3

                                                                    4
A_Sequence        50 S--TLAKKGQ-----EQSAAKEEPEHTFRKALTL-SLTNPKAILFYVSFF     91
                     |   :::|       |   | |: |  |::  |   :     :| :|: :
B_Sequence        50 SARNMSRKTDIEVMAEAEEAIEDGERKFKQRNTAWKIFIESGVLTFVAEW     99
                                                                    4

                                                   5
A_Sequence        92 ---VQFIDMDYAHTGVSFAILAVILEMISFCYMTLLIFSGAALAHFLSEK    138
                         ||  :  | |  |  ::|  :   : | : : :  | |:|  :||
B_Sequence       100 GDRTQFATVTLAATKDSLGVMAGGIVGHAICAL-IAVIGGRAIASHISE-    147
                                                   5

                                   6
A_Sequence       139 KRLAKLGNSMVGLLFL    154
                     : :  :|    ||||:
B_Sequence       148 RTITIIG----GLLFI    159
                                   6
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     18.42 +/- 6.49
Standard score (Z):      13.0
Precise score (Z):  13.0
```

Supplemental Figure 3. GSAT comparisons with CaCA2. (A) CadD vs. CaCA2. (B) LysE vs. CaCA2. (C) RhtB vs. CaCA2, cont.

S4A

```
# 1: A_Sequence: Kor1 (2.A.76.1.5 homologue)
# 2: B_Sequence: Cce1 (2.A.107.1.2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 105
# Identity:      26/105 (24.8%)
# Similarity:    58/105 (55.2%)
# Gaps:           9/105 ( 8.6%)
# Score: 80.0
#=======================================
                         3                                          4
A_Sequence         1 LYLAYLGINMLRGAWAARRRTAAQAPAQTLSNIHTHDNV-FRHALLLSLS      49
                     : || :||||:: :    | :: :    |:::::| ::: |::  :|:::
B_Sequence         1 ILLAIIGINMIKES----RNSSCEVAVDTVADVNTDNSLSFKNMFVLAVA      46
                         3                                          4

                                                            5
A_Sequence        50 NPKAALFFLSFFIPFVNPRYPHPALSFFILAAVMQTLSMCYLATLALAGD      99
                         ||  :     |:|     ||:||  :  |  ||||  :   :: |:
B_Sequence        47 TSIDAL-AVGITFAFLNVNI-IPAVSF--IGIVTFTLSMIGVRIGSVFGE      92
                                                            5

A_Sequence       100 KLLAK    104
                     |  ::
B_Sequence        93 KFKSR     97
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     11.53 +/- 4.99
Standard score (Z):      14.0
Precise score (Z):  13.7
```

Supplemental Figure 4. GSAT comparisons with MntP. (A) RhtB vs. MntP. (B) CadD vs. MntP. (C) CaCA2 vs. MntP.

S4B

```
# 1: A_Sequence: Hku1 (2.A.77.1.5 homologue)
# 2: B_Sequence: Dvu1 (2.A.107.1.2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 137
# Identity:      41/137 (29.9%)
# Similarity:    69/137 (50.4%)
# Gaps:          16/137 (11.7%)
# Score: 110.0
#=======================================
                                  3                                  4
A_Sequence         1 -DKWIVGLLGLIPLFIGIKFALSGEDEDETEEIREKIEQDKSKNLLWTVV     49
                      | |:    ||: |:||::  :  |  :||||  ::   | :: |  |::
B_Sequence         1 WDHWLA--FGLL-LYIGVR--MMREAFEETEENDDRC--DPTRGL--TLI     41
                                  3                                  4
                                                           5
A_Sequence        50 LLTIASGGDNLGVYIPYFSSLNWSKIIIVLIIFAIGIAILCELSRSLSKI     99
                     :| :|:  | | | :    ||:   | ||     ||:  |   :  |
B_Sequence        42 MLAVATSIDALAVGL----SLSVLGIDIVTPAIVIGVVCLLFTATGLHLG     87
                                                           5


                                          6
A_Sequence       100 PMVS--EIIEKYEKIIVPVVFIALGIYIMYENGTIQT    134
                      |:|  | : :   :   || | :|: |:||:|   |
B_Sequence        88 RMLSRAESLGRRAALAGGVVLIGIGLRILYEHGVFDT    124
                                          6
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     14.66 +/- 6.04
Standard score (Z):      16.0
Precise score (Z):  15.7
```

Supplemental Figure 4. GSAT comparisons with MntP. (A) RhtB vs. MntP. (B) CadD vs. MntP. (C) CaCA2 vs. MntP, cont.

S4C

```
# 1: A_Sequence: Csp2 (2.A.106.1.1 homologue)
# 2: B_Sequence: Eco2 (2.A.107.1.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 152
# Identity:      47/152 (30.9%)
# Similarity:    70/152 (46.1%)
# Gaps:          15/152 ( 9.9%)
# Score: 134.0
#=======================================
                         2                                 3
A_Sequence         1 AALASMTLLSVLMGQAISFLPKHYI----HWAEIALFLGFGLKLIYDASQ     46
                     |   |:  |: |:| ||  : :|||    ||    | |  ||::|| | |
B_Sequence         1 AVFGSVETLTPLIGWAIGSVAQHYIADWDHWIAFTLLLLLGLRMIYGALQ     50
                         2                                 3
                                                             4
A_Sequence        47 MPSQSQGTVIKEAAEAVDQIPQSGNR-----LTKLLARYPQIGIWLQAFS     91
                      | |  |   :::|||  :  ||| |     :   :|    |   :
B_Sequence        51 -PEQPAG---EQSAEAQPESGQSGRRPPSPLMLVAIAFATSIDSMIVGVG     96
                                                             4
                                                   5
A_Sequence        92 MTFLAEWGDRTQISTIALASS-YNVIGVTTGAILGHGICSVIAVIGGKLV    140
                     : || |         | ||::    ||:  |: ||  |     ::|| ::
B_Sequence        97 LAFL-EVNILLTALAIGLATTIMAAIGLRLGSFLGSAIGKRAEILGGLVL    145
                                  5                                 6

A_Sequence       141 AG     142
                      |
B_Sequence       146 IG     147
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     22.13 +/- 7.43
Standard score (Z): 15.0
Precise score (Z):  15.1
```

Supplemental Figure 4. GSAT comparisons with MntP. (A) RhtB vs. MntP. (B) CadD vs. MntP. (C) CaCA2 vs. MntP, cont.

S5A

```
# 1: A_Sequence: Lbr1 (2.A.77.1.1 homologue)
# 2: B_Sequence: Sma2 (2.A.108.2.4 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 207
# Identity:      55/207 (26.6%)
# Similarity:   101/207 (48.8%)
# Gaps:          31/207 (15.0%)
# Score: 138.0
#=======================================
                         1                         2
A_Sequence         1 IDYLIILMVIFGA--TPKRHRFLVYLGDFLGTAILVLTSYLMAVILGFV-     47
                     :: |:|:: : |     |: | | ::  :|| |: || | : |::| |:
B_Sequence         1 VEALLIVLALIGTLKASKQKRGLKWV--YLGAALGVLASVVTAIMLQFLF     48
                         2                         3
                                          3
A_Sequence        48 PA-------EWLLGFLGLIPILM--GVKLLIFGEKEDDDLIENEIQKKTN     88
                     ||       | | | :|:  : |  || : :   | :    :| ::|: |
B_Sequence        49 PALTSGNNREMLEGAVGIFAVFMMIGVGVWLH-SKANISAWQNYMEKQLN     97
                                          4
                         4                                          5
A_Sequence        89 VILK---------VAIITIATCGADNIGIYVPLFTQISPTN--IPILLVT    127
                     :::          :: : :   ||: |  || :   ||  |  : ||
B_Sequence        98 LVMSTGSFVSMFALSFLAVFREGAETILFYVGILPNISLQNLLLGILAAV    147
                         5                                          6
                                                      6
A_Sequence       128 FFIMMTLFCYLGYLLSKIPTIGNILE--KWSRYITAVVYIGLGIYILWES    175
                       :||  | ::     ||| |  : :   |: || |   :|: |: |  :
B_Sequence       148 LILMMLAFVFI-KSSEKIP-IHRVFQLLTWTIYILAFKMLGVSIHALQLT    195
                                                      7
A_Sequence       176 GTL-THL    181
                       | ||:
B_Sequence       196 NALPTHV    202
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     27.14 +/- 8.23
Standard score (Z):      13.0
Precise score (Z):  13.5
```

Supplemental Figure 5. GSAT comparisons with ILT. (A) CadD vs. ILT. (B) RhtB vs. ILT. (C) CaCA2 vs. ILT.

S5B

```
# 1: A_Sequence: Aau1 (2.A.76.1.5 homologue)
# 2: B_Sequence: Eli1 (2.A.108.2.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 104
# Identity:      25/104 (24.0%)
# Similarity:    48/104 (46.2%)
# Gaps:           5/104 ( 4.8%)
# Score: 84.0
#=======================================
                                  4                                   5
A_Sequence         1 KKNESALSMFQRGIWVNLLNPKAIVFFLA-FMPQFIRPDQPLLQQYAVLT     49
                     |  :     |  | |   |   |: :  | ::       :  :: :  |
B_Sequence         1 KAGQDRALRFVHGGWTGALVAGALTWLAATYLLDISGAGRESIEAFGSLI     50
                                  2                                   3
                                                           6
A_Sequence        50 ATVIIIDIMVMWFFFAFAARSFQRFTHDQKGQKVLNRVFGCLFVLVGILL     99
                     | :::: : | |      | ::||:  |: | | |:|  | |: | ||:
B_Sequence        51 AALVLLSVGV-WMHGKSQADNWQRYIRDKLG-KALSR--GSLWFLFGIVF     96
                                                           4

A_Sequence       100 AVIH    103
                      |::
B_Sequence        97 LVVY    100
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     11.31 +/- 5.29
Standard score (Z):      14.0
Precise score (Z):  13.7
```

Supplemental Figure 5. GSAT comparisons with ILT. (A) CadD vs. ILT. (B) RhtB vs. ILT. (C) CaCA2 vs. ILT, cont.

S5C

```
# 1: A_Sequence: Aan1 (2.A.106.1.1 homologue)
# 2: B_Sequence: Sso1 (2.A.108.3.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 82
# Identity:      28/82 (34.1%)
# Similarity:    47/82 (57.3%)
# Gaps:           2/82 ( 2.4%)
# Score: 104.0
#========================================
                     1                    2                         3
A_Sequence         1 IAAILAMKHARLVIFLGAVSALAVMTVLSAAMGYALPALMPRTYTHYASA      50
                     ||||    :   : |: ||  :|:: : :  :|  |  |:|  |   |||
B_Sequence         1 IAAIYHNIYKNNLPFIYAVLGVAIVLIPTFTLG-KLIYLVPLNYVLLASA      49
                     1                    2                         3

A_Sequence        51 LLFFYFGCRMLKDASSMSGSGVSEELGEVEEE      82
                     :: |||| |::: |   |  |: :: || :||
B_Sequence        50 VILFYFGYRLIRSA-RRSFKGIKKKGGEEKEE      80
#----------------------------------------
============ FINISHED =============
Average Quality (AQ)      13.74 +/- 5.89
Standard score (Z):       15.0
Precise score (Z):  15.3
```

Supplemental Figure 5. GSAT comparisons with ILT. (A) CadD vs. ILT. (B) RhtB vs. ILT. (C) CaCA2 vs. ILT, cont.

S6A

```
# 1: A_Sequence: Pre2 (2.A.76.1.5 homologue)
# 2: B_Sequence: Lfr1 (2.A.109.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 120
# Identity:      39/120 (32.5%)
# Similarity:    55/120 (45.8%)
# Gaps:          16/120 (13.3%)
# Score: 91.0
#=======================================
                         1                                          2
A_Sequence         1 NFWTYLAGLVLIIIVPGPNSLYVLKTSTS-SGTRFGYRAAL--GVFTGDG     47
                     | |  :  ||:|  :   ::  ||   |    |:     :|  |:: |
B_Sequence         1 NDWLIIFSLVVIECLLSVDNAVVLAAQTQVLPTKKWQEESLFYGMW-GAY     49
                         1                                          2
                                                   3
A_Sequence        48 ILIFLSFIGV-ASVIKASPTLFMIVRYLGAAYLLYLGCKILYSTFM--HK     94
                     |  ||  |||   :||     | |:: |||||| ||     |:     ||
B_Sequence        50 IFRFL-IIGVGVYLIK-----FWIIKVLGAAYLFYLAFSFFYNMHQNRHK     93
                                                   3

A_Sequence        95 KSNQDGTDTISIKTENHFTR    114
                     ||:   |  :    :||  |
B_Sequence        94 KSH---THQVKPNKKNHTRR    110
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     13.27 +/- 5.74
Standard score (Z):      14.0
Precise score (Z):  13.5
```

Supplemental Figure 6. GSAT comparisons with TerC. (A) RhtB vs. TerC. (B) CadD vs. TerC. (C) LysE vs. TerC (D) MntP vs. TerC. (E) ILT vs. TerC. (F) CaCA2 vs. TerC.

S6B

```
# 1: A_Sequence: Osp1 (2.A.77.1.1 homologue)
# 2: B_Sequence: Bsp1 (2.A.109.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 118
# Identity:      36/118 (30.5%)
# Similarity:    60/118 (50.8%)
# Gaps:          11/118 ( 9.3%)
# Score: 104.0
#=======================================
                         1                                 2
A_Sequence         1 SIDYIVIL--VVLFAQNERRKRAVRDIFLGQYIGFTILIAISLLAAFGLT     48
                     |||   :|  :::  : | ||:|::    | |  |     |||:  |
B_Sequence         1 SIDNAAMLASMIMKLKKEDRKKALKYGIFGAYF-FR---GISLI--FASI     44
                         1                                 2
                                  3                                  4
A_Sequence        49 LIPQHWIGLL-GLVPIFIGLKVLFEKE--DDDDQEEIIDTNRFTSFILSV     95
                     ||   |: || ||  ::||:   |:|:    : :: ||  | |   |:|:
B_Sequence        45 LIKIWWLKLLGGLYLVYIGISHFFKKKLIKKNSKKNIILRNSFWKIIISI     94
                                  3                                  4


A_Sequence        96 AVIMLAAGGDNLGVYIPY    113
                      :: |    ||:   | :
B_Sequence        95 EIMDLTFSIDNIFATIAF    112
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)      16.41 +/- 6.46
Standard score (Z):       14.0
Precise score (Z):  13.6
```

Supplemental Figure 6. GSAT comparisons with TerC. (A) RhtB vs. TerC. (B) CadD vs. TerC. (C) LysE vs. TerC (D) MntP vs. TerC. (E) ILT vs. TerC. (F) CaCA2 vs. TerC, cont.

S6C

```
# 1: A_Sequence: Pfr1 (2.A.75.1.1 homologue)
# 2: B_Sequence: Rpa3 (2.A.109.1.5 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 103
# Identity:      31/103 (30.1%)
# Similarity:    51/103 (49.5%)
# Gaps:          11/103 (10.7%)
# Score: 112.0
#=======================================
                     3                                              4
A_Sequence         1 YLCWFAWRSFRSALRPQSD--DALTGQGPDAGALRPIVGTTL-ALTWLNP     47
                     ::||  ||  ||  :  :|  |||   |  :|| |  :|  :  :|  :
B_Sequence         1 WVCWKMWRELRSQSQHDADALDALNDDGTASGAPRKTLGQAVWQITLADI     50
                     3                                              4
                                                   5
A_Sequence        48 HVYLDTMVMLGGLANQHPGLTRWAFAGGAMLGSALWFAALGLGARALSRP     97
                      : || :: : | | :|| :         : | ||  | :|| |  :::
B_Sequence        51 SMSLDNVLAVAGAAREHPII--------LVFGLALSIALMGLAASFIAKL     92
                                                   5

A_Sequence        98 LSK    100
                     | |
B_Sequence        93 LQK     95
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)      16.06 +/- 6.58
Standard score (Z):       15.0
Precise score (Z):  14.6
```

Supplemental Figure 6. GSAT comparisons with TerC. (A) RhtB vs. TerC. (B) CadD vs. TerC. (C) LysE vs. TerC (D) MntP vs. TerC. (E) ILT vs. TerC. (F) CaCA2 vs. TerC, cont.

S6D

```
# 1: A_Sequence: Lmi1 (2.A.107.1.1 homologue)
# 2: B_Sequence: Ddo1 (2.A.109.5.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 188
# Identity:       48/188 (25.5%)
# Similarity:     94/188 (50.0%)
# Gaps:           29/188 (15.4%)
# Score: 132.0
#=========================================

                          2                                 3
A_Sequence          1 LSQALGIGILFGV---VEATTPLIGWLLGSAASRFVASIDHWVAFVLLAG     47
                      |:  : | :|||:   |  : |   | :   ||     |  | | :|:||
B_Sequence          1 LAMGMRIALLFGISWLVALSAPF--WHIN--ASWITGGIS-WQAVILIAG     45
                          2                                 3

A_Sequence         48 LGIHMVWKSFQPLEPDCDDQTDAPYDTGVQLGADGSALRTGRLLPAGLLS     97
                       || ::|||   :    |       :||::   :  : ::   |   ::
B_Sequence         46 -GIFLIWKSVHEIHEKVD-------ETGLE--EEEISKKSSTTLGNAIVQ     85

                          4                                 5
A_Sequence         98 MLLTSVATSIDAM--AVGVTLAFVDVPIGQVALVIGLCTTMMVTLGVML-    144
                      : : ::  | |::  |||:|    | |    ||:| :   :::::|:|:
B_Sequence         86 IAVINLVFSFDSILTAVGMTNGLSDNPTD--ALII-MVIAVVISVGIMML    132
                          4                                 5
                                           6
A_Sequence        145 -GRLLGTLVGRR--AEMLGGIVLIVIGTVILYE--HLA     177
                          :|  : :    ::||   ||:|| ::: |  ||:
B_Sequence        133 FANPVGNFIAKHPSLQILGLSFLILIGFMLIAEGAHLS     170
                                           6
#-----------------------------------------
============ FINISHED =============
Average Quality (AQ)     26.22 +/- 7.83
Standard score (Z):      14.0
Precise score (Z):  13.5
```

Supplemental Figure 6. GSAT comparisons with TerC. (A) RhtB vs. TerC. (B) CadD vs. TerC. (C) LysE vs. TerC (D) MntP vs. TerC. (E) ILT vs. TerC. (F) CaCA2 vs. TerC, cont.

S6E

```
# 1: A_Sequence: Npe1 (2.A.108.2.1 homologue)
# 2: B_Sequence: Cte1 (2.A.109.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 202
# Identity:      57/202 (28.2%)
# Similarity:    91/202 (45.0%)
# Gaps:          19/202 ( 9.4%)
# Score: 125.0
#=======================================
                                                                    2
A_Sequence         1 SSFVAAFTILVREGLEAI---LIVIAMITFLAKADRRDVLPYVHGGWIAA     47
                     || :  | ::| ||| ::    ::  |:  | :  |   | |   | : |
B_Sequence         1 SSLLVIFNLIVIEGLLSVDNAAVLATMVLDLPQKQRPAALTY---GILGA     47
                                  1                                 2
                                                   3
A_Sequence        48 -LFAGAGTWAAATWLITISGASRELTEGFGGVFAALVLLWVGIWMH-GKS     95
                      || |   : || :|:     |      |||::  | |:|   | : |
B_Sequence        48 YLFRGLFLFFAA-FLV-----SAWWLRPFGGLY-LLYLVW-NWWNNRGSK     89
                                                   3
                                                             4
A_Sequence        96 NADAWQRYIRD-KLGRALNRRSAWFLFALAFIVVYREVFETILFYAAIWS    144
                     : ||     || :| | ::||   |   : |: :    |     :||:
B_Sequence        90 DGDAMCTEKRDNRLYRFVSRRIGPFWATVLFVEMMDIAFSIDNVFAAVAF    139
                                                             4
                                  5                                 6
A_Sequence       145 QGNGGAVVAGAFAAIAVLAVIAFVMLRHSRTLPIGKFFAYSSALIAVLAV    194
                       |   |  | |  | |:  :|:  :|     |  :  ||   ::||| :
B_Sequence       140 TDNLILVCTGVFIGILVMRFVAYGFIRLMEEYPFLESCAY--IVLAVLGL    187
                                  5                                 6


A_Sequence       195 VL    196
                      |
B_Sequence       188 RL    189
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     22.32 +/- 7.82
Standard score (Z):      13.0
Precise score (Z):  13.1
```

Supplemental Figure 6. GSAT comparisons with TerC. (A) RhtB vs. TerC. (B) CadD vs. TerC. (C) LysE vs. TerC (D) MntP vs. TerC. (E) ILT vs. TerC. (F) CaCA2 vs. TerC, cont.

S6F

```
# 1: A_Sequence: Ptr2 (2.A.106.1.1 homologue)
# 2: B_Sequence: Lga1 (2.A.109.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 194
# Identity:      53/194 (27.3%)
# Similarity:    85/194 (43.8%)
# Gaps:          31/194 (16.0%)
# Score: 111.0
#=======================================
                                  1                                 2
A_Sequence         1 GGFWNAFTSSVAMIIATEIGDKTFFIAA---VLSMK-HSRSAVFFGAILA     46
                     |  |    : : |     : |    :||   ||  |   | :: :| :
B_Sequence         1 GQDWMMILTLILMECLLSV-DNAVVLAAQTQVLPTKDEQRKSLVYG-LWG     48
                                  1                                 2

                                                   3
A_Sequence        47 LIVMTVLSTAMGMMLPNFIPKEYTHLLGGLLFLYFGCKLIYDSRQMEAGK     96
                       :   :   :|  | ||       |||||  ||   |  || |
B_Sequence        49 AYLFRFIVIGIGTYLINFWE---IKLLGGLYLLYLVYKYFYDVRHP----     91
                                                   3

                                                                      4
A_Sequence        97 TSEELEEVEEELLQQGKKKADLEEGSRSNRPPSKKQMGWNQVV-IQSLTL    145
                           :| ::  :  ||:|  :: |:: :      : |  |: |:|: :
B_Sequence        92 -----AQVAKK--EAAKKEAHKKKNSKTRK--HHLSLFWRTVISIESMDI    132
                                                                      4

                                                   5
A_Sequence       146 TFVAEWGDRSQIATIALAASKNPIGVTIGGCVGHSLC-TGLAVV    188
                      |  :    | :|  ||| | ||: | :|| :|  ||  |:| |
B_Sequence       133 VFSID----SVLA--ALAMSNNPVVVLVGGMIG-ILCMRGVAEV    169
                                                   5

#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     13.76 +/- 5.99
Standard score (Z):      16.0
Precise score (Z):  16.2
```

Supplemental Figure 6. GSAT comparisons with TerC. (A) RhtB vs. TerC. (B) CadD vs. TerC. (C) LysE vs. TerC (D) MntP vs. TerC. (E) ILT vs. TerC. (F) CaCA2 vs. TerC, cont.

S7A

```
# 1: A_Sequence: Spl1 (2.A.75.1.1 homologue)
# 2: B_Sequence: Ogr1 (2.A.95.1.3 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 157
# Identity:      50/157 (31.8%)
# Similarity:    74/157 (47.1%)
# Gaps:          14/157 ( 8.9%)
# Score: 124.0
#========================================
                         1                         2
A_Sequence         1 LIVAIGAQNTFV-LTQGIRKQHRFVVALICSL-CDAFLISAGVAG--LGS     46
                     :|  ||    || ||||:  : |  :|:  :|   | |::  | |  |
B_Sequence         1 IIDPIGLTPLFVALTQGMPDRQRRAIAVRATLVAVAVLLAFAVFGEALLG     50
                         1                         2
                                  3
A_Sequence        47 LIEQSPTLLRLAGGGGALFLFIYGLKCLFSALQAEQELGETESNPTSRRQ     96
                      :  |    |:|||   : ||:  |  ||   || :|  :|  :||
B_Sequence        51 FVGISMAAFRIAGG---VLLFLTALDMLFQRRQARRE--DTADDPTEDPS     95
                                  3
                          4                                         5
A_Sequence        97 VILTILAI-TLCNPNVYLDTVVLLGGISATFVGQGRYLFGAGAISMSFIW    145
                     |    |||  |  |   : |::|| | | :  |    | |     :: ::
B_Sequence        96 VF--PLAIPLLAGPGA-IATIILLTGQSESVAGFAAVL-GVMVAVLTIVF    141
                          4                                         5

A_Sequence       146 FFILSYG    152
                      | |: |
B_Sequence       142 LFFLAAG    148
#----------------------------------------
============ FINISHED =============
Average Quality (AQ)      20.31 +/- 7.39
Standard score (Z):       14.0
Precise score (Z):  14.0
```

Supplemental Figure 7. GSAT comparisons with NAAT. (A) LysE vs. NAAT. (B) RhtB vs. NAAT. (C) CadD vs. NAAT (D) MntP vs. NAAT. (E) TerC vs. NAAT.

S7B

```
# 1: A_Sequence = Pag1 (2.A.76.1.2 homologue)
# 2: B_Sequence = Cba1 (2.A.95.1.4 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 167
# Identity:      44/167 (26.3%)
# Similarity:    79/167 (47.3%)
# Gaps:          15/167 ( 9.0%)
# Score: 115.0
#=======================================
                         1                                 2
A_Sequence         1 ALVHLVALMSP-GPDFFFVS-QTAASRSRKEAMMGVLGITLGIVVWAGV-     47
                     : : |:||::| |   ||:|  |  :   |   : :  |::  ||  |:
B_Sequence         1 SFISLLALINPIGAIPFFISLTTQQTEEEKRHTIKIAAISVATVV--GIS     48
                         1                                 2

                                          3
A_Sequence        48 ALMGLHLILEKMAWLHQVIMVGGGLYLLWMGWQLMCSARQRHKQPQQDEP     97
                     ||:|   |:|        : ||||| :: |   :: :   | |   ::|
B_Sequence        49 ALLG-QQIIEFFNISVASLQVGGGLIMIMMAMNMLNAQTSRTKATPEEED     97
                                          3

                                   4
A_Sequence        98 VVELPKRGMSFLKGLLTNLSNPKAIIYFGSVFSLFVGDDVGSAERWGLFL    147
                       |  |  :: :   |  |: |      ||: :: |    |  : |   |
B_Sequence        98 EAE-AKASIAVVPLALPLLTGP------GSISTVIV--YAGKTQHWYQLL    138
                                   4

                         5
A_Sequence       148 LIIGETFAWFALVAAIF    164
                     :::|   |  |:|  :|
B_Sequence       139 ILVGIGVALGAVVYIVF    155
                         5
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     17.31 +/- 6.50
Standard score (Z):      15.0
Precise score (Z):  15.0
```

Supplemental Figure 7. GSAT comparisons with NAAT. (A) LysE vs. NAAT. (B) RhtB vs. NAAT. (C) CadD vs. NAAT (D) MntP vs. NAAT. (E) TerC vs. NAAT, cont.

S7C

```
# 1: A_Sequence: Msp1 (2.A.77.1.4 homologue)
# 2: B_Sequence: Orf7 (2.A.95.1.4 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 152
# Identity:      42/152 (27.6%)
# Similarity:    68/152 (44.7%)
# Gaps:           9/152 ( 5.9%)
# Score: 97.0
#=======================================
                     1                             2
A_Sequence         1 FLLAAFFANPEFRAKDVVLGQYLGFIVLLT--ISSLAYFVQF--IIPSNW     46
                     ||      ||: | |         |:|| |  |:    |  |   :|:
B_Sequence         1 FLAVTTGQNPQKRRKTARKASLTAFVVLTTFAIAGTFIFKMFGITLPAFE     50
                     1                             2
                         3                                          4
A_Sequence        47 ISLLGVIPIMIGIRSFLHLKK-PQTDYSGENRDFSKYKEGQMMLPVTLVT     95
                     |:  ||| ::||:   |  |: |  : |||  : :  ||   ::|: :
B_Sequence        51 IA-GGVILLLIGL-DMLEAKRSPTQESSGETAEAAS-KEDVGIVPLGIPM     97
                         3                                          4
                                                   5
A_Sequence        96 LANGGDNLGVYMPLFASMGPFDL-FLTAIIFLIMVGVWCFLGYKLVNNRV    144
                     ||  |    | : :  :  |: :  : | | :  |  :  ||      |:
B_Sequence        98 LAGPGAITSVMVLVGQAQNPWQVGTIIAAIAITAVSCYVVLGAATRVARI    147
                                                   5

A_Sequence       145 LG    146
                     ||
B_Sequence       148 LG    149
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     13.03 +/- 5.85
Standard score (Z):      14.0
Precise score (Z):  14.4
```

Supplemental Figure 7. GSAT comparisons with NAAT. (A) LysE vs. NAAT. (B) RhtB vs. NAAT. (C) CadD vs. NAAT (D) MntP vs. NAAT. (E) TerC vs. NAAT, cont.

S7D

```
# 1: A_Sequence: Asu1(2.A.107.2.1 homologue)
# 2: B_Sequence: Csh1(2.A.95.1.5 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 124
# Identity:      33/124 (26.6%)
# Similarity:    64/124 (51.6%)
# Gaps:           6/124 ( 4.8%)
# Score: 108.0
#=======================================
                                 3                                 4
A_Sequence         1 YISEFDHWIA-FALLCVIGINMIKMSVTNENSDDDPSD--FSL-RHLTML     46
                     ::||    ||   :| :| : |:     : |  | |:|  | :   : ::
B_Sequence         1 HLSETSLGIAGGVILFLIALRMV-FPAPHGNGADHPADEPFVVPLAIPLI     49
                                 3                                 4

                                                  5
A_Sequence        47 GVATSIDALAMGVSFAFLKVNIWTAAAIIGITTTILSL-FGVKAGHWLGD     95
                        :::  : : ||    ::  | ||  : :    |:| |  |  ||||:
B_Sequence        50 AGPSALATVLLLVSREPARLWEWVAALALTMVVCALTLAFAEKISHWLGE     99
                                                  5

                                 6
A_Sequence        96 RIHKQAELLGGIILIAMGVKVLIE    119
                     |:    | | |::| |: |::|::
B_Sequence       100 RVTTAFERLMGLVLTAIAVQMLLD    123
                                 6
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     16.70 +/- 6.06
Standard score (Z):      15.0
Precise score (Z):  15.1
```

Supplemental Figure 7. GSAT comparisons with NAAT. (A) LysE vs. NAAT. (B) RhtB vs. NAAT. (C) CadD vs. NAAT (D) MntP vs. NAAT. (E) TerC vs. NAAT, cont.

S7E

```
# 1: A_Sequence: Gka1 (2.A.109.1.5 homologue)
# 2: B_Sequence: Dgi1 (2.A.95.1.5 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 106
# Identity:      29/106 (27.4%)
# Similarity:    52/106 (49.1%)
# Gaps:           6/106 ( 5.7%)
# Score: 93.0
#=======================================
                          1                                  2
A_Sequence          1 LIIGIDVILGGDNAVVIALAS-RNLPEQKRNVAIIVGTALAIAVRIVLTV       49
                      | : : :|: |   | : ::  |  | :::   |      |:|: |:
B_Sequence          1 LAVPLFLIMDGLGNVPVCMSMLRRFPPRRQQRIIFRELCFALAISILFCF       50
                          1                                  2
                                            3
A_Sequence         50 AVVWLLTI----P-FLQLAGGVVLFWIALKLIGQKDEKPTMIKAEPSLWK       94
                         |||      |  |:|||||||| |:::::   : | |    :||
B_Sequence         51 FGDWLLKFLGLGPSTLRLAGGVVLFVISMRMVFPDESKETADPEDPSALA      100
                                            3
                          4
A_Sequence         95 AIQTIV    100
                      | :  :
B_Sequence        101 AEEPFI    106
                          4
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)      12.39 +/- 5.29
Standard score (Z):       15.0
Precise score (Z):  15.2
```

Supplemental Figure 7. GSAT comparisons with NAAT. (A) LysE vs. NAAT. (B) RhtB vs. NAAT. (C) CadD vs. NAAT (D) MntP vs. NAAT. (E) TerC vs. NAAT, cont.

S8A

```
# 1: A_Sequence: Aur1 (2.A.76.1.2 homologue)
# 2: B_Sequence: Bco1 (2.A.113.1.9 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 209
# Identity:      52/209 (24.9%)
# Similarity:    93/209 (44.5%)
# Gaps:          18/209 ( 8.6%)
# Score: 128.0
#========================================
                         1                                          2
A_Sequence         1 FIALITLMFIQFCALITPGPDFFLVSQTAISRSRREAVFVVLGITVGVMF     50
                     ||::: | |:         ||  :   |  |||:: :   : |:  |:
B_Sequence         1 FISVLALGFVLGIKHAIE-PDHIIAVSTIASRSKKLSQSSLAGVFWGIGH     49
                         1                                          2

                                                   3
A_Sequence        51 WAILALMGLNIIFEK----MAWLKQILLVIGGIYLCWLGFQMLRSAFSKQ     96
                      | | ::|: ::  |      |   :  :: || | :||   | ||| :
B_Sequence        50 TATLFIVGICLLIIKGEIPEKWAMSLEFLV-GIMLVYLGITTL-SAFKRV     97
                                                   3

                                                           4
A_Sequence        97 KVQNTNTPIDLPKTETKF-FLKGLLTNLSNPKAVIYFGS-VFSLFLANPA    144
                     ::   |     |  :  : ::| :     :  |    ||    |   :
B_Sequence        98 RI---NHHYHEPGHKRNYSYIKSVCIGFVHGLA----GSGAMVLLTMSTV    140
                                                           4

                                          5
A_Sequence       145 LDHVHSLLFIII-AVETLIWFLFVVFVFSLPSFKSAYQ-NVAKWIDGVSG    192
                        | | ::|:|  : |:   ||   :  :|   || :  | | :  ::|
B_Sequence       141 KSVVESAIYILIFGIGTIFGMLFFTTILGIPFIISAKKVEVNKTLTQITG    190
                                          5

                           6
A_Sequence       193 GIFTAFGIY    201
                      | | ||||
B_Sequence       191 AISTVFGIY    199
                           6
#----------------------------------------
============ FINISHED =============
Average Quality (AQ)      23.09 +/- 7.61
Standard score (Z):       14.0
Precise score (Z):  13.8
```

Supplemental Figure 8. GSAT comparisons with NicO. (A) RhtB vs. NicO. (B) CadD vs. NicO. (C) TerC vs. NicO (D) NAAT vs. NicO.

S8B

```
# 1: A_Sequence = Acy3 (2.A.77.1.4 homologue)
# 2: B_Sequence = Gar1 (2.A.113.1.9 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 195
# Identity:      55/195 (28.2%)
# Similarity:    94/195 (48.2%)
# Gaps:          23/195 (11.8%)
# Score: 129.0
#=======================================
                           1                                        2
A_Sequence         1 NIITPILTGV-FAFI-ATNIDDIVILLVFFSQVNEN--FRPWQIVMGQYL     46
                     : :  :|||: |  | | ::| || :  |  | |:|     :    |
B_Sequence         1 DFLAAVLTGIMFGIIHAFDVDHIVAMATFSEQKNKNKQILTYAFKWGTGH     50
                           1                                        2
                                                           3
A_Sequence        47 GFTILVIFSLPGFFGGLILPPAWIG----LLGLIPIGIGISSLVNKEKEQ     92
                     |  ||::  :   | |  ||  ::     ::|:: | :|:  ||   ::
B_Sequence        51 G-GILLLLGMLLIFIGFQLPNWFVHYSEIMVGVLLIYLGVKLLVLLHRKG     99
                                                           3
                                                   4
A_Sequence        93 LADVPEEIISPATSINNYSLTPQIYTVAAITVANGSDNISIYIPLFSSIS    142
                        ||| :   | |:| :  || ::    | : :|    :  : |  ::
B_Sequence       100 TFSVPESLDLAARSLNKHDHTP-LF----IGMLHGVAGSAPLLALLPNML    144
                                                   4
                                  5                             6
A_Sequence       143 FNSFLLIIGLFFF--LLGVWC--YV--AYQL-THQK--KVADFFT    178
                        ||| | ||    | |::|  |:  :||:   ||  |:|  ||
B_Sequence       145 ETQFLLHISLFSIGCLFGMFCFGYIFGSYQVYIKQKKEKLAKAFT    189
                                  5                             6
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)      18.67 +/- 7.32
Standard score (Z):       15.0
Precise score (Z):  15.1
```

Supplemental Figure 8. GSAT comparisons with NicO. (A) RhtB vs. NicO. (B) CadD vs. NicO. (C) TerC vs. NicO (D) NAAT vs. NicO, cont.

S8C

```
# 1: A_Sequence: Bdi1 (2.A.109.1.5 homologue)
# 2: B_Sequence: Cul1 (2.A.113.1.9 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 196
# Identity:      55/196 (28.1%)
# Similarity:    83/196 (42.3%)
# Gaps:          19/196 ( 9.7%)
# Score: 117.0
#=======================================
                                  1                        2
A_Sequence         1 ASALGKVLMIDLVLAGDNAVAVGLAAAALPQEQRRKAILIGLA-----AA     45
                     |  :| :| :   |  |: |||  :  || :    :   ||       |
B_Sequence         1 AMGIGILLGLRHALDADHVVAV--STMALEERNLLRGGWIGFCWGVGHAL     48
                                  1                        2
                                          3
A_Sequence        46 VVMRIGLALIT--VQLLAIVGLLLAG--GFLLLWVCWKMWRELREQATHD     91
                     |:   | |||   ::|  :||  | |  | :|: :    || :|    |
B_Sequence        49 VLFLFGGALILSGIRLPEVVGRWLEGGVGVMLILIALGSWRRMRRSKLHI     98
                                          3
                                                                    4
A_Sequence        92 QAEAEAEIERAMAIEHGGGPSPEEL----GLK-RKTFGAALIQIMIADLT    136
                         : : ||     |    ||  |    | |   :|    :  :
B_Sequence        99 HVH-QHDGERYHTHFHVHDDSPRHLEKHHGWKGSHSFLIGTVHGLAGTGA    147
                                                                    4
                                          5
A_Sequence       137 MSLDNVLAVAGASHEHPWIMVFGL--ILSIALMGLAATFIAKLLNR    180
                     : :  : ||:       ::  |||  |||: |  |: | | |||||
B_Sequence       148 VMVLTIAAVSDPLQRIAYLASFGLGTILSMTLFSLSLTLITKLLNR    193
                                          5
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     18.97 +/- 7.06
Standard score (Z):      14.0
Precise score (Z):  13.9
```

Supplemental Figure 8. GSAT comparisons with NicO. (A) RhtB vs. NicO. (B) CadD vs. NicO. (C) TerC vs. NicO (D) NAAT vs. NicO, cont.

S8D

```
# 1: A_Sequence: Mfo1 (2.A.95.1.4 homologue)
# 2: B_Sequence: Orf5 (2.A.113.2.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 128
# Identity:      33/128 (25.8%)
# Similarity:    66/128 (51.6%)
# Gaps:          11/128 ( 8.6%)
# Score: 109.0
#=======================================
                                                                    4
A_Sequence         1 FKIAWDMLHAEMSKTKHSPREEIDMRMGSVAVVPLAIPLLAGPGAITTTI     50
                     :: | :  |||  : : : ||   :  | : :  |   |:  |||||  :
B_Sequence         1 YQDAHERQHAEDIRRRFAGRE---VTTGQIILFGLTGGLIPCPGAITVLL     47
                                                                    4

                                                   5
A_Sequence        51 ILME-KAQSLANKTIVISSI--ILTMIVSGLILSASDIVVKKLKVSGINA     97
                     : :: |  :| :  ::  ||   |||: || :::|  :   : : ||  :
B_Sequence        48 LCLQLKRVALGSVLVLCFSIGLALTMVASG-VIAALSVKYAERRFSGFGS     96
                                                   5

                                          6
A_Sequence        98 IVR----IMGLILAAISVQIIFSGAYGL    121
                     :||      ||::  : : :  || : |
B_Sequence        97 LVRKAPYASGLVILCVGLYVALSGWHSL    124
                                          6

#---------------------------------------
============ FINISHED =============
Average Quality (AQ)      18.86 +/- 6.66
Standard score (Z):       14.0
Precise score (Z):  13.5
```

Supplemental Figure 8. GSAT comparisons with NicO. (A) RhtB vs. NicO. (B) CadD vs. NicO. (C) TerC vs. NicO (D) NAAT vs. NicO, cont.

S9A

```
# 1: A_Sequence: Hgr1 (2.A.76.1.5 homologue)
# 2: B_Sequence: Ssp3 (2.A.116.1.7 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 188
# Identity:      60/188 (31.9%)
# Similarity:    91/188 (48.4%)
# Gaps:          27/188 (14.4%)
# Score: 138.0
#=======================================
                          1                                 2
A_Sequence         1 IGIVLLPGPNSLFVLSVATA-RGVRVGYHAACGVF----LGDSILL-LFT      44
                     : | ::  ||  :  : :||| | |||      ||     || :|:| : |
B_Sequence         1 LAITMMAGPQIMSAVILATAQRAVRVSLGFVTGVLIATSLGVAIMLGIAT      50
                          1                                 2

                                           3
A_Sequence        45 ALGAA---SLLRGYPALFMVVKYVGAAYLFWVGMNLAWSAWRKWRAAGIA      91
                     ||| |          ::  |::||  | |      ||  | | ||
B_Sequence        51 ALGGAVDFGSSGDKSSVGRVIQYVLVALLI-----LA--ALRNWR----K      89
                                           3

                                                            4
A_Sequence        92 TQLVEPTA-LAAAQSAHLLAPFQRALVISLLNPKAILFLLSFFVQFIDPA     140
                      : |||   | |  ||     |:  |:: || |  :: :|:  |  :|
B_Sequence        90 RETVEPPKWLHALMSADTRKAFETGLLVVLLMPSDLMVMLTVGVH-LDQG     138
                                                            4

                                           5
A_Sequence       141 YDT--PAIPFLILSVIVMAFSAVYLSVLIVAGARLADA     176
                     : :   |:||: |: :| | : : | ||:  | | | |
B_Sequence       139 HSSFVDALPFIALTTLVAA-TPLLLRVLL--GRRAASA     173
                                           5

#---------------------------------------
============ FINISHED =============
Average Quality (AQ)      26.84 +/- 7.65
Standard score (Z):       15.0
Precise score (Z):  14.5
```

S9 Fig. GSAT comparisons with GAP. (A) RhtB vs. GAP.

S10A

```
# 1: A_Sequence: Btr2 (2.A.76.1.2 homologue)
# 2: B_Sequence: Cba1 (5.A.1.2.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 196
# Identity:      52/196 (26.5%)
# Similarity:    90/196 (45.9%)
# Gaps:          24/196 (12.2%)
# Score: 134.0
#=======================================
                                                   2
A_Sequence         1 YVCRKAMADSRRNAMLGALGIALG--VGFWAVIVLFGLT--FLNHTIPNF      46
                     |:    : :  ::  | ||   ||  :||  : ::||:   |:      :
B_Sequence         1 YITGTTLEEELQDKKLFALSRTLGFVLGFTIIFMIFGILAGFVGQAFIRY      50
                                                   2

                                  3
A_Sequence        47 QFYLMLLGGSYLAYCGIKMVQVRKSVEIDENLKSQANEKSPL----W-KE      91
                     :  |  :||  :   |: || : |     | |  | | :||     |
B_Sequence        51 RNVLTKIGGIIIVLFGLNMVGLLKL----EFLNKQRNVRSPKEVKNWFSS      96
                                  3

                         4                                          5
A_Sequence        92 ILGGLAINLS-NPKVVVFFSSVL--AGYVANISAFKDILAVLAILMGSTL     138
                     || |:|      | :     ::|   |  | :|  | |: :||  :|
B_Sequence        97 ILMGMAFAAGWTPCIGPVLGTILIYVGTTATVS--KGIILLLAYSIG-LA     143
                         4                                          5

                                                           6
A_Sequence       139 IWFWTVAILFSQNKIRRFYAKNNR---YLDNAAGVVFILFGLKLIY     181
                     | |   |:|   |:  :|  |: :   |:   :||| |: |: :::
B_Sequence       144 IPFLLTALLI--NQFSKFLMKSEKVLPYIVKISGVVIIVVGVLIVF     187
                                                           6

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      23.74 +/- 7.86
Standard score (Z):       14.0
Precise score (Z):  14.0
```

Supplemental Figure 10. GSAT comparisons with DsbD. (A) RhtB vs. DsbD. (B) CaCA2 vs. DsbD. (C) MntP vs. DsbD. (D) NAAT vs. DsbD. (E) GAP vs. DsbD.

S10B

```
# 1: A_Sequence = Sne3 (2.A.106.1.2 homologue)
# 2: B_Sequence = Orf5 (5.A.1.2.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 161
# Identity:      52/161 (32.3%)
# Similarity:    82/161 (50.9%)
# Gaps:          22/161 (13.7%)
# Score: 130.0
#=======================================



                         1                                          2
A_Sequence         1 LLGIILAFLIVDGIAILAGEWITDIAPRELIKMLSGAIFI-IFGLVTLIF     49
                     |:|  : |: :   : | | :      ::|:: : ||||| ||||: | |
B_Sequence         1 LVGFSVIFIFLGYSSSLVGTFFYQY--QDLLRQI-GAIFIVIFGLMILGF     47
                         2                                          3

                                                   3
A_Sequence        50 RNKREEIK-TKYHFEN-P--FYSGFI--LIFVSEWGDKTQIATG---LFA     90
                        :  :|  |  |:| |  ::  |:  | | : |   |   ||   : |
B_Sequence        48 FTPKFLMKEKKLQFKNRPAGYFGTFLIGLAFAAGWTPCTGPITGAVFMMA     97
                                                   4

                                          4
A_Sequence        91 TQYNG------LMVLTGVIIALSLLSVIAIYSGKFISDKVTRETLTKLTG    134
                      |  |      |: : |  |   |||:  |   |:|  |  | |:||: |
B_Sequence        98 AQNPGSGMWYMLVYVLGFAIPFFLLSIF-ITRVKWI-QKYNR-TITKVGG    144
                                  5

                         5
A_Sequence       135 FLFISMGVLFF    145
                     :| |::|:| |
B_Sequence       145 YLMIALGILLF    155
                         6
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     26.20 +/- 7.85
Standard score (Z):      13.0
Precise score (Z):  13.2
```

Supplemental Figure 10. GSAT comparisons with DsbD. (A) RhtB vs. DsbD. (B) CaCA2 vs. DsbD. (C) MntP vs. DsbD. (D) NAAT vs. DsbD. (E) GAP vs. DsbD, cont.

S10C

```
# 1: A_Sequence: Cac1 (2.A.107.2.1 homologue)
# 2: B_Sequence: Dsp2 (5.A.1.2.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 154
# Identity:      47/154 (30.5%)
# Similarity:    79/154 (51.3%)
# Gaps:          14/154 ( 9.1%)
# Score: 146.0
#=======================================
                         2                                 3
A_Sequence         1 FAISFGFFQFLCTFIGAYSGFLFNTYITYVPQIIGGMIIAFVGAFM---I     47
                     | : |    |   |  :: |  |: |   : :::||::|| :| ||   |
B_Sequence         1 FILGFSIIFFALGFSASWVGSFFSEYRDLI-RMLGGVLIAVMGLFMLGLI     49
                         2                                 3

                                                   4
A_Sequence        48 KEGFDNKEEKLLLNFKMYFVLGISVSIDAAVVGFT-MFNKISSNYVILGD     96
                     | ||  ||::| :  | :  || ||   |   |:|     |  : : |
B_Sequence        50 KPGFMMKEKRLEVGRKRWGYLGSSVIGMAFAAGWTPCVGPILVSVLALAA     99
                                                   4

                                  5                                 6
A_Sequence        97 S------VFIGIVTLILSIIAFIISRYLKRIQLVCKYADYI---GGIILV    137
                     |       :|   ||  :|  ||:: :| | : : ||:: :   || ::|
B_Sequence       100 SNPSAGLAYITAYTLGFAIPFFIMAFFLGRTRWILKYSNSLMKAGGALMV    149
                                  5                                 6


A_Sequence       138 IFGL    141
                     :||:
B_Sequence       150 VFGV    153

#---------------------------------------
============ FINISHED =============
Average Quality (AQ)      28.64 +/- 7.96
Standard score (Z):       15.0
Precise score (Z):  14.7
```

Supplemental Figure 10. GSAT comparisons with DsbD. (A) RhtB vs. DsbD. (B) CaCA2 vs. DsbD. (C) MntP vs. DsbD. (D) NAAT vs. DsbD. (E) GAP vs. DsbD, cont.

S10D

```
# 1: A_Sequence: Pfu1 (2.A.95.1.4 homologue)
# 2: B_Sequence: Dto1 (5.A.1.2.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 213
# Identity:      57/213 (26.8%)
# Similarity:   104/213 (48.8%)
# Gaps:          25/213 (11.7%)
# Score: 140.0
#=======================================
                         1                                          2
A_Sequence           1 GLFAITNP--IGAVPVFIAVTRNLSPEKRKEIARKTSTTVMVTLLVFALV     48
                       || :  :|  :  :| : :    || :: ||  |:|   | ::  || :
B_Sequence           1 GLLSFFSPCILPLIPAYFSFITGLSLDELKENKRQTRQKVFLS-TVFYVA     49
                         1                                          2

                                                           3
A_Sequence          49 G-EWIFKFFGASTDAFS------------IAGGIILFRMSLEMLSGKLSS     85
                       |  :||  ||||                 | |||||    | :| | ::
B_Sequence          50 GFSFIFILFGASASFLGGLASQYAWVVRYIGGGIILV-FGLHLL-GIINI     97
                                                           3

                                                   4
A_Sequence          86 VKIS-EEEEHISEEAVTLEEVAIIPLAIPLLSGP--GAITTTMLYMAKSS    132
                          : |:: |: |: : |    :| :|      |  | :  ::| :| :
B_Sequence          98 KGFNFEKKIHVKEKPLHLMGTFVIGMAFGAGWSPCIGPLLGSILIVAGNQ    147
                                                   4

                                  5
A_Sequence         133 TMIEKSIVLLVVVAIGITV-WIILSA-ANRIHQ--KLGTIGIKVMTRMMG    178
                         : | : || | : |: | ::|||   | | :  |  |  |:|: :: |
B_Sequence         148 ETVLKGVFLLAVYSAGLAVPFLILSVFINSILEIMKRATKFIRVLNKISG    197
                                  5

                         6
A_Sequence         179 LILASMAVQMVIN    191
                       ::| :: : :| :
B_Sequence         198 ILLIAIGLLLVFD    210
                         6
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     22.82 +/- 7.67
Standard score (Z):      15.0
Precise score (Z):  15.3
```

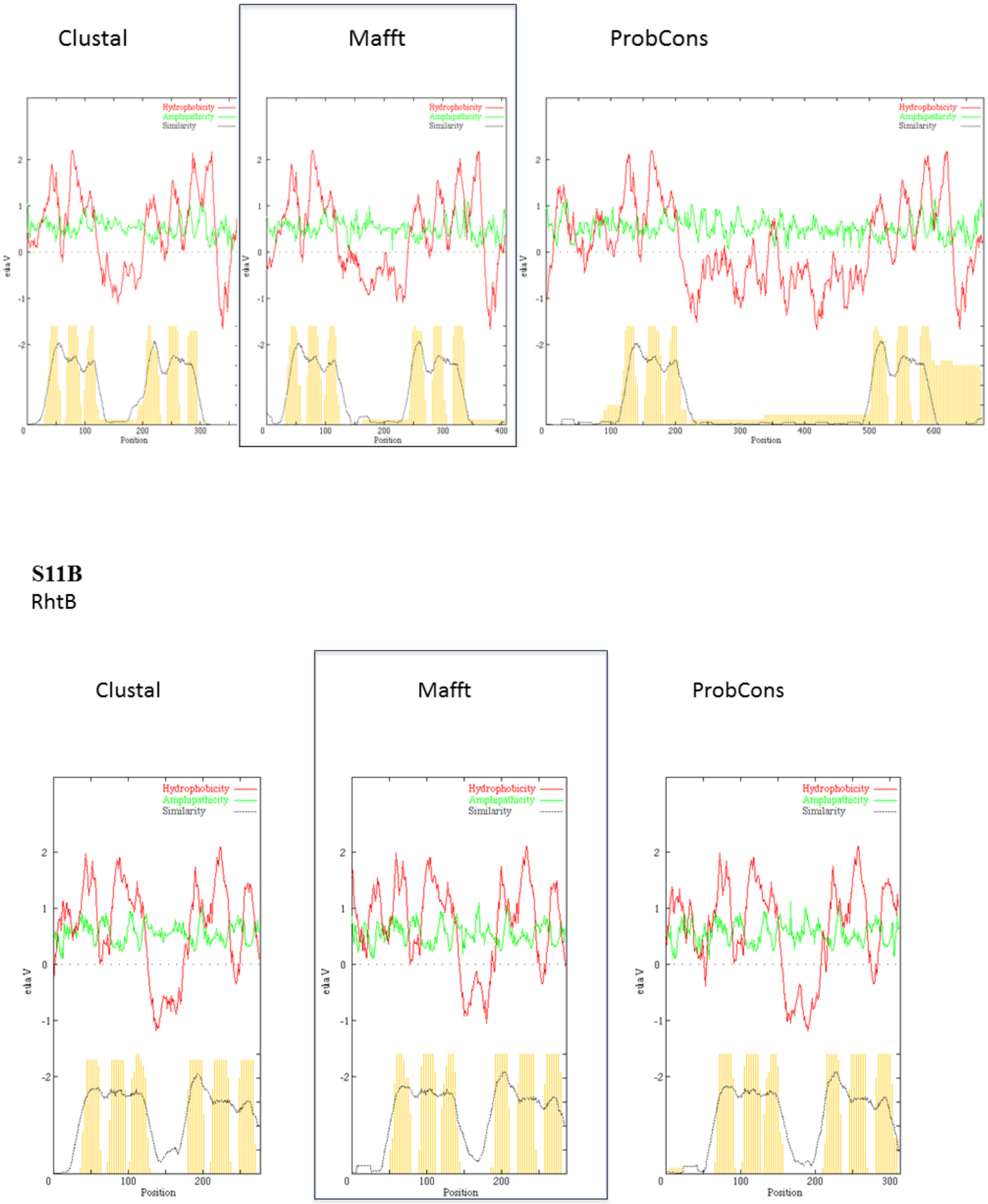
Supplemental Figure 10. GSAT comparisons with DsbD. (A) RhtB vs. DsbD. (B) CaCA2 vs. DsbD. (C) MntP vs. DsbD. (D) NAAT vs. DsbD. (E) GAP vs. DsbD, cont.

S10E

```
# 1: A_Sequence: Sni1 (2.A.116.1.4 homologue)
# 2: B_Sequence: Psp5 (5.A.1.2.1 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0

# Length: 194
# Identity:      46/194 (23.7%)
# Similarity:    97/194 (50.0%)
# Gaps:          15/194 ( 7.7%)
# Score: 117.0
#=======================================
                                  2                                3
A_Sequence         1 KPRPTSLAFLAGWVVGLVGLTVVFIEASSLAGGEQHTRPAWMSWVRIILG     50
                     |  |  |  ::  |: ::| :::|:                |  |: | |
B_Sequence         1 KNSPNKLTVISQTVLFILGFSILFVLLGISVSTVSRLLSEHMRLVQQIGG     50
                                  2                                3
                                                             4
A_Sequence        51 AALIVFGVYRF-VTRHR--HTEQPRWMRPFAKLTPG-RAG--LTGVAVAV     94
                     | ::|||::   : | :  ::|: |:: |    :|| :||  : |:| ||
B_Sequence        51 ALIVVFGLHMTGLLRIKLLYSEK-RYL-PSG--SPGKKAGALVLGMAFAV     96
                                                             4
                                                                    5
A_Sequence        95 VRPEVLALVATAGLEIGAGGLSTAGAWTCGVLFIAVAASTVAIPVLAYAI    144
                          :  : :: | | || ::| |    ||| ::: |  :|:| |  |:
B_Sequence        97 GWTPCIGPILSSIL-IYAGSMATLGK---GVLLLSMYALGLAVPFLLSAV    142
                                                                    5
                                                   6
A_Sequence       145 AGERLDPTMARIKDWMDRNLGAMEAVVLVVIGLMVIEKGISSLS    188
                       : |   : ::   : : :     ||::::|::|    :   |
B_Sequence       143 LIDNLTAYLRKVTKHLPK-ISVASGVVMMLMGVLVFTNQLEVFS    185
                                                   6
#---------------------------------------
============ FINISHED =============
Average Quality (AQ)     21.78 +/- 7.27
Standard score (Z):      13.0
Precise score (Z):  13.1
```

Supplemental Figure 10. GSAT comparisons with DsbD. (A) RhtB vs. DsbD. (B) CaCA2 vs. DsbD. (C) MntP vs. DsbD. (D) NAAT vs. DsbD. (E) GAP vs. DsbD, cont.

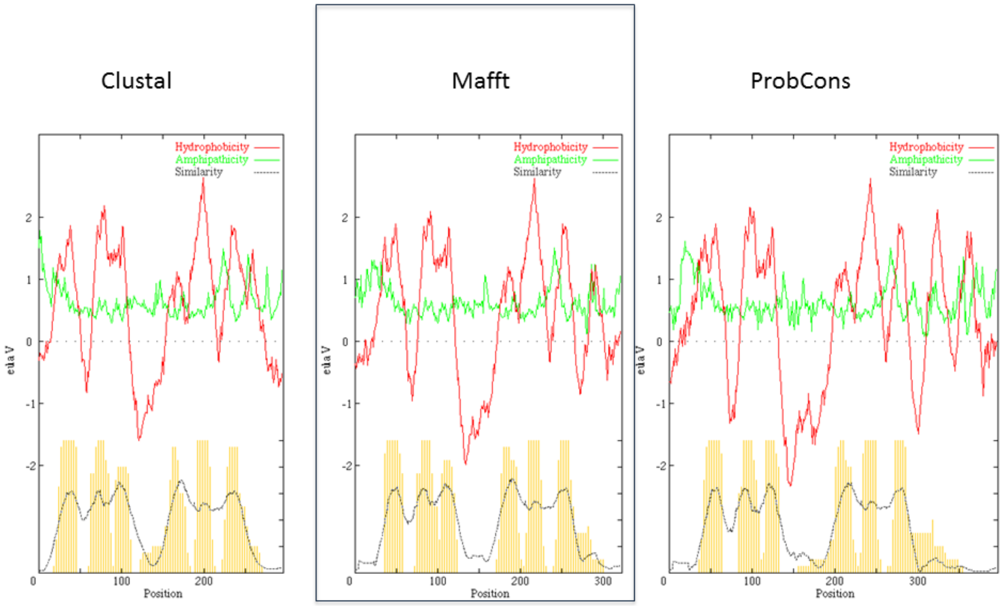
S11A
LysE



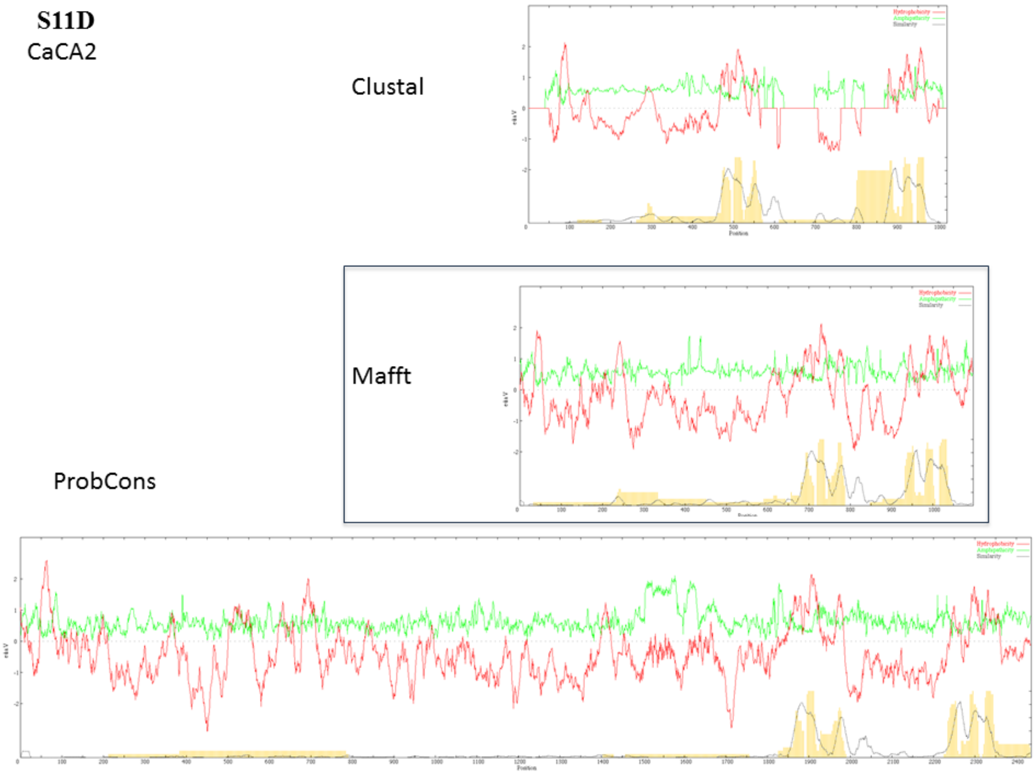
S11B
RhtB

Supplemental Figure 11. AveHAS plots of each family based on multiple alignments generated using three different programs. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) NAAT. (G) NicO. (H) GAP. (I) DsbD. (J) ILT. (K) TerC.

S11C
CadD

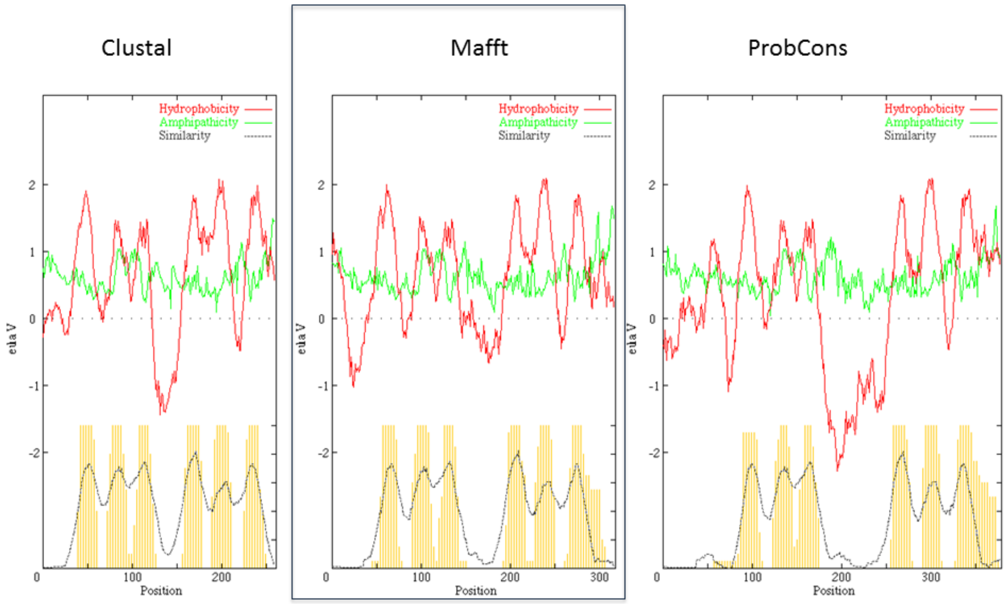


S11D
CaCA2

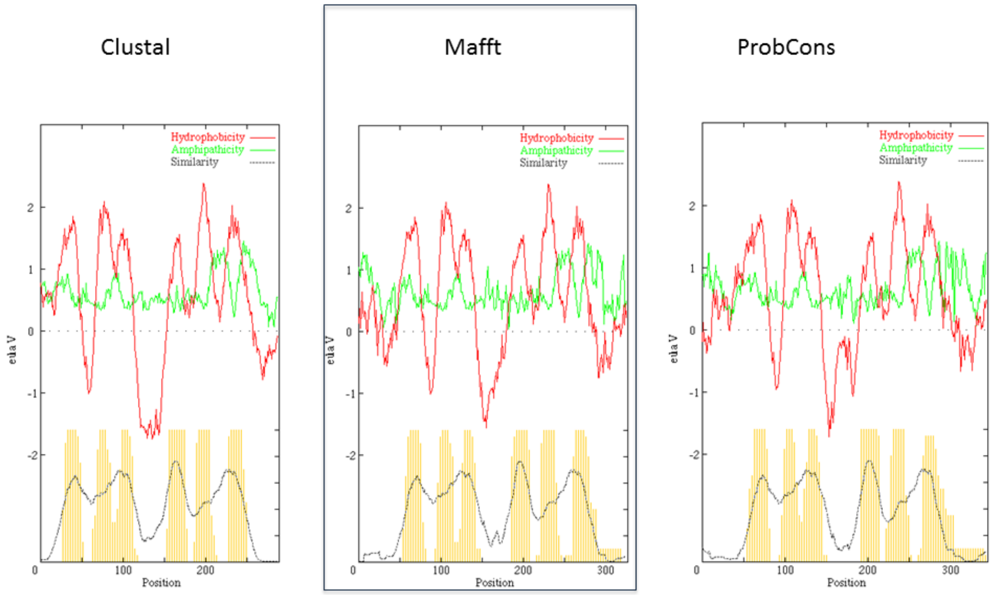


Supplemental Figure 11. AveHAS plots of each family based on multiple alignments generated using three different programs. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) NAAT. (G) NicO. (H) GAP. (I) DsbD. (J) ILT. (K) TerC, cont.

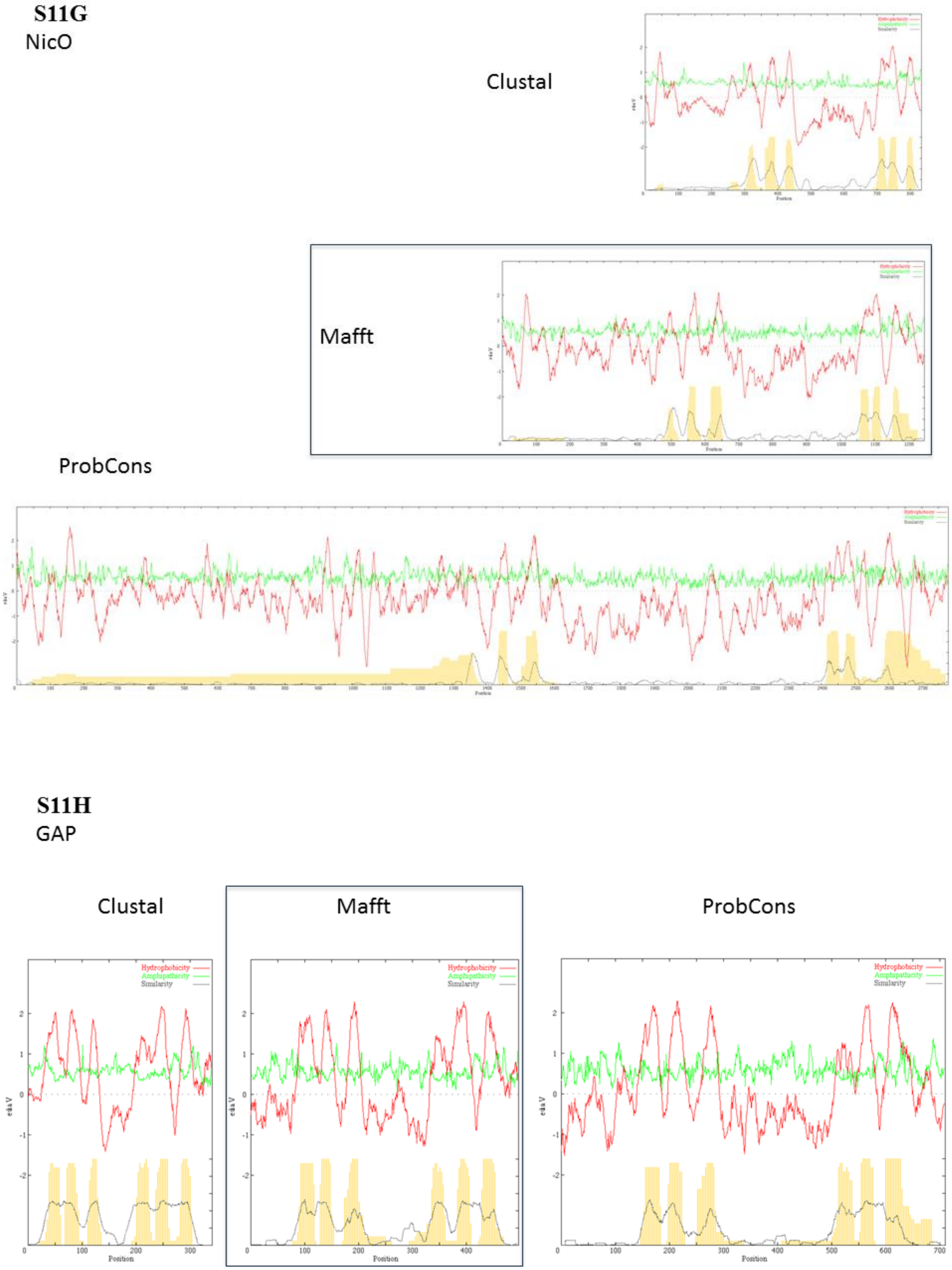
S11E
MntP



S11F
NAAT



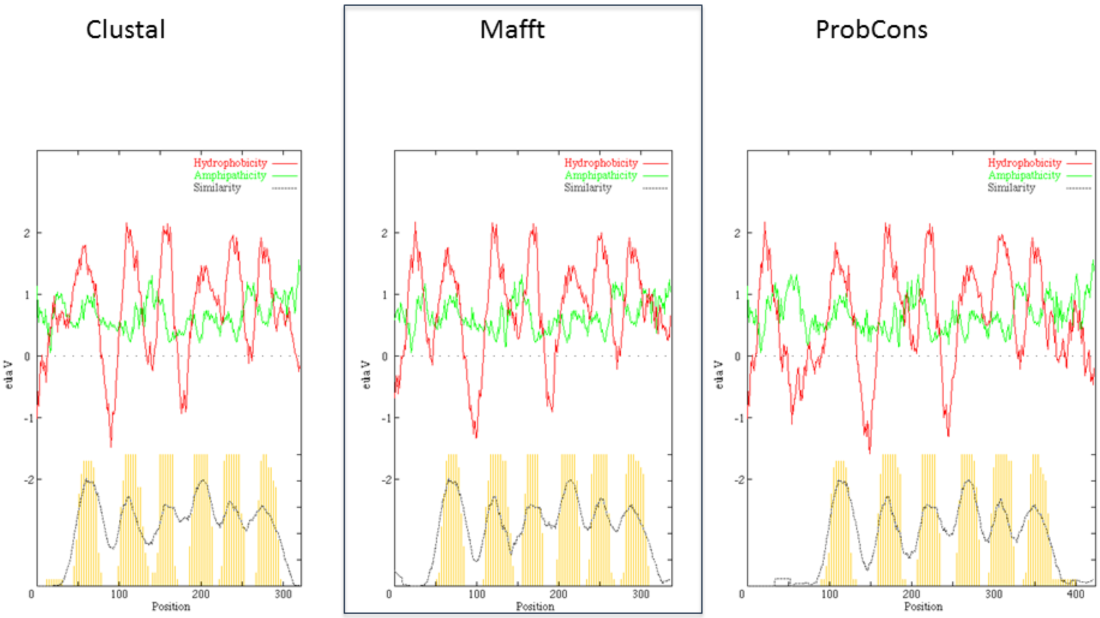
Supplemental Figure 11. AveHAS plots of each family based on multiple alignments generated using three different programs. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) NAAT. (G) NicO. (H) GAP. (I) DsbD. (J) ILT. (K) TerC, cont.


S11G
NicO

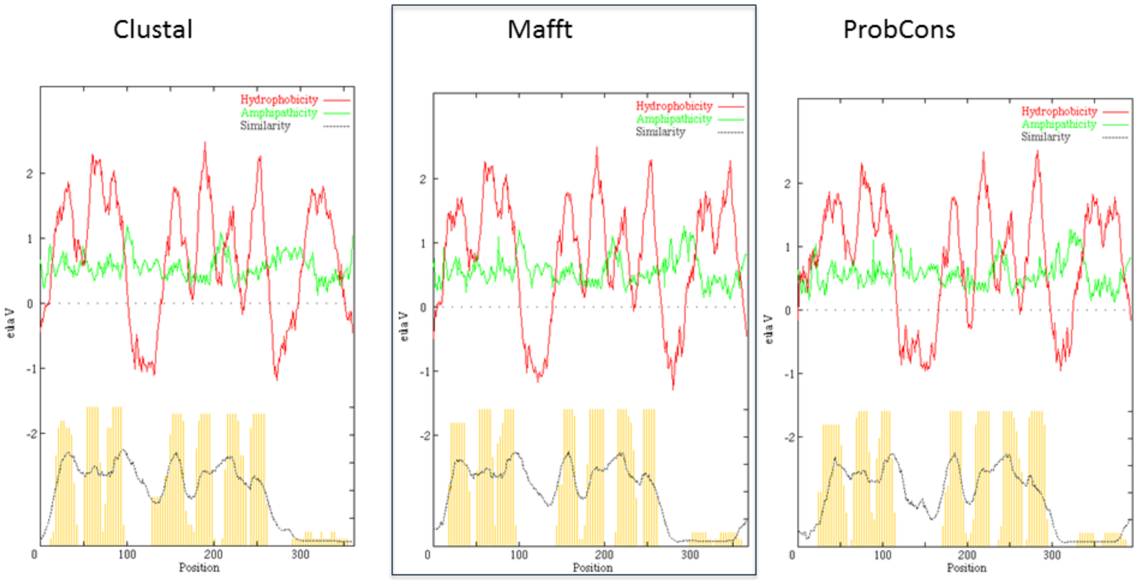
S11H
GAP

Supplemental Figure 11. AveHAS plots of each family based on multiple alignments generated using three different programs. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) NAAT. (G) NicO. (H) GAP. (I) DsbD. (J) ILT. (K) TerC, cont.

S11I
DsbD

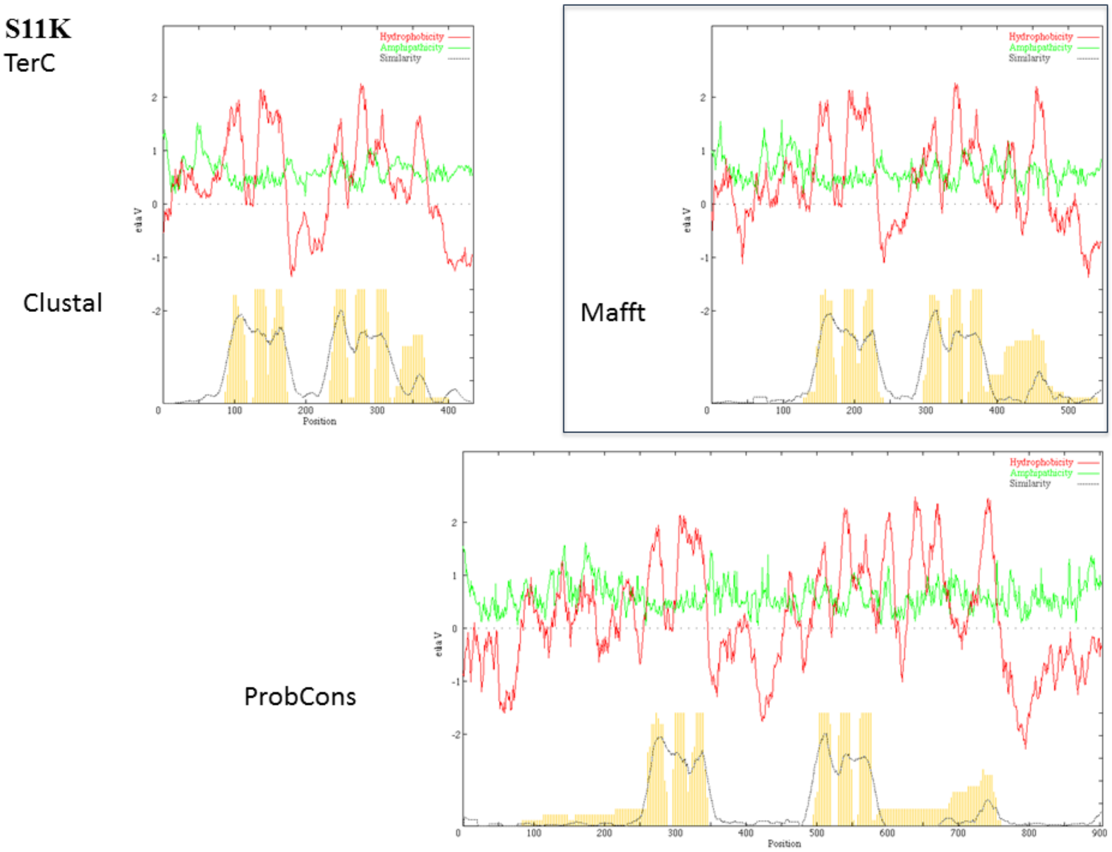


S11J
ILT



Supplemental Figure 11. AveHAS plots of each family based on multiple alignments generated using three different programs. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) NAAT. (G) NicO. (H) GAP. (I) DsbD. (J) ILT. (K) TerC, cont.

S11K
TerC



Supplemental Figure 11. AveHAS plots of each family based on multiple alignments generated using three different programs. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) NAAT. (G) NicO. (H) GAP. (I) DsbD. (J) ILT. (K) TerC, cont.

S12A

```
# 1: A_Sequence: Ssp2 TMS #1-3 (Q2JWH3 ; 2.A.106 homologue)
# 2: B_Sequence: Ssp2 TMS #4-6
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 118
# Identity:      30/118 (25.4%)
# Similarity:    39/118 (33.1%)
# Gaps:          29/118 (24.6%)
# Score: 111.0
#
#
#========================================
                                                     1  ExGD(K/R)(T/S)
A_Sequence         1 ----------------------MWAGFASSLLLVTVAEFGDKTFFTPLIL     28
                                            ||    :  |  :|||||||    : |
B_Sequence         1 EEEEEALRLVEQAEAKGAGRGGAWAVVWEAFSLTALAEFGDKTQIATVSL     50
                                                     4  ExGD(K/R)(T/S)
                                  2                                  3
A_Sequence        29 AMRHPRRWVFLGTWLALAAMTLLAVVAGKVLFELLPPLGVRVLSAGVFAA     78
                     |  ||   |: |  |    |  |||| |: |   :    |  :  |:|
B_Sequence        51 AATHPGLSVWAGATLGHGLMVGLAVVGGRFLAAHISERAVHWVGGGLFLL    100
                                  5                                  6

A_Sequence        79 FGLRMLWQAYQMTPQQEK     96
                     | |   |:
B_Sequence       101 FALVTSWELLG-------    111


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      19.10 +/- 6.83
Standard score (Z):       13.0
Precise score (Z):        13.5
```

Supplemental Figure 12. Identification of internal repeats in the CaCA2 family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three CaCA2 homologues with assigned UniProt accession numbers. (A) Q2JWH3. (B) I7M883. (C) K4DX00.

S12B

```
# 1: A_Sequence: Tth1 TMS #1-3 (I7M883 ; 2.A.106 homologue)
# 2: B_Sequence: Tth1 TMS #4-6
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 172
# Identity:      34/172 (19.8%)
# Similarity:    58/172 (33.7%)
# Gaps:          48/172 (27.9%)
# Score: 102.0
#
#
#=======================================

A_Sequence         1 MKVLYILIISFLLLSSINTKEPNNEKGNSSEKSLLNSFNDDQILQSHGSF     50
                              : |   | : |: ||:  || |          : : : |
B_Sequence         1 ---------NDLKEKSTSDKQQNNQ-ANSQENEKKKKKKQIKGIAAPGYV     40

                           1   ExGD(K/R)(T/S)                      2
A_Sequence        51 IG--SFISTSVSEIGDKTFIMTAILSSKYNRFWVFVGSVGSMLIMTLISC     98
                     |   :|:|    | |||: | |  :|: |:  :||:|:|   :   |::
B_Sequence        41 IAMQTFVSNFFGEWGDKSQISTIAISASYDFVFVFLGTVVGQIFCILLAL     90
                           4   ExGD(K/R)(T/S)                      5
                                        3
A_Sequence        99 LLGS-LTEYFIPLVYVKFISSALFLIFGLKMLYEVYTDTVDDEDDEAEEE    147
                     : |  | : |     :  :   ||:||    ||
B_Sequence        91 IGGQVLAKQFSEKT-MALLGGILFIIFSFITLYTTLNK------------    127
                                        6

A_Sequence       148 VEELEKRLSKIVTKPKTETDQN    169

B_Sequence       127 ----------------------    127


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      18.33 +/- 7.38
Standard score (Z): 11.0
Precise score (Z):  11.3
```

Supplemental Figure 12. Identification of internal repeats in the CaCA2 family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three CaCA2 homologues with assigned UniProt accession numbers. (A) Q2JWH3. (B) I7M883. (C) K4DX00, cont.

S12C

```
# 1: A_Sequence: Tcr1 TMS #1-3 (K4DX00 ; 2.A.106 homologue)
# 2: B_Sequence: Tcr1 TMS #4-6
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 137
# Identity:      25/137 (18.2%)
# Similarity:    47/137 (34.3%)
# Gaps:          26/137 (19.0%)
# Score: 52.0
#
#
#=======================================
                                               1  ExGD(K/R)(T/S)
A_Sequence         1 -------MAIHATRRW--TEGLLSS-FSMILVSEIGDKTFFIACLMAMRH     40
                             |  | | |     :::  |::  |:| ||::      :|
B_Sequence         1 TGSISSTGAGCARRHWFAFHPVMAEVFALTFVAEWGDRSQLATIALAAAK     50
                                               4  ExGD(K/R)(T/S)
                                 2                          3
A_Sequence        41 SKVLVFLGAIGALAGMTVLSALMGLVVPSVLSVRVTKMLAVVLFFGFGGK     90
                     :   | :| :   |  |  :: | | :    :|:|   ::   ||  |
B_Sequence        51 NPFAVTIGGVLGHAVCTGVAVLCGNMTARYVSMRSVNIVGGGLFIVFALA    100
                                 5                          6
A_Sequence        91 ILYDEFAKRGQGDAESDDEMTEAAAIIRKKDPNDAVE    127
                      ||:        |  |   :  :
B_Sequence       101 TLYELITNTHHID-EMQQQKEK---------------    121


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       14.58 +/- 6.61
Standard score (Z): 6.0
Precise score (Z):  5.7
```

Supplemental Figure 12. Identification of internal repeats in the CaCA2 family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three CaCA2 homologues with assigned UniProt accession numbers. (A) Q2JWH3. (B) I7M883. (C) K4DX00, cont.

S13A

```
# 1: A_Sequence: Nps1 TMS #1-3 (Q8YX33 ; 2.A.108 homologue)
# 2: B_Sequence: Nps1 TMS #4-6,7
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 150
# Identity:      30/150 (20.0%)
# Similarity:    50/150 (33.3%)
# Gaps:          54/150 (36.0%)
# Score: 99.0
#
#
#=======================================
                                             1 (D/E)xxE
A_Sequence         1 -----------MNWEIVLASFASSLIELVEILGIVIVVGRL-AGWRNALV     38
                                 ||  |: :|  :|:: ||:   |: :|   : |  |:
B_Sequence         1 EQLARDENESGWNWFAVVTTFKGALLDSVEVAIAVVTLGAAQSQWLEAIS     50
                                             4 (D/E)xxE
                           2                            3
A_Sequence        39 GSGAGIALTLLLSLVLGKSLTLIPVNILRIVAGVLLLLFGQKWTRSIVRY     88
                     |:|      |:|: :    |  :||  ::  | :||: ||  |
B_Sequence        51 GAGFATFSLLVLAFLFRTPLQQVPVKPMKFTAAMLLMGFGLYWLG-----     95
                           5                            6

A_Sequence        89 YAGLPKKRKGGGEDSLE---------------------------------    105
                      |||  :  |   | |
B_Sequence        96 -AGLNVEWPG---DELAIIWLPLAWGVGMAIASTIWRWRVSLDKPEEAIG    141
                                               7

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       20.11 +/- 7.38
Standard score (Z): 11.0
Precise score (Z):  10.7
```

Supplemental Figure 13. Identification of internal repeats in the ILT family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three ILT homologues with assigned UniProt accession numbers. (A) Q8YX33. (B) K9Q6B8. (C) J2KV33.

S13B

```
# 1: A_Sequence: Nps3 TMS #1-3 (K9Q6B8 ; 2.A.108 homologue)
# 2: B_Sequence: Nps3 TMS #4-6,7
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 151
# Identity:      30/151 (19.9%)
# Similarity:    47/151 (31.1%)
# Gaps:          58/151 (38.4%)
# Score: 96.0
#
#
#=======================================
                                               1(D/E)xxE
A_Sequence         1 -------------MNWEIFLASFVGSLIELVEILGLVLIVGKLAG-WRNA      36
                                   ||     :| |:|:: ||:   |: :|   | |  |
B_Sequence         1 LETELANTGNQLGWNWFAIATTFKGALLDSVEVAIAVVTLGATGGKWLEA      50
                                               4 (D/E)xxE
                           2                                3
A_Sequence        37 FVGA-GSGIGLTLLASLILGTSLTIIPVDILRIVAGVFLLAFGQKWTRSI      85
                       ||  :  || ::|  :  | |  :|:  ::  | : |: ||  |
B_Sequence        51 AGGASAAAFGLVVVA-FLFRTPLNQVPIKPMKFTAAMLLMGFGIYWLSE-      98
                           5                                6

A_Sequence        86 VKYYAGIPKKRKDEEDD---------------------------------     102
                          |   | |   ||
B_Sequence        99 -----GF--KIKLPGDDWAIVWLPIVWGCLMAVSALLLRWQVGLQPKEIV     141
                                               7

A_Sequence       102 -    102

B_Sequence       142 S    142


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       21.97 +/- 7.85
Standard score (Z): 9.0
Precise score (Z):  9.4
```

Supplemental Figure 13. Identification of internal repeats in the ILT family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three ILT homologues with assigned UniProt accession numbers. (A) Q8YX33. (B) K9Q6B8. (C) J2KV33, cont.

S13C

```
# 1: A_Sequence: Rsp3 TMS #1-3 (J2KV33 ; 2.A.108 homologue)
# 2: B_Sequence: Rsp3 TMS #4-6,7
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 140
# Identity:      37/140 (26.4%)
# Similarity:    56/140 (40.0%)
# Gaps:          30/140 (21.4%)
# Score: 84.0
#
#
#=======================================
                                        1 (D/E)xxE
A_Sequence         1 MTTITSITSTMA--ASFLGSFVEVVEAFTIILAVGVTQSWRPAFIGTGLA     48
                       :       :|  |:|    :| ||   |::| |      | : | |
B_Sequence         1 --SADRRADFLAGTAAFKAVLLEGVEVVFIVIATGARPGMLP-YAGLGAL     47
                                        4 (D/E)xxE
                           2                         3
A_Sequence        49 LSVLAVLV---LIFGPLLGLIPIDILQFTIGTLLILFGMRWLRKAI----     91
                     :: :||||   |:  | |  :| : |:| :| ||  ||: |: : |
B_Sequence        48 IACIAVLVIGLLVHKP-LSSVPENTLKFIVGLLLTAFGIFWIGEGIGTPW     96
                           5                         6

A_Sequence        92 ----LRASGFIALHDEEKAFASETDALARQ----------    117
                         |   |  ||     ||:       ||
B_Sequence        97 PGEDLSLIGIFAL---LAAFSFIAVRWLRQYHHAQTEPAR    133
                                 7

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       22.15 +/- 7.74
Standard score (Z): 8.0
Precise score (Z):  8.0
```

Supplemental Figure 13. Identification of internal repeats in the ILT family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three ILT homologues with assigned UniProt accession numbers. (A) Q8YX33. (B) K9Q6B8. (C) J2KV33, cont.

S14A

```
# 1: A_Sequence: Ceu1 TMS #1-3 (A8SU47 ; 2.A.107 homologue)
# 2: B_Sequence: Ceu1 TMS #4-6
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 98
# Identity:      22/98 (22.4%)
# Similarity:    42/98 (42.9%)
# Gaps:           8/98 ( 8.2%)
# Score: 82.0
#
#
#=======================================
                                 1 Fully conserved D
A_Sequence         1 ---MSIVELFMLAVGLSMDAFAVSICKGLSLRDIKVKHMVIAGVWFGGFQ     47
                        ||   :|:|||  |:|| || :    :   :   ::|:| :: |
B_Sequence         1 NADMSAKVMFLLAVATSIDALAVGV--SFAFLKLSTLYIVLAVIFIGCIT     48
                                 4 Fully conserved D
                           2                          3
A_Sequence        48 ALMPTLGYVLGSFFADLVSKWSHWIAFVLLLFIGGSMIKESFGGEEEV     95
                      :    |  :|| |       :     ::|: ||  :: :  |
B_Sequence        49 FIFSAAGVKIGSIFGTKYKSKAELAGGIILILIGIKVVLDGLGIL---     93
                           5                          6

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       24.98 +/- 7.01
Standard score (Z): 8.0
Precise score (Z):  8.1
```

Supplemental Figure 14. Identification of internal repeats in the MntP family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three MntP homologues with assigned UniProt accession numbers. (A) A8SU47. (B) R9SLI6. (C) C6JCY1.

S14B

```
# 1: A_Sequence: Rsp2 TMS #1-3 (R9SLI6 ; 2.A.107 homologue)
# 2: B_Sequence: Rsp2 TMS #4-6
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 102
# Identity:      22/102 (21.6%)
# Similarity:    40/102 (39.2%)
# Gaps:          20/102 (19.6%)
# Score: 64.0
#
#
#=======================================
                                        1 Fully conserved D
A_Sequence         1 ----------MNIFELFILAIGLSMDAFAVSVCKGLSLGRINAKHMCIAG     40
                               |:|  :||||:  |:|| ||    |::   :  : :  |
B_Sequence         1 SKEEEHVNADMDIKSMFILAVATSIDALAV----GVTFAFLKVE-IVSAV     45
                                        4 Fully conserved D
                           2                                3
A_Sequence        41 AWFGGFQALMPLVGYFGGRFFADKVTRYSHWVAFVLLVFIGAGMIKE---     87
                     :: |    :    |   |  |  |    :     ::|: ||  :: |
B_Sequence        46 SFIGVITFVCSAAGVKIGSLFGMKYKSKAELCGGIILILIGTKILLEGLG     95
                           5                                6

A_Sequence        87 --     87

B_Sequence        96 MI     97


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       19.03 +/- 6.10
Standard score (Z): 7.0
Precise score (Z):  7.4
```

Supplemental Figure 14. Identification of internal repeats in the MntP family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three MntP homologues with assigned UniProt accession numbers. (A) A8SU47. (B) R9SLI6. (C) C6JCY1, cont.

S14C

```
# 1: A_Sequence: Msp1 TMS #1-3 (C6JCY1 ; 2.A.107 homologue)
# 2: B_Sequence: Msp1 TMS #4-6
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 103
# Identity:      22/103 (21.4%)
# Similarity:    43/103 (41.7%)
# Gaps:          20/103 (19.4%)
# Score: 67.0
#
#
#=======================================
                                        1 Fully conserved D
A_Sequence         1 -----------MDIVSTLLIAVALAMDAFSVSLTKGFTLKNITLKQILWF     39
                                      :|:||| ::|||:| :|    ||   |  ::
B_Sequence         1 FSDDLDDDEDTFSFAELILLAVATSIDAFAVGVTYA-VLKIDILIPVIII     49
                                        4 Fully conserved D
                           2                                 3
A_Sequence        40 GVFFGGFQSLMPILGWTLGVQLQLIVSEVAPWIAFILLVLIGANMIRES-     88
                     |:    |  :  |:|  || ::     :    :  ::|:|:|  :: |
B_Sequence        50 GLV--AF--IFTIIGIYLGKKIGDYFGDKFEILGGVILILLGCRILLEGL     95
                           5                                 6

A_Sequence        88 ---     88

B_Sequence        96 GFL     98


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      21.80 +/- 6.60
Standard score (Z): 7.0
Precise score (Z):  6.9
```

Supplemental Figure 14. Identification of internal repeats in the MntP family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three MntP homologues with assigned UniProt accession numbers. (A) A8SU47. (B) R9SLI6. (C) C6JCY1.

S15A

```
# 1: A_Sequence: Gth1 TMS #1-3 (A4IKQ1 ; 2.A.109 homologue)
# 2: B_Sequence: Gth1 TMS #4-6,7
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 141
# Identity:      29/141 (20.6%)
# Similarity:    52/141 (36.9%)
# Gaps:          50/141 (35.5%)
# Score: 98.0
#
#
#=======================================
                                       1 DxxxxxDN
A_Sequence         1 MSVDLFSPEFWTALLSIVIIDLVLAGDNAIVIGLAARNLPKHQQKKAVIW     50
                               | |: :|:| | ::  || : :  ||     |     ||
B_Sequence         1 -------GSLWEAVRTIIIADALMGLDNVLAVAGAA-----HGHFLLVIL     38
                                       4 DxxxxxDN
                           2                         3
A_Sequence        51 GTVGAVVIRAM-ATIFVVWLLKIPGLLLVGGLLLVWIAYKLLVEE---KG     96
                     | : :| |    :|: : |: : | :: :|  :| | | |::|:|   ||
B_Sequence        39 GLLISVPIMVWGSTLILKWIERFPIIITIGAGILAWTASKMIVDEPFLKG     88
                           5                         6

A_Sequence        97 H---DDIEAG-------------------------------    103
                     :     |: |
B_Sequence        89 YFANPVIKYGFELLLVAAVIAIGTQKKRKAAKKPHLKVANE    129
                                 7

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       23.59 +/- 7.93
Standard score (Z): 9.0
Precise score (Z):  9.4
```

Supplemental Figure 15. Identification of internal repeats in the TerC family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three TerC homologues with assigned UniProt accession numbers. (A) A4IKQ1. (B) G8M4S7. (C) R9LI44.

S15B

```
# 1: A_Sequence: Bsp2 TMS #1-3 (G8M4S7 ; 2.A.109 homologue)
# 2: B_Sequence: Bsp2 TMS #4-6,7
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 143
# Identity:      30/143 (21.0%)
# Similarity:    44/143 (30.8%)
# Gaps:          50/143 (35.0%)
# Score: 92.0
#
#
#=======================================
                                      1 DxxxxxDN
A_Sequence         1 MLEFFSTLHWGAVVQIIVIDILLGGDNAVVIALACRNLPDRQRTRGIVLG     50
                          |   | ||  |:: | ::  || : || |      | |   :: |
B_Sequence         1 -----SDRLWAAVKTIVIADAVMSLDNVIAIAGAAEAADPRHRLALVIFG     45
                                      4 DxxxxxDN
                           2                         3
A_Sequence        51 TLGAILLRVILIAFAVMLLD-VPFLKFVGGVLLLWIGVKLMQPDHDEHHI     99
                      : :| | |      : |||  | :  :|  || ||   |:    |:  |
B_Sequence        46 LIVSIPLIVWGSTLVLKLLDRFPVVVLLGAALLGWIAGGLI---IDDPFI     92
                           5                         6

A_Sequence       100 DA-----------------------------------------    101
                     |
B_Sequence        93 DRWPALNTDIVGYAARVAGALFVVGVGWLLRRRALADGNRATG    135
                                        7

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       21.13 +/- 7.79
Standard score (Z): 9.0
Precise score (Z):  9.1
```

Supplemental Figure 15. Identification of internal repeats in the TerC family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three TerC homologues with assigned UniProt accession numbers. (A) A4IKQ1. (B) G8M4S7. (C) R9LI44, cont.

S15C

```
# 1: A_Sequence: Pba1 TMS #1-3 (R9LI44 ; 2.A.109 homologue)
# 2: B_Sequence: Pba1 TMS #4-6,7
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 143
# Identity:      25/143 (17.5%)
# Similarity:    47/143 (32.9%)
# Gaps:          53/143 (37.1%)
# Score: 82.0
#
#
#=======================================
                                     1 DxxxxxDN
A_Sequence         1 MDLLSPEFWMALLSIVLIDLVLAGDNAIVIGLAARNVPQQDQKKVIVWGT     50
                           : | |: :|:: | ::  || : :  ||      |   ||: |
B_Sequence         1 -----NQMWAAIRTIIIADAMMGLDNVLAVAGAAHG----DTLLVII-GL     40
                                     4 DxxxxxDN
                          2                          3
A_Sequence        51 LGAILIRVVMTLLVVQLL-NIPGLRLAGGLALVWIAYKLLIEEK-SHEIK     98
                       :: | |  : ::::|    | :   |   | | | |:::||   |:
B_Sequence        41 AVSVPIMVWGSTMILKLTERFPIVITIGAAVLAWTASKMIVEEPLIHDWF     90
                          5                          6

A_Sequence        99 AG-----------------------------------------   100
                     |
B_Sequence        91 ASPWIKYGFELLVIAAVVLLGNLMKKRKARLHQAKAMPQTNGS   133
                          7

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       23.22 +/- 7.53
Standard score (Z): 8.0
Precise score (Z):  7.8
```

Supplemental Figure 15. Identification of internal repeats in the TerC family. GSAT comparisons between TMS#1-3 and TMS#4-6 for three TerC homologues with assigned UniProt accession numbers. (A) A4IKQ1. (B) G8M4S7. (C) R9LI44, cont.

S16A

```
# 1: A_Sequence: Ame2 (MC homologue)
# 2: B_Sequence: Spl1 (LysE homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 195
# Identity:      16/195 ( 8.2%)
# Similarity:    35/195 (17.9%)
# Gaps:         131/195 (67.2%)
# Score: 38.0
#
#
#=======================================
                         4
A_Sequence         1 VLGLYRGFNVSVQGIIIYRAAYFGFYDTTKNLLPDPKKTPLHITFLIAQT     50

B_Sequence         0 --------------------------------------------------      0

                         5
A_Sequence        51 VTTLAGIISYPFDTVRRRMMMQSGLKRAE----VMYKNTLDCWIKTAKTE     96
                                        ::  |:::       :  :  | :: :|
B_Sequence         1 -------------------VLTQGIRKQHRFVVALICSLCDAFLISAGVA     31
                                         1                 2

A_Sequence        97 GIAAFFKGSLSNI-LRGTGGALVLTLYDSIKDILEKSLRK----------    135
                     |: :  : | : : | | |||| | :|  :|  |  :|:
B_Sequence        32 GLGSLIEQSPTLLRLAGGGGALFLFIY-GLK-CLFSALQAEQELGETESN     79
                                          3

A_Sequence       135 ---------------------------------------------    135

B_Sequence        80 PTSRRQVILTILAITLCNPNVYLDTVVLLGGISATFVGQGRYLFG    124
                                  4

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      14.44 +/- 5.82
Standard score (Z):       4.0
Precise score (Z):        4.1
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD.

S16B

```
# 1: A_Sequence: Pmo1 (MC homologue)
# 2: B_Sequence: Hgr1 (RhtB homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 89
# Identity:      26/89 (29.2%)
# Similarity:    39/89 (43.8%)
# Gaps:           7/89 ( 7.9%)
# Score: 70.0
#
#
#=======================================
                                  3
A_Sequence         1 TQFWRYFIGNLASGGAAGDTSLCFVYTLDFARTRLAADIGKGAGQREFNG     50
                     |  | | :| :      |  || || ::  ||      :|  |    |
B_Sequence         1 TDLWTYVLGAIGIVLLPGPNSL-FVLSVATAR---GVRVGYHAACGVF--     44
                                  1                                 2
                                          4
A_Sequence        51 LGDCLVKIFKADGIMGLYRGFGVSVQGIIIYRAAFFGFY     89
                     ||| :: :| | |   | ||:      :: |  | : |:
B_Sequence        45 LGDSILLLFTALGAASLLRGYPALFM-VVKYVGAAYLFW     82
                                                   3

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      16.98 +/- 5.99
Standard score (Z):       9.0
Precise score (Z):        8.8
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16C

```
# 1: A_Sequence: Oga1 (MC homologue)
# 2: B_Sequence: Sro1 (CadD homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 87
# Identity:      26/87 (29.9%)
# Similarity:    39/87 (44.8%)
# Gaps:          14/87 (16.1%)
# Score: 59.0
#
#
#=======================================
                           5
A_Sequence         1 ISWMIAQSATAVAGLTSYPFDTVHHCMMMQSVRKGTGIMYTGATDCWRKI     50
                     |: ::| || |  |||  | |      ::  |  | |:        |  :
B_Sequence         1 IAILVAVSAVAALGLTVVP-D--RWAGLLGLVPFGMGV--------WGLV     39
                           2                          3
                                        6
A_Sequence        51 LRDEGGKAFFKGAXSSVLRGVGGAFVLVLYDEIKKYT     87
                      :|:||:|   |  :| :  | |  :    | |  ||
B_Sequence        40 RKDDGGEA---GPVASGVVSVAGVTLANGADNISVYT     73
                                        4

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       13.78 +/- 5.35
Standard score (Z): 8.0
Precise score (Z):  8.5
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16D

```
# 1: A_Sequence: Isc1 (MC homologue)
# 2: B_Sequence: Ghi1 (CaCA2 homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 97
# Identity:      30/97 (30.9%)
# Similarity:    51/97 (52.6%)
# Gaps:          10/97 (10.3%)
# Score: 75.0
#
#
#=======================================
                                         4
A_Sequence         1 LGNCLTKIFKSDGL-MGLYRGFG--VSVQGIIIYRAA-YFGF--FDTAKG      44
                     :|: |  :  :| | : : | ||  :  : : :: || :|||  :   :|
B_Sequence         1 IGSTLGMV-AADALAIAIGRAFGRHLPERTVALFAAALFFGFGIWLLTQG      49
                           5                          6
                                         5
A_Sequence        45 MLPDPKNTPLVISWLIAQTVTTVAGIMSYPFDTVRRRMMMQSGRAKA      91
                     :| |    |::|  | |  |  ||||      | ||| : :: | :|
B_Sequence        50 LL-D-ATVPVLIGTLTAVAVVMVAGI-GVIVSTHRRRQLEKAIRTRA      93
                                         7

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       13.98 +/- 5.83
Standard score (Z): 10.0
Precise score (Z):  10.5
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16E

```
# 1: A_Sequence: Mbr1 (MC homologue)
# 2: B_Sequence: Cst1 (MntP homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 79
# Identity:      22/79 (27.8%)
# Similarity:    39/79 (49.4%)
# Gaps:           7/79 ( 8.9%)
# Score: 62.0
#
#
#=======================================
                                 2
A_Sequence         1 FWRYFA-GNLASGGAAGATSLCFVYPLD----FA-RTRLAADVGKGSAQR     44
                     |: :|  | |: |  :|:|   |: |:|    |   | :   : | | :
B_Sequence         1 FFGFFQWGMLSLGWLSGSTFRTFIEPVDHWIAFVLLTFIGVKMWKESTEE     50
                           2                   3
                                        3
A_Sequence        45 MLP-DPKNVHIFISWMIAQSVTAVAGLVS      72
                       | |  :| : ::  :| |: | |  :|
B_Sequence        51 AEPLDLTSVKLMLTLSVATSIDAFAAGIS      79
                                        4

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      13.34 +/- 5.38
Standard score (Z): 9.0
Precise score (Z):  9.1
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16F

```
# 1: A_Sequence: Cmi2 (MC homologue)
# 2: B_Sequence: Aho1 (ILT homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 142
# Identity:      36/142 (25.4%)
# Similarity:    58/142 (40.8%)
# Gaps:           7/142 ( 4.9%)
# Score: 65.0
#
#
#=======================================
                                  2
A_Sequence         1 QGFLSFWRGNLANVIRYFPTQALNF-AFKDKYKQIFMSGIDKK---TQFG     46
                     | :| : |  : : :       | | ||   |::|| : :  :   ||
B_Sequence         1 QRWLGYIRDKVDSALGRGTVWTLAFVAFISVYREIFETILFYQALWTQVD     50
                                          3
                                  3
A_Sequence        47 KWFLANLASGGAAGATSLCFVYPLDFARTRLAADVGKGNEERQFKGLADC     96
                         | |  |  |   :|  |  | | |  :   :|          |
B_Sequence        51 GQTQAFLFYGIGAAVLALA-VVSLLFFRVGMTLPLGVFFRVTSLVLLVLS     99
                                  4
                                          4
A_Sequence        97 LAKIGKRDGIQGLYQGFAVSVNGIIVYRASYFGCYDTIKGIL    138
                     :  :||  ||  | :   :||  : |    : | | |::|:|
B_Sequence       100 VILLGK--GIAALQEAGLISVMHLAVPTVDWLGVYPTVQGLL    139
                         5

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     13.51 +/- 5.69
Standard score (Z):      9.0
Precise score (Z):       9.1
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16G

```
# 1: A_Sequence: Rsy1 (MC homologue)
# 2: B_Sequence: Sya2 (TerC homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 144
# Identity:      28/144 (19.4%)
# Similarity:    48/144 (33.3%)
# Gaps:          50/144 (34.7%)
# Score: 38.0
#
#
#=======================================
                                                                   1
A_Sequence         1 ANVIRYFPTQALNFGFKDKYKKIFLDNVDKRTQFWRYFAGNLASGGAAGA     50
                                   |:|  ::::      :  :   : ||  |  | | |
B_Sequence         1 ------------WVGWK-MWRELRAHGEPEDAE---HMAGKAAPKGFAQA     34
                                  3

A_Sequence        51 TSLCFVYPLDFARTRLAAD--VGKAGAGREFNGLGDCLAKIFKSDGLKGL     98
                           : :  |   :: |  :  ||| ||  |:   :  :  |  | |:
B_Sequence        35 -----AWAVAIADVSMSLDNVLAVAGAAREHPGI--LVIGLVLSVALMGV     77
                                           4                         5
                                  2
A_Sequence        99 YQGFNVSVQGIIIYRA-AYFGI----------------------    119
                         |:  : |  ||| ||||:
B_Sequence        78 --AANLLARVIERYRAVAYFGLIVILYVAGKMIYEGAIDPATGL    119
                                           6

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      12.70 +/- 5.66
Standard score (Z):       4.0
Precise score (Z):        4.4
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16H

```
# 1: A_Sequence: Rsy1 (MC homologue)
# 2: B_Sequence: Orf9 (NAAT homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 75
# Identity:      19/75 (25.3%)
# Similarity:    35/75 (46.7%)
# Gaps:           3/75 ( 4.0%)
# Score: 63.0
#
#
#=======================================
                           2                                3
A_Sequence         1 SVQGIIIYRAAYFGIYDTAKGMLPDPKNTHIFVSWMIAQSVTAVAGF-GS     49
                     ::   |:| | |  ::||  |::        | ||:: :|   :  | |
B_Sequence         1 AISSTIVYGARYPSMFDT-MGIIL-TIIAFCFCSWLLFRSAPLLVRFLGQ     48
                           4                    5

A_Sequence        50 YPFDTVRRRMMMQSGRKGAEIMYSG     74
                        : : | | :  |  | | : :|
B_Sequence        49 TGINVITRIMGLILGALGIEFIANG     73
                                 6

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)       12.39 +/- 5.04
Standard score (Z): 10.0
Precise score (Z):  10.0
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16I

```
# 1: A_Sequence: Cfe1 (MC homologue)
# 2: B_Sequence: Bsm1 (NicO homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 114
# Identity:      26/114 (22.8%)
# Similarity:    46/114 (40.4%)
# Gaps:           8/114 ( 7.0%)
# Score: 67.0
#
#
#=======================================
                                          2
A_Sequence         1 SYRGIFHAFSTIYQQEGFLAFYRGVSLTVLVYMN-LEKIWNGPRDRFSLF     49
                     :|:|| :  |      | :    | :  ||: |: :|| | |    : ||
B_Sequence         1 TYKGIPYVKSLFI---GIIHGLAGSAAMVLLTMSTVEKAWEGL--LYILF     45
                                          3

A_Sequence        50 QNFANVCLAAAVTQTLSFPFDTVKRKMQAQSPYLPHCGGVDVHFSGAVDC     99
                          |      |  :  ||    ||::  : ::   | :   |   :
B_Sequence        46 FGAGTVLGMLCFTTLIGIPFTLSARKIRIHNAFIQITGFISTVF--GIHY     93
                         4                                 5
                                  3
A_Sequence       100 FRQVVKAQGVLGLW    113
                        :   :|:  ||
B_Sequence        94 MYNLGVTEGLFKLW    107


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     14.43 +/- 5.72
Standard score (Z):      9.0
Precise score (Z):       9.3
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16J

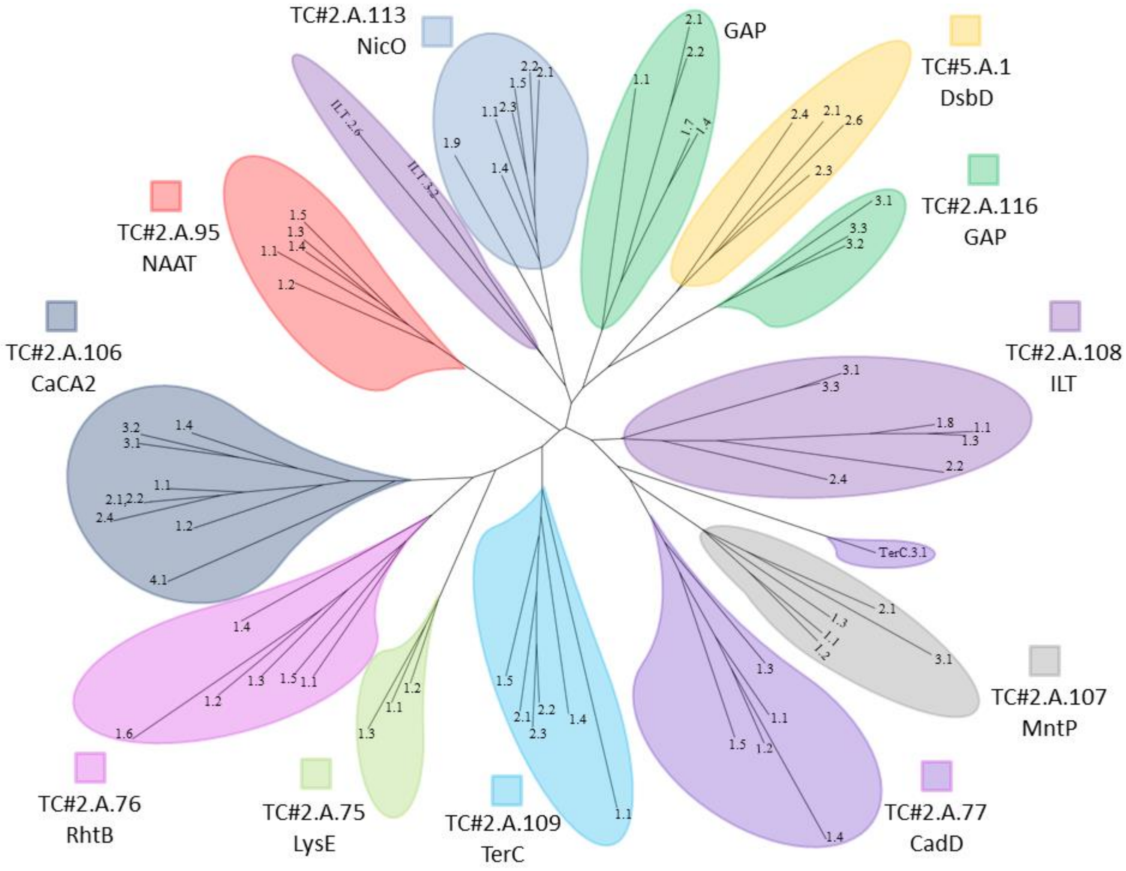
```
# 1: A_Sequence: Cmi2 (NicO homologue)
# 2: B_Sequence: Msp16 (GAP homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 153
# Identity:      19/153 (12.4%)
# Similarity:    30/153 (19.6%)
# Gaps:          78/153 (51.0%)
# Score: 40.0
#
#
#=======================================

A_Sequence         1 -----------------------------------------QGFLSFWRG      9
                                                              :|  | |
B_Sequence         1 PRVQMAAGVIVLLVAAAVAVGLGGTKAGRRGQLATRTSRLMEGH-SLWIA     49
                                  3
                          2                                         3
A_Sequence        10 NLANVIRYFPTQALNFAFKDKYKQIFMSGIDKKTQFGKWFLANLASGGAA     59
                      :| :    |    :  :      |  ||    || |   | |: : |
B_Sequence        50 GVAGLGIALP----SVDYLAALTIIIASGAAAATQVGALLLFNVVAFGLV     95
                          4                                         5
                                                                      4
A_Sequence        60 GATSLCFVYPLDFARTRLAADVGKGNEERQFKGLADCLAKIGKRDGIQGL    109
                         :|::   |  |  |:|
B_Sequence        96 EIPLICYLVAPDRTRAMLSAL-----------------------------    116


A_Sequence       110 YQG    112

B_Sequence       116 ---    116


#---------------------------------------

============ FINISHED =============
Average Quality (AQ)     10.38 +/- 5.16
Standard score (Z):      6.0
Precise score (Z):       5.8
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

S16K

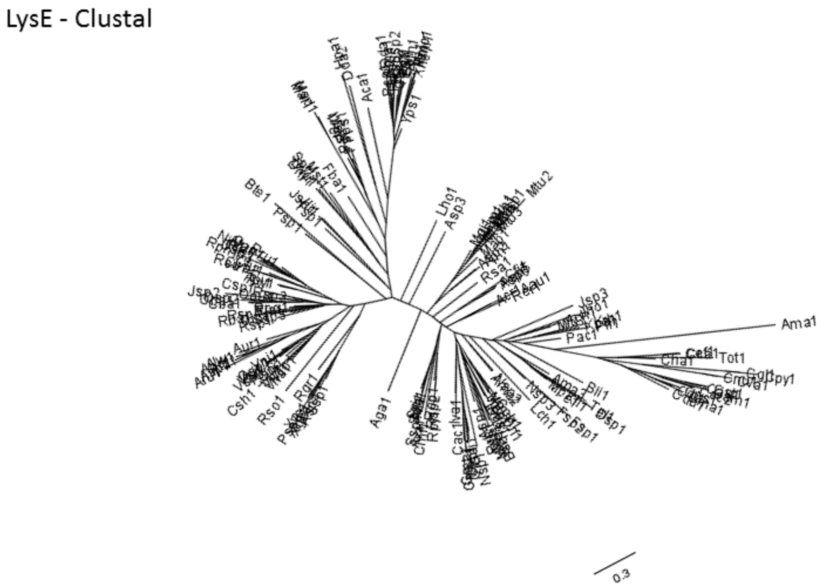
```
# 1: A_Sequence: Cmy1 (MC homologue)
# 2: B_Sequence: Glo1 (DsbD homologue)
# Matrix: EBLOSUM62
# Gap_penalty: 8.0
# Extend_penalty: 2.0
#
# Length: 116
# Identity:      29/116 (25.0%)
# Similarity:    51/116 (44.0%)
# Gaps:          16/116 (13.8%)
# Score: 69.0
#
#
#=======================================
                                                   5
A_Sequence         1 SYFGCYDTIKGLLP--NPKQTPFVLSFLIAQAVTTFSGI-LSYPFDTVRR      47
                     :| | :  : |||   :|   | : |::      :|| :   :|   |||
B_Sequence         1 TYIGAF--VAGLLSFLSPCVLPLIPSYITYITGLSFSDLDAEHPTHVVRR      48
                                  1

A_Sequence        48 RMMMQSGEAERQYKGTIDCFFKIYKQEGLKAFFRGAF----SNILRGTGG      93
                     : |: |          :  |  ::   |  |  : |:|      ::|  ||
B_Sequence        49 KTMLHS-------LAFVSGFTVVFVLLGASATYIGSFLQQHMELVRKLGG      91
                                  2
                         6
A_Sequence        94 ALVLVLYDKIKELVNL    109
                      |::|    :  || |
B_Sequence        92 ILIIVFGIHVTGLVPL    107
                         3

#---------------------------------------

============ FINISHED =============
Average Quality (AQ)      13.12 +/- 5.64
Standard score (Z):       10.0
Precise score (Z):        9.9
```

Supplemental Figure 16. GSAT comparisons with MC, the negative control. (A) LysE. (B) RhtB. (C) CadD. (D) CaCA2. (E) MntP. (F) ILT. (G) TerC. (H) NAAT. (I) NicO. (J) GAP. (K) DsbD, cont.

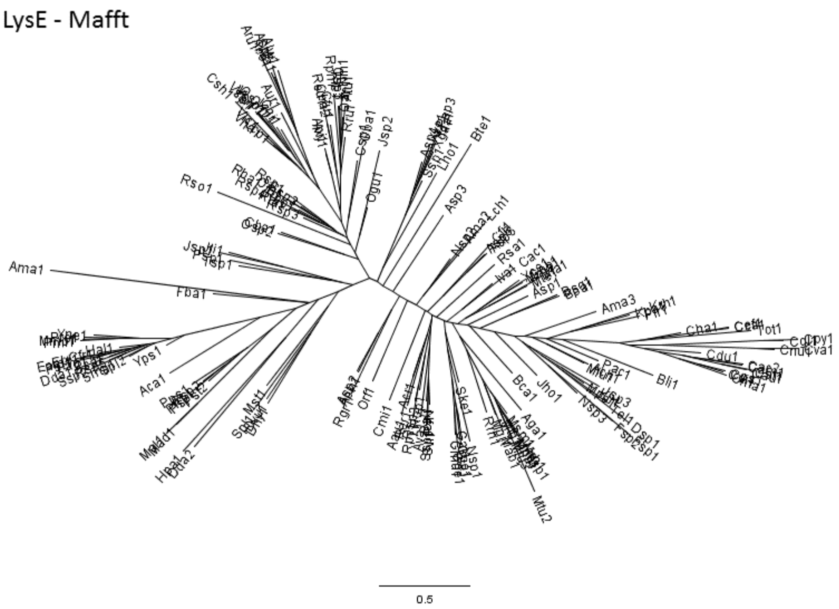


Supplemental Figure 17. RAxML Phylogenetic Tree of the LysE Superfamily based on a multiple alignment generated with Mafft. The Mafft-homologs function was set to retrieve 200 homologs at a threshold E-value of 1e^{-20} by BLAST (Using UniProt) for each query sequence to improve the accuracy of aligning a small number of distantly related sequences.

S18A



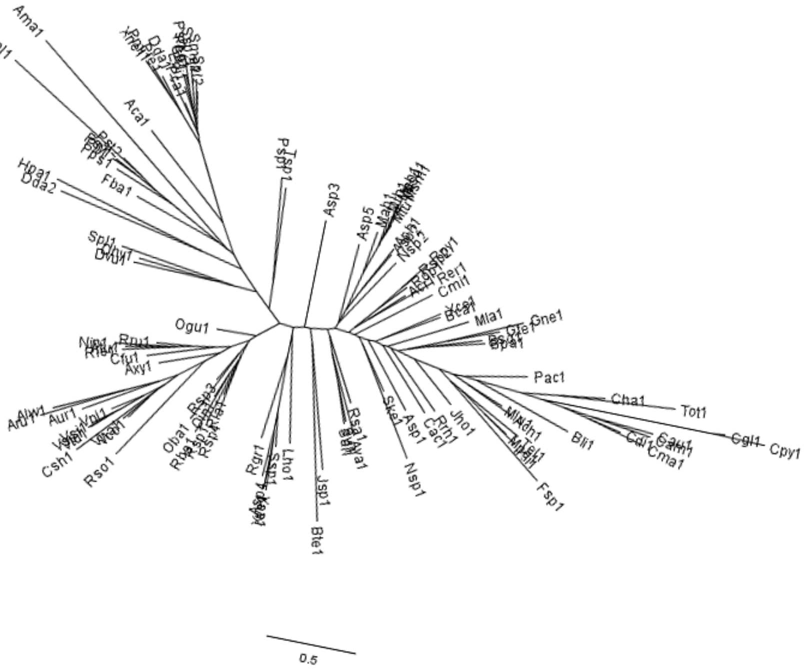
S18B



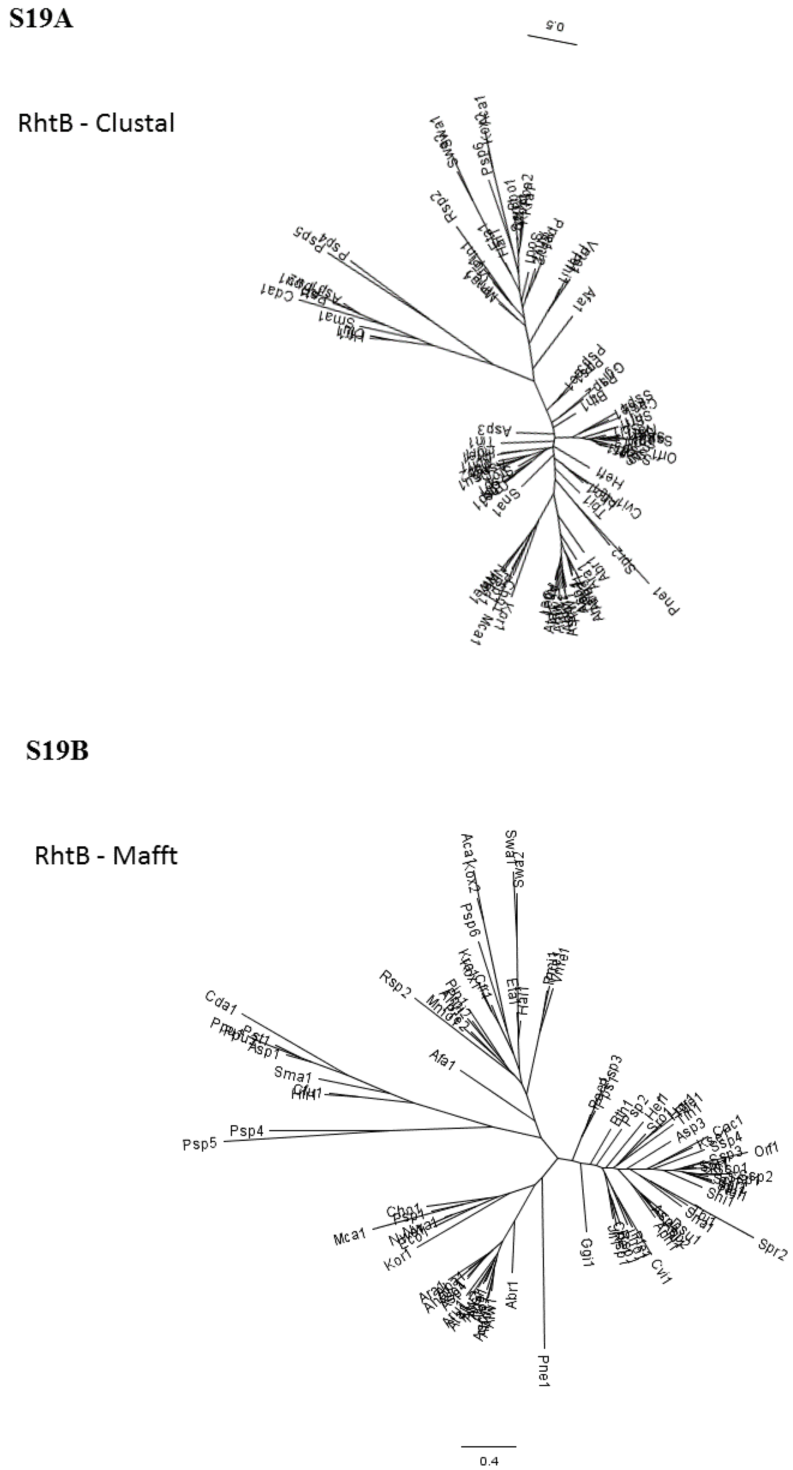
Supplemental Figure 18. Phylogenetic Trees of the LysE Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

S18C

LysE - ProbCons



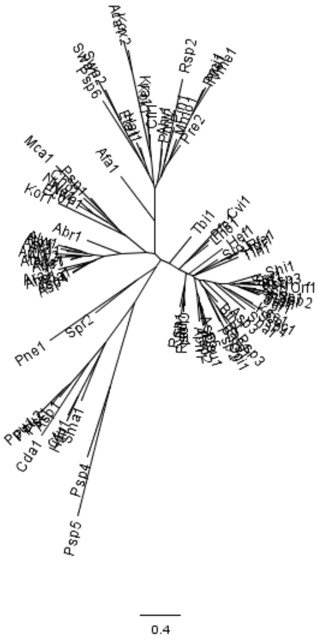
Supplemental Figure 18. Phylogenetic Trees of the LysE Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.



Supplemental Figure 19. Phylogenetic Trees of the RhtB Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

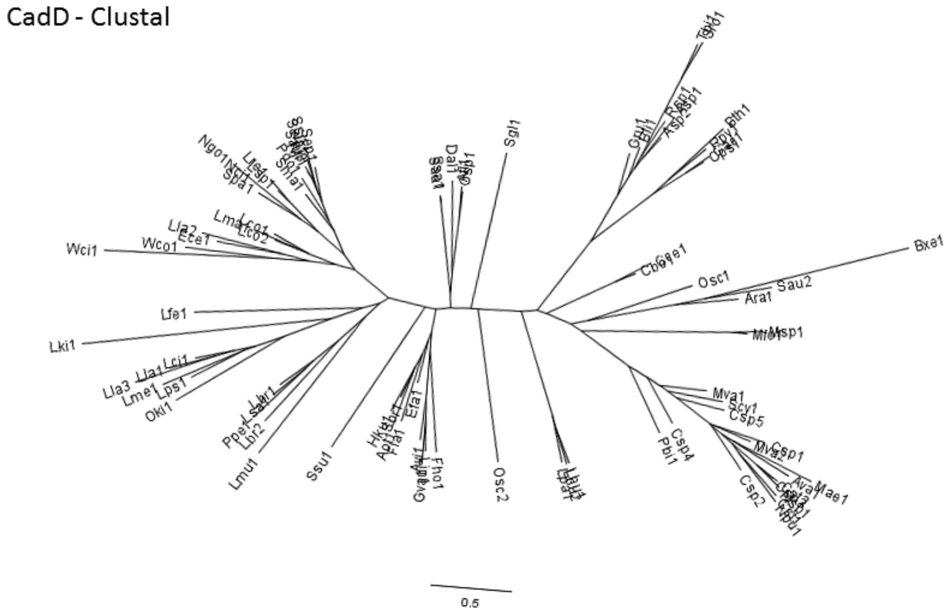
S19C

RhtB - ProbCons

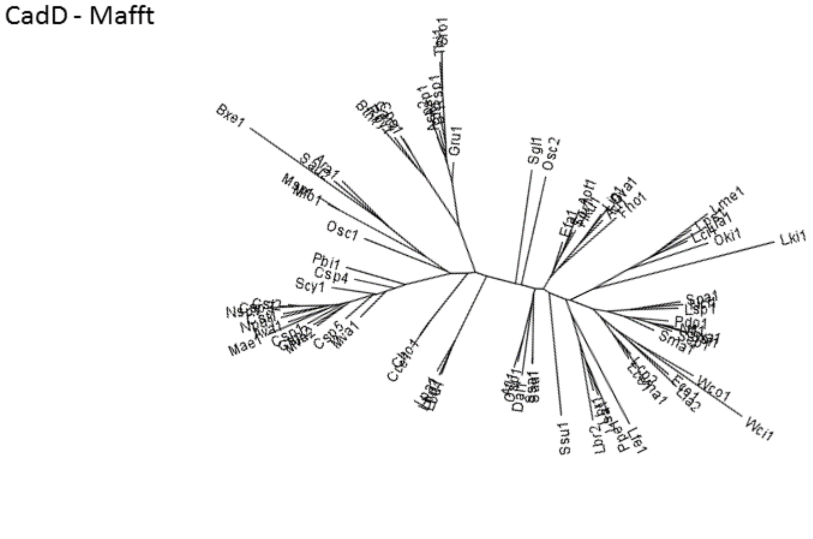


Supplemental Figure 19. Phylogenetic Trees of the RhtB Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.

S20A



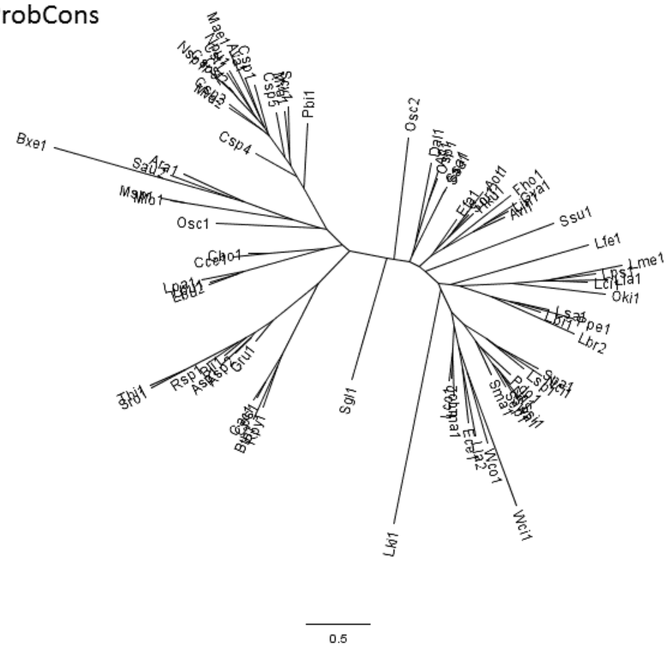
S20B



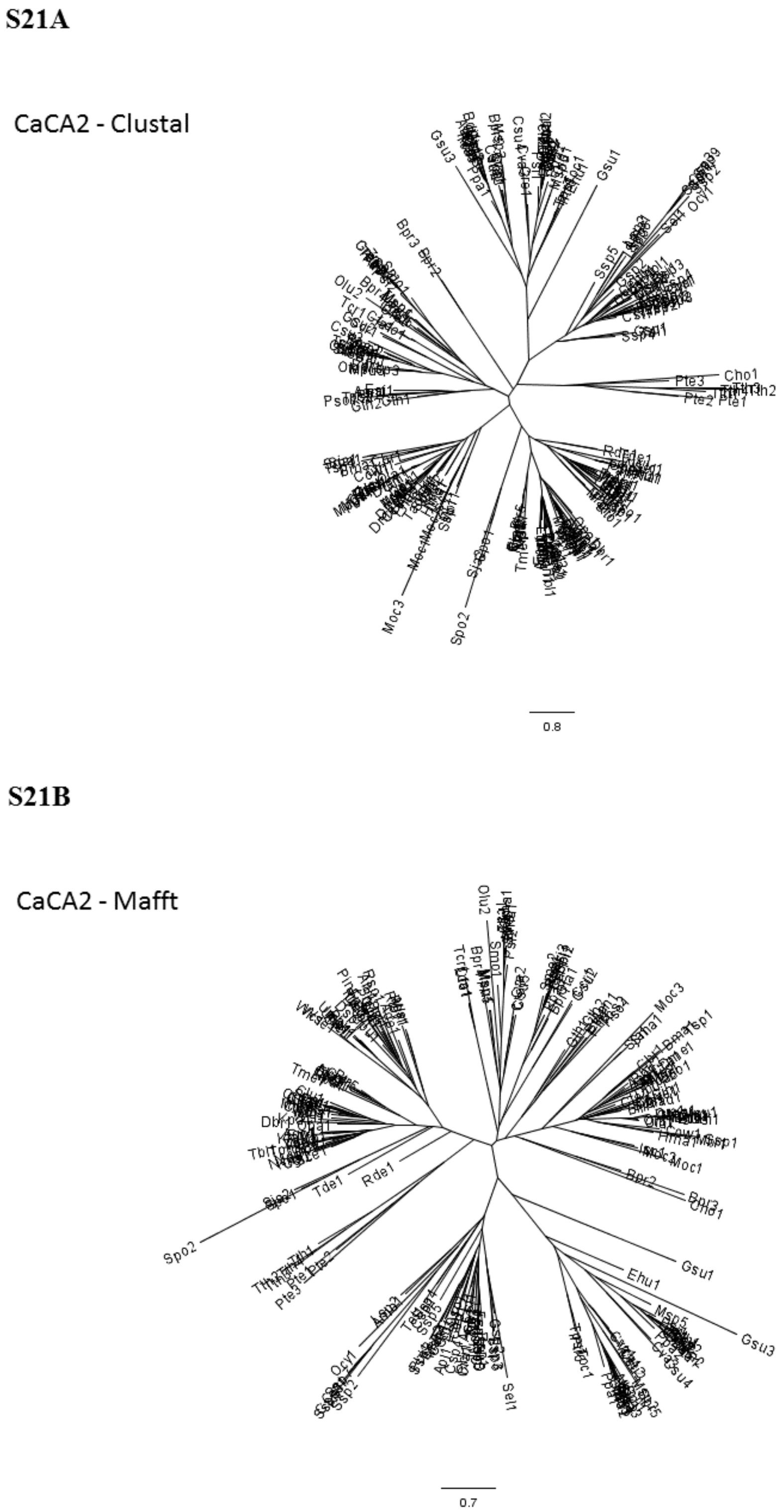
Supplemental Figure 20. Phylogenetic Trees of the CadD Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

S20C

CadD - ProbCons



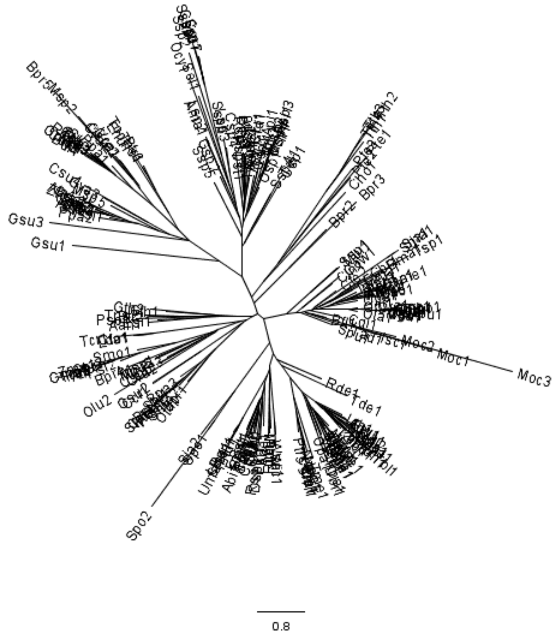
Supplemental Figure 20. Phylogenetic Trees of the CadD Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.



Supplemental Figure 21. Phylogenetic Trees of the CaCA2 Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

S21C

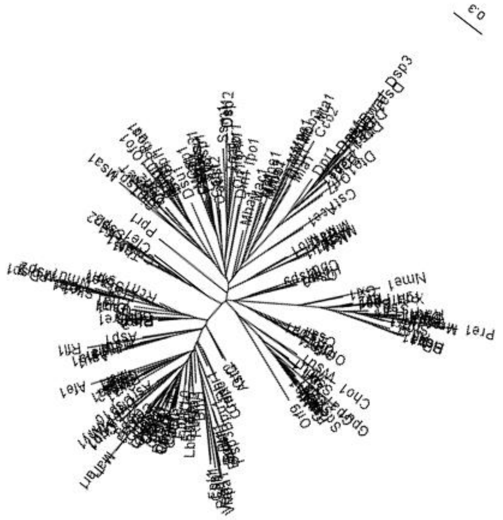
CaCA2 - ProbCons



Supplemental Figure 21. Phylogenetic Trees of the CaCA2 Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.

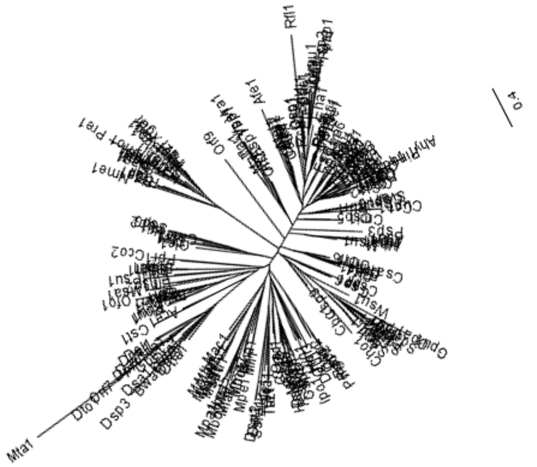
S22A

MntP - Clustal



S22B

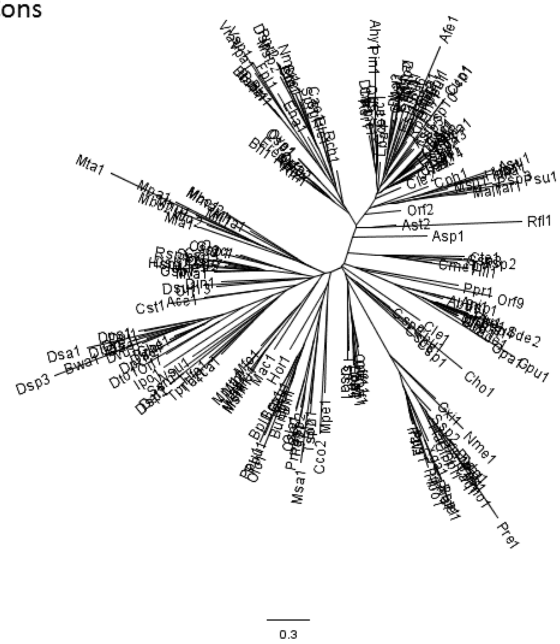
MntP - Mafft



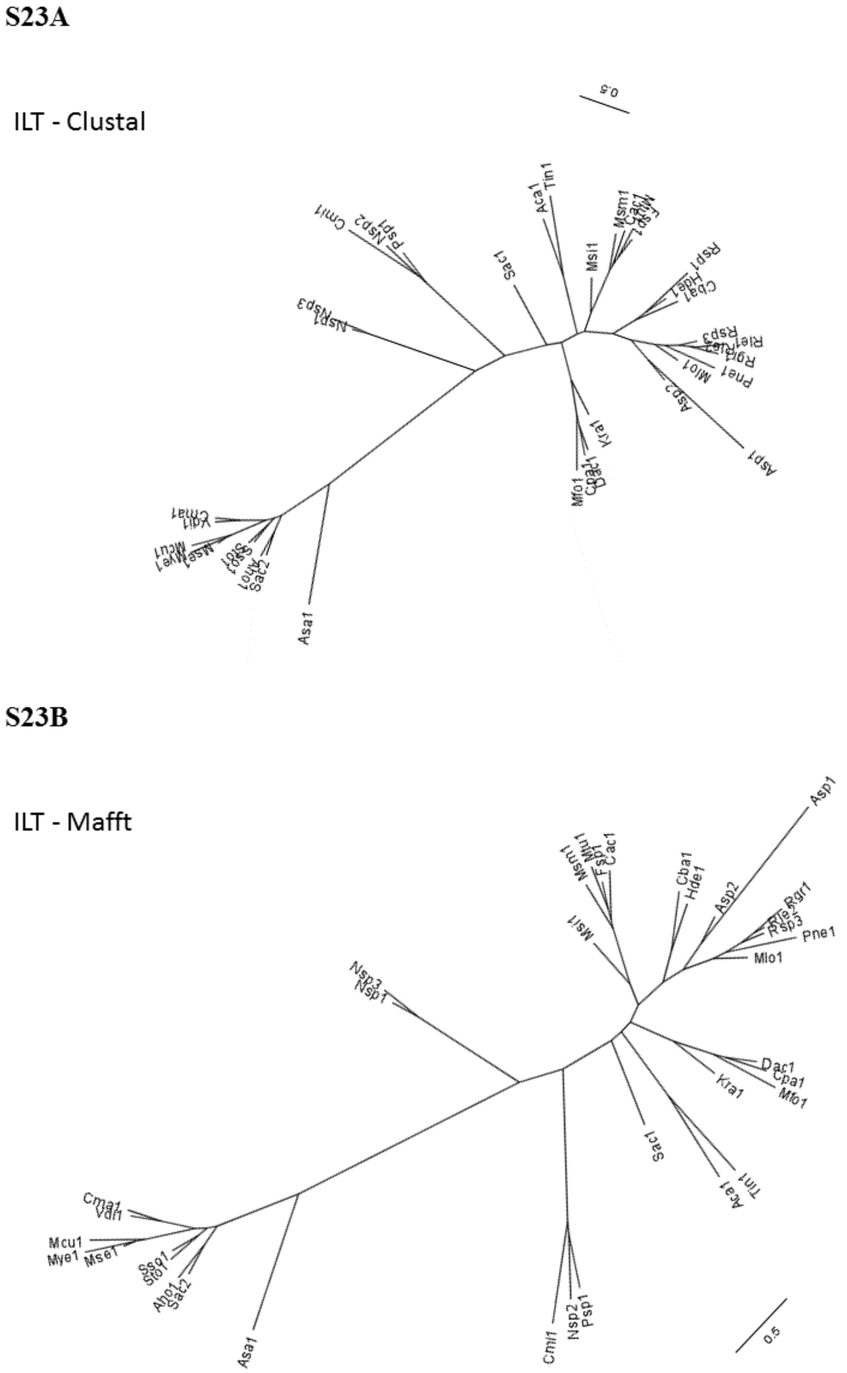
Supplemental Figure 22. Phylogenetic Trees of the MntP Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

S22C

MntP - ProbCons

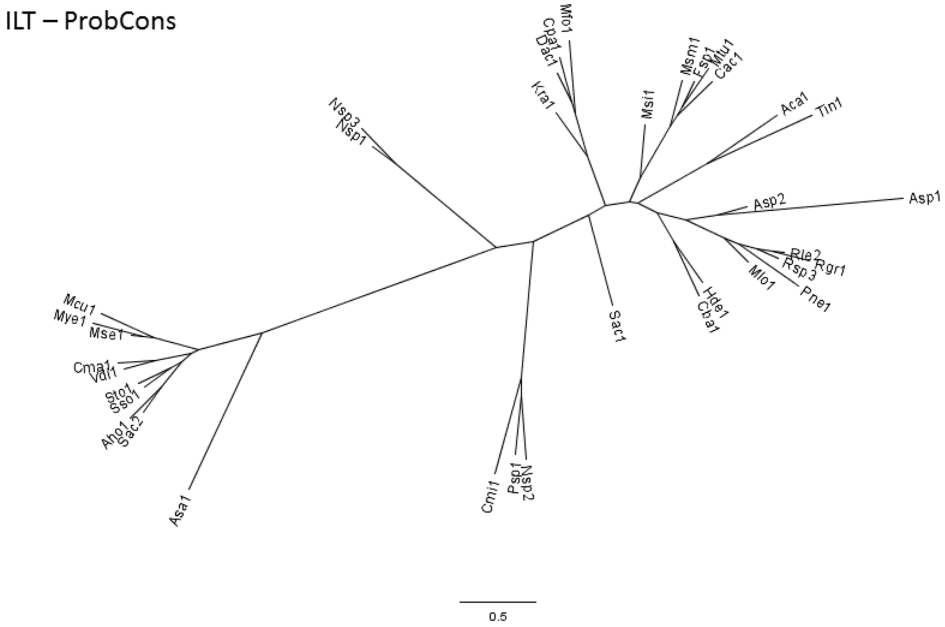


Supplemental Figure 22. Phylogenetic Trees of the MntP Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.

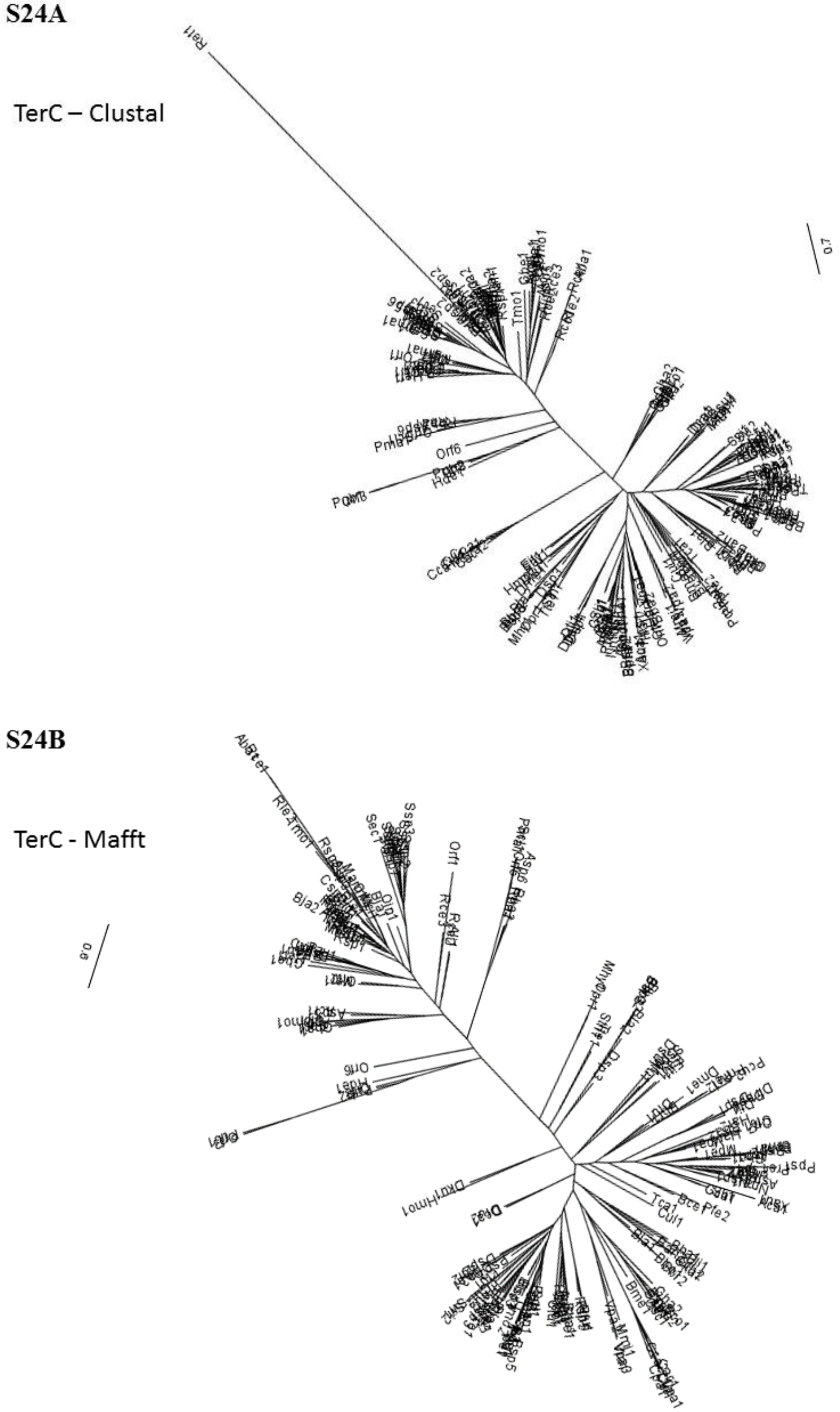


Supplemental Figure 23. Phylogenetic Trees of the ILT Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

S23C

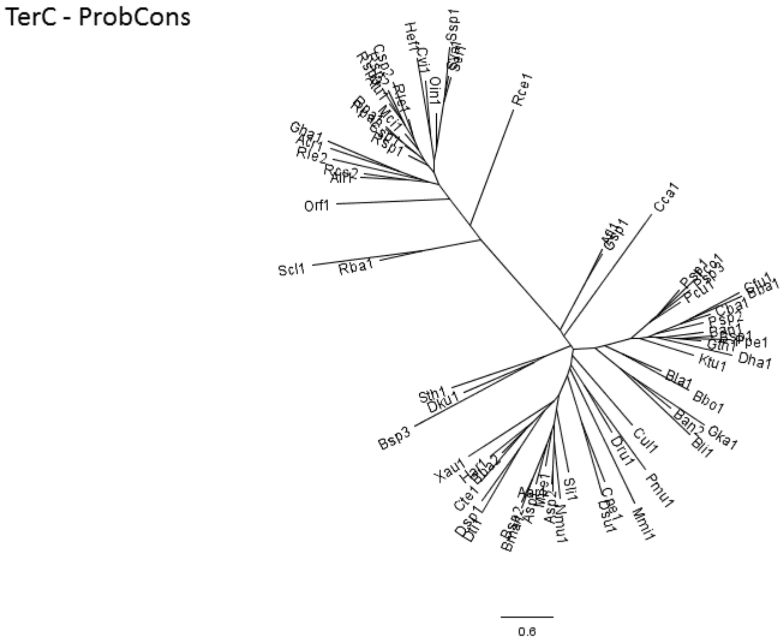


Supplemental Figure 23. Phylogenetic Trees of the ILT Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.



Supplemental Figure 24. Phylogenetic Trees of the TerC Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

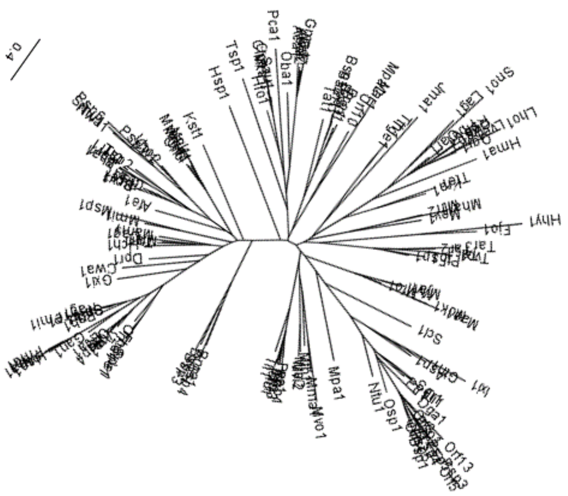
S24C


TerC - ProbCons

Supplemental Figure 24. Phylogenetic Trees of the TerC Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.

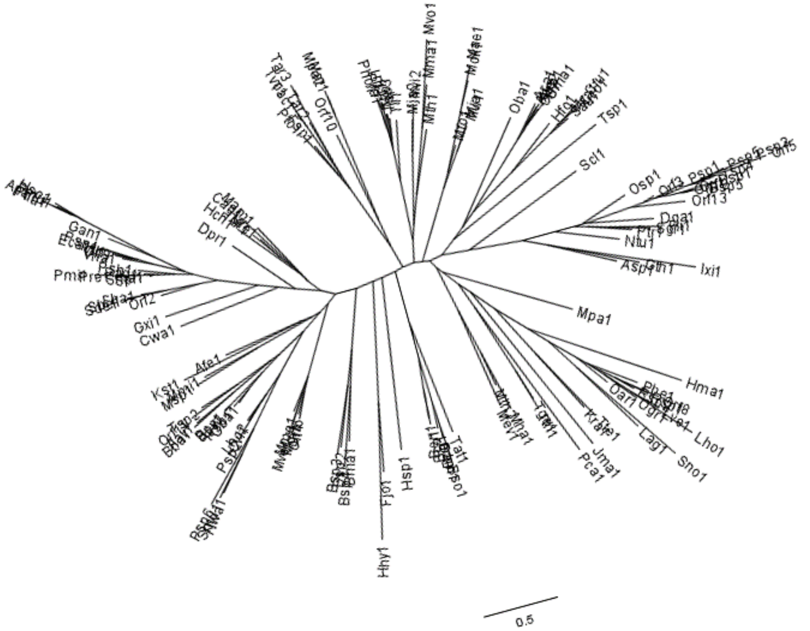
S25A

NAAT– Clustal



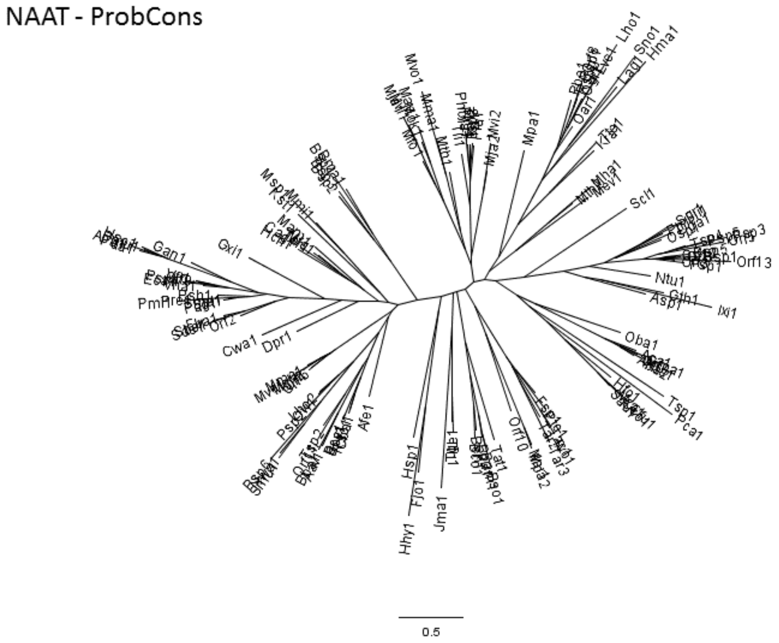
S25B

NAAT - Mafft



Supplemental Figure 25. Phylogenetic Trees of the NAAT Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

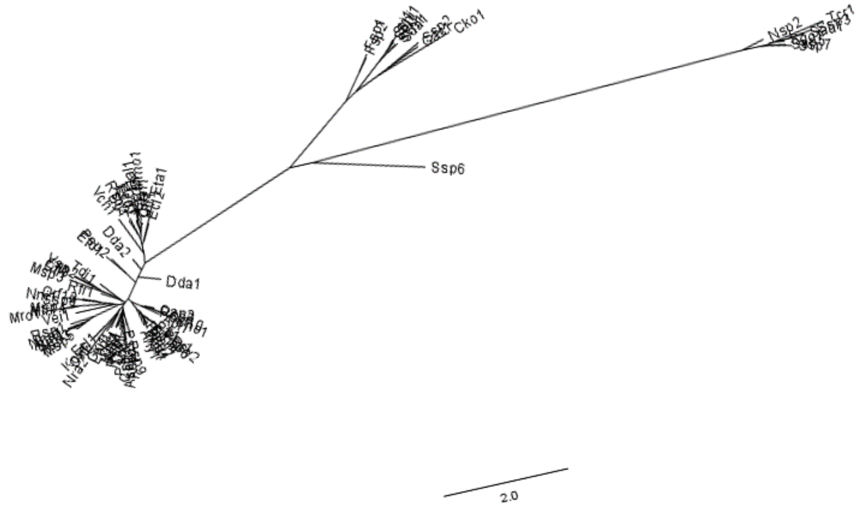
S25C



Supplemental Figure 25. Phylogenetic Trees of the NAAT Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.

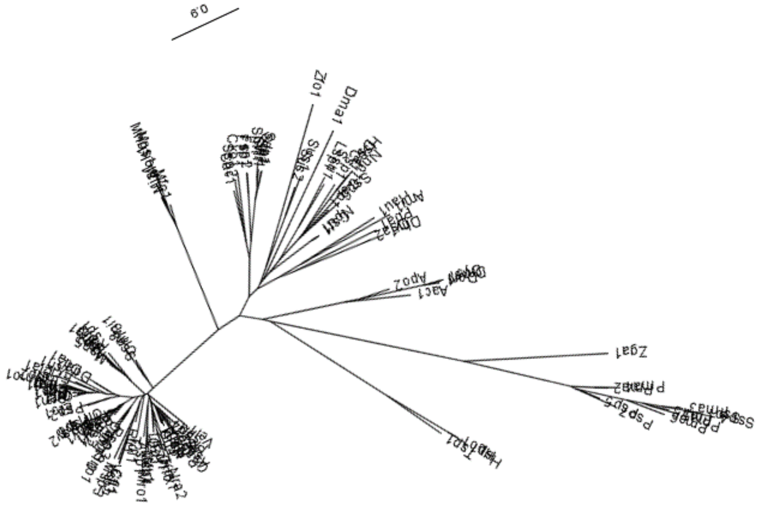
S26A

NicO – Clustal



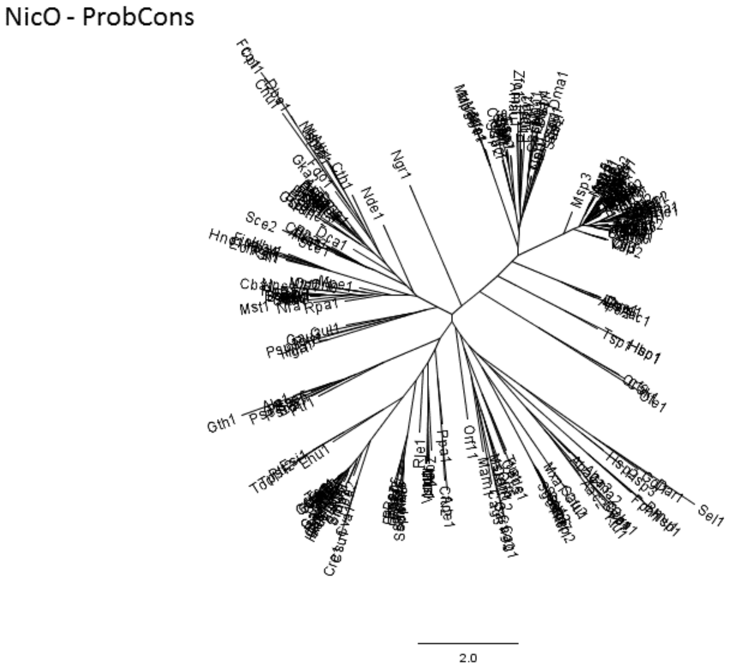
S26B

NicO - Mafft



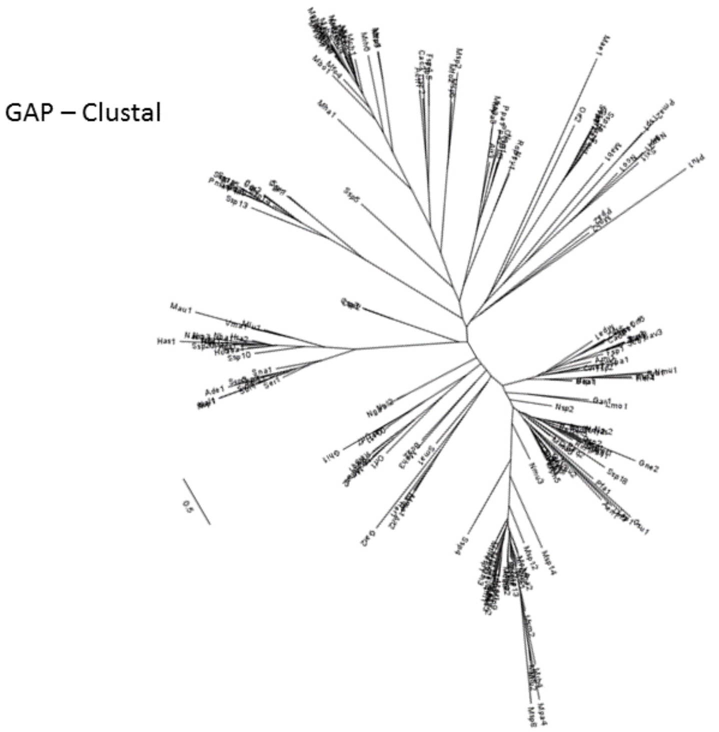
Supplemental Figure 26. Phylogenetic Trees of the NicO Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

S26C

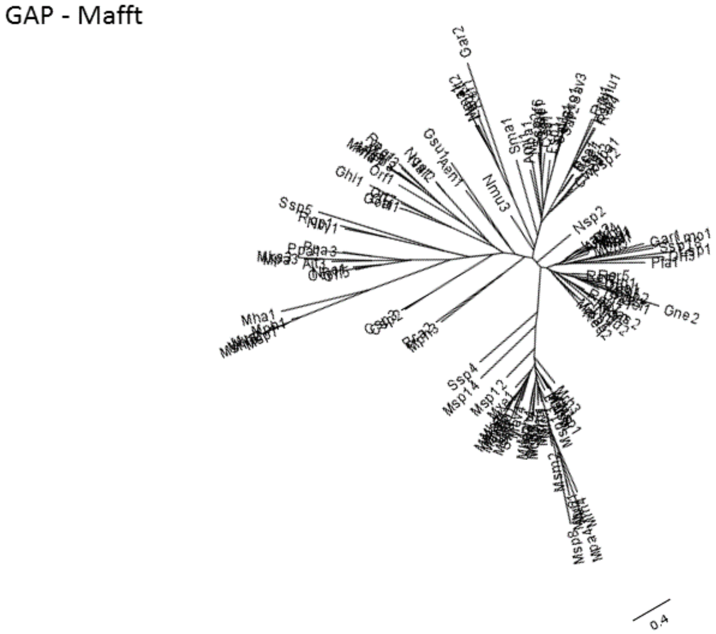


Supplemental Figure 26. Phylogenetic Trees of the NicO Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.

S27A

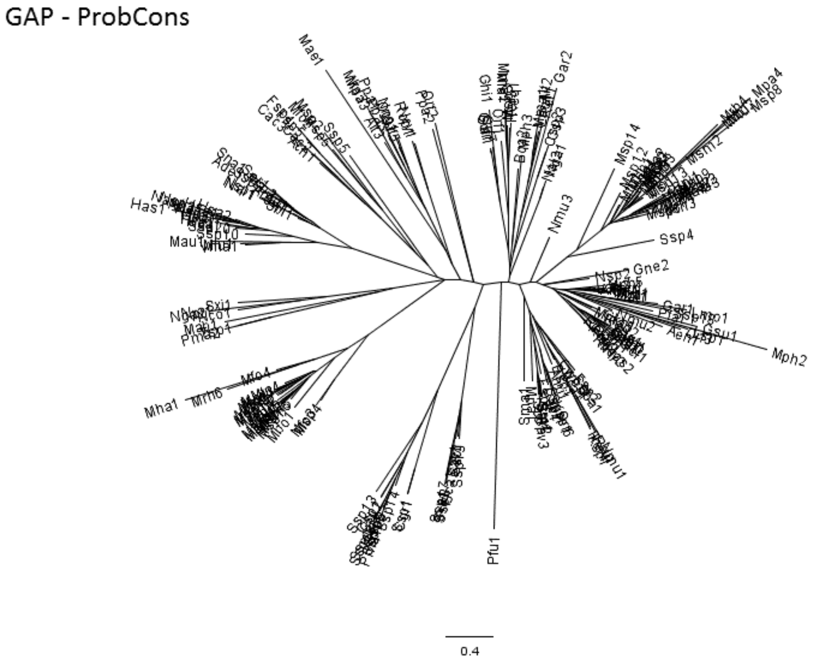


S27B



Supplemental Figure 27. Phylogenetic Trees of the GAP Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

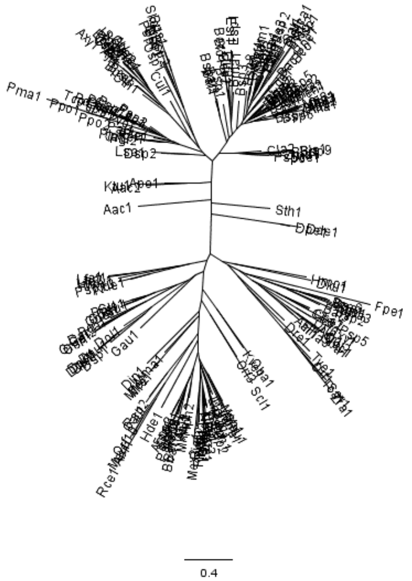
S27C



Supplemental Figure 27. Phylogenetic Trees of the GAP Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont.

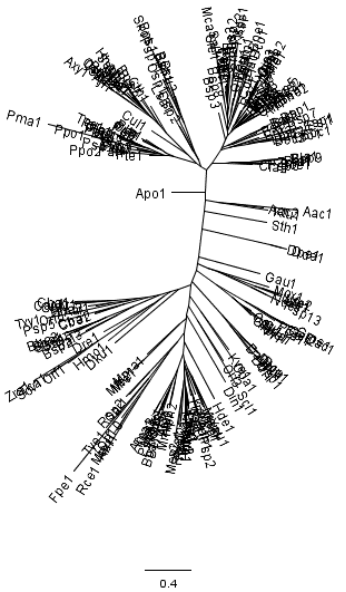
S28A

DsbD – Clustal



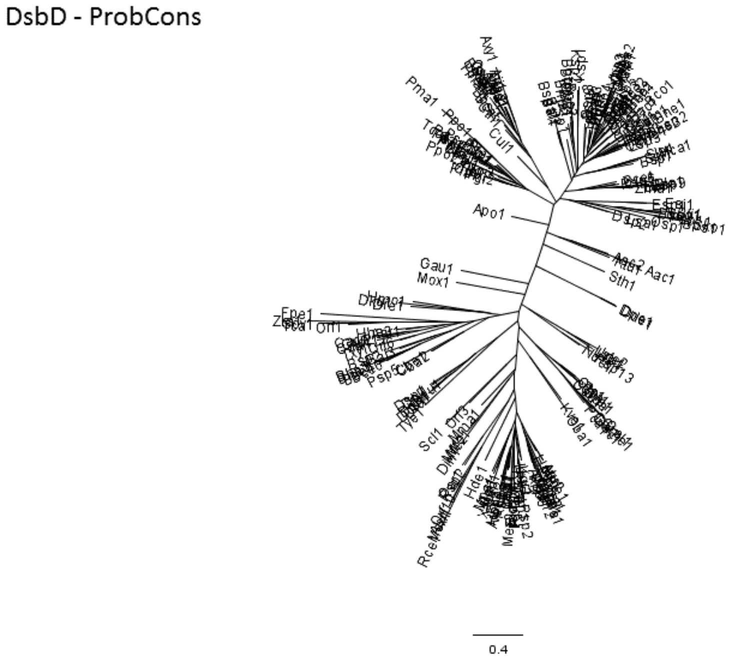
S28B

DsbD - Mafft



Supplemental Figure 28. Phylogenetic Trees of the DsbD Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons.

S28C



Supplemental Figure 28. Phylogenetic Trees of the DsbD Family based on multiple alignments generated with (A) ClustalX, (B) Mafft, (C) ProbCons, cont..

References

1. Vrljic M, Garg J, Bellmann A, Wachi S, Freudl R, Malecki MJ, Sahm H, Kozina VJ, Eggeling L, Saier MH Jr, Eggeling L, Saier MH Jr. (1999) The LysE superfamily: topology of the lysine exporter LysE of Corynebacterium glutamicum, a paradyme for a novel superfamily of transmembrane solute translocators. J Mol Microbiol Biotechnol 1: 327-336.

2. Zakataeva NP, Aleshin VV, Tokmakova IL, Troshin PV, Livshits VA (1999) The novel transmembrane Escherichia coli proteins involved in the amino acid efflux. FEBS Lett 452: 228-232.

3. Crupper SS, Worrell V, Stewart GC, Iandolo JJ (1999) Cloning and expression of cadD, a new cadmium resistance gene of Staphylococcus aureus. J Bacteriol 181: 4071-4075.

4. Bellmann A, Vrljic M, Patek M, Sahm H, Kramer R, Eggeling L. (2001) Expression control and specificity of the basic amino acid exporter LysE of Corynebacterium glutamicum. Microbiology 147: 1765-1774.

5. Broer S, Kramer R (1991) Lysine excretion by Corynebacterium glutamicum. 1. Identification of a specific secretion carrier system. European journal of biochemistry / FEBS 202: 131-135.

6. Broer S, Kramer R (1991) Lysine excretion by Corynebacterium glutamicum. 2. Energetics and mechanism of the transport system. European journal of biochemistry / FEBS 202: 137-143.

7. Vrljic M, Garg J, Bellmann A, Wachi S, Freudl R, Malecki MJ, Sahm H, Kozina VJ, Eggeling L, Saier MH Jr, Eggeling L, Saier MH Jr. (1999) The LysE superfamily: topology of the lysine exporter LysE of Corynebacterium glutamicum, a paradyme for a novel superfamily of transmembrane solute translocators. Journal of molecular microbiology and biotechnology 1: 327-336.

8. Vrljic M, Sahm H, Eggeling L (1996) A new type of transporter with a new type of cellular function: L-lysine export from Corynebacterium glutamicum. Molecular microbiology 22: 815-826.

9. Stabler N, Oikawa T, Bott M, Eggeling L (2011) Corynebacterium glutamicum as a host for synthesis and export of D-Amino Acids. Journal of bacteriology 193: 1702-1709.

10. Marbaniang CN, Gowrishankar J (2012) Transcriptional cross-regulation between Gram-negative and gram-positive bacteria, demonstrated using ArgP-argO of

Escherichia coli and LysG-lysE of Corynebacterium glutamicum. Journal of bacteriology 194: 5657-5666.

11. Nandineni MR, Gowrishankar J (2004) Evidence for an arginine exporter encoded by yggA (argO) that is regulated by the LysR-type transcriptional regulator ArgP in Escherichia coli. Journal of bacteriology 186: 3539-3546.

12. Burkovski A, Kramer R (2002) Bacterial amino acid transport proteins: occurrence, functions, and significance for biotechnological applications. Applied microbiology and biotechnology 58: 265-274.

13. Eggeling L, Sahm H (2003) New ubiquitous translocators: amino acid export by Corynebacterium glutamicum and Escherichia coli. Archives of microbiology 180: 155-160.

14. Schwartz JH, Maas WK (1960) Analysis of the inhibition of growth produced by canavanine in Escherichia coli. Journal of bacteriology 79: 794-799.

15. Hori H, Yoneyama H, Tobe R, Ando T, Isogai E, Katsumata R. (2011) Inducible L-alanine exporter encoded by the novel gene ygaW (alaE) in Escherichia coli. Applied and environmental microbiology 77: 4027-4034.

16. Vicente CM, Santos-Aberturas J, Guerra SM, Payero TD, Martín JF, Aparicio JF. (2009) PimT, an amino acid exporter controls polyene production via secretion of the quorum sensing pimaricin-inducer PI-factor in Streptomyces natalensis. Microbial cell factories 8: 33.

17. Kutukova EA, Livshits VA, Altman IP , Ptitsyn LR, Zyiatdinov MH, Tokmakova IL, Zakataeva NP. (2005) The yeaS (leuE) gene of Escherichia coli encodes an exporter of leucine, and the Lrp protein regulates its expression. FEBS letters 579: 4629-4634.

18. Franke I, Resch A, Dassler T, Maier T, Bock A (2003) YfiK from Escherichia coli promotes export of O-acetylserine and cysteine. Journal of bacteriology 185: 1161-1166.

19. Braun SD, Hofmann J, Wensing A, Ullrich MS , Weingart H, Völksch B, Spiteller D. (2010) Identification of the biosynthetic gene cluster for 3-methylarginine, a toxin produced by Pseudomonas syringae pv. syringae 22d/93. Applied and environmental microbiology 76: 2500-2508.

20. te Welscher YM, Jones L, van Leeuwen MR, Dijksterhuis J, de Kruijff B, Eitzen G, Breukink E. (2010) Natamycin inhibits vacuole fusion at the priming phase via a specific interaction with ergosterol. Antimicrobial agents and chemotherapy 54: 2618-2625.

21. Begg SL, Eijkelkamp BA, Luo Z, Counago RM, Morey JR, Maher MJ, Ong CL, McEwan AG, Kobe B, O'Mara ML, Paton JC, McDevitt CA. (2015) Dysregulation of transition metal ion homeostasis is the molecular basis for cadmium toxicity in Streptococcus pneumoniae. Nature communications 6: 6418.

22. Chen YY, Feng CW, Chiu CF, Burne RA (2008) cadDX operon of Streptococcus salivarius 57.I. Applied and environmental microbiology 74: 1642-1645.

23. Crupper SS, Worrell V, Stewart GC, Iandolo JJ (1999) Cloning and expression of cadD, a new cadmium resistance gene of Staphylococcus aureus. Journal of bacteriology 181: 4071-4075.

24. Ciobanu C, Slencu BG, Cuciureanu R (2012) Estimation of dietary intake of cadmium and lead through food consumption. Revista medico-chirurgicala a Societatii de Medici si Naturalisti din Iasi 116: 617-623.

25. Demaegd D, Foulquier F, Colinet AS, Gremillon L, Legrand D, Mariot P, Peiter E, Van Schaftingen E, Matthijs G, Morsomme P. (2013) Newly characterized Golgi-localized family of proteins is involved in calcium and pH homeostasis in yeast and human cells. Proc Natl Acad Sci U S A 110: 6859-6864.

26. Foulquier F, Amyere M, Jaeken J, Zeevaert R, Schollen E, Race V, Bammens R, Morelle W, Rosnoblet C, Legrand D, Demaegd D, Buist N, Cheillan D, Guffon N, Morsomme P, Annaert W, Freeze HH, Van Schaftingen E, Vikkula M, Matthijs G. (2012) TMEM165 deficiency causes a congenital disorder of glycosylation. Am J Hum Genet 91: 15-26.

27. Demaegd D, Foulquier F, Colinet AS, Gremillon L, Legrand D, Mariot P, Peiter E, Van Schaftingen E, Matthijs G, Morsomme P. (2013) Newly characterized Golgi-localized family of proteins is involved in calcium and pH homeostasis in yeast and human cells. Proceedings of the National Academy of Sciences of the United States of America 110: 6859-6864.

28. Foulquier F, Amyere M, Jaeken J, Zeevaert R, Schollen E, Race V, Bammens R, Morelle W, Rosnoblet C, Legrand D, Demaegd D, Buist N, Cheillan D, Guffon N, Morsomme P, Annaert W, Freeze HH, Van Schaftingen E, Vikkula M, Matthijs G. (2012) TMEM165 deficiency causes a congenital disorder of glycosylation. American journal of human genetics 91: 15-26.

29. Waters LS, Sandoval M, Storz G (2011) The Escherichia coli MntR miniregulon includes genes encoding a small protein and an efflux pump required for manganese homeostasis. J Bacteriol 193: 5887-5897.

30. Kaur G, Sengupta S, Kumar V, Kumari A, Ghosh A, Parrack P, Dutta D. (2014) Novel MntR-independent mechanism of manganese homeostasis in Escherichia coli by the ribosome-associated protein HflX. J Bacteriol 196: 2587-2597.

31. Guedon E, Moore CM, Que Q, Wang T, Ye RW, Helmann JD. (2003) The global transcriptional response of Bacillus subtilis to manganese involves the MntR, Fur, TnrA and sigmaB regulons. Molecular microbiology 49: 1477-1491.

32. Kehres DG, Maguire ME (2003) Emerging themes in manganese transport, biochemistry and pathogenesis in bacteria. FEMS microbiology reviews 27: 263-290.

33. Debut AJ, Dumay QC, Barabote RD, Saier MH, Jr. (2006) The iron/lead transporter superfamily of Fe/Pb2+ uptake systems. J Mol Microbiol Biotechnol 11: 1-9.

34. Singh A, Severance S, Kaur N, Wiltsie W, Kosman DJ (2006) Assembly, activation, and trafficking of the Fet3p.Ftr1p high affinity iron permease complex in Saccharomyces cerevisiae. The Journal of biological chemistry 281: 13355-13364.

35. Koch D, Chan AC, Murphy ME, Lilie H, Grass G, Nies DH. (2011) Characterization of a dipartite iron uptake system from uropathogenic Escherichia coli strain F11. The Journal of biological chemistry 286: 25317-25330.

36. Troxell B, Hassan HM (2013) Transcriptional regulation by Ferric Uptake Regulator (Fur) in pathogenic bacteria. Frontiers in cellular and infection microbiology 3: 59.

37. Burian J, Tu N, Kl'ucar L, Guller L, Lloyd-Jones G, Stuchlík S, Fejdi P, Siekel P, Turna J. (1998) In vivo and in vitro cloning and phenotype characterization of tellurite resistance determinant conferred by plasmid pTE53 of a clinical isolate of Escherichia coli. Folia Microbiol (Praha) 43: 589-599.

38. Anantharaman V, Iyer LM, Aravind L (2012) Ter-dependent stress response systems: novel pathways related to metal sensing, production of a nucleoside-like metabolite, and DNA-processing. Mol Biosyst 8: 3142-3165.

39. Azeddoug H, Reysset G (1994) Cloning and sequencing of a chromosomal fragment from Clostridium acetobutylicum strain ABKn8 conferring chemical-damaging agents and UV resistance to E. coli recA strains. Current microbiology 29: 229-235.

40. Ponnusamy D, Hartson SD, Clinkenbeard KD (2011) Intracellular Yersinia pestis expresses general stress response and tellurite resistance proteins in mouse macrophages. Veterinary microbiology 150: 146-151.

41. Sanssouci E, Lerat S, Grondin G, Shareck F, Beaulieu C (2011) tdd8: a TerD domain-encoding gene involved in Streptomyces coelicolor differentiation. Antonie van Leeuwenhoek 100: 385-398.

42. Kwon KC, Cho MH (2008) Deletion of the chloroplast-localized AtTerC gene product in Arabidopsis thaliana leads to loss of the thylakoid membrane and to seedling lethality. The Plant journal : for cell and molecular biology 55: 428-442.

43. Suzina NE, Duda VI, Anisimova LA, Dmitriev VV, Boronin AM (1995) Cytological aspects of resistance to potassium tellurite conferred on Pseudomonas cells by plasmids. Archives of microbiology 163: 282-285.

44. Akahane S, Kamata H, Yagisawa H, Hirata H (2003) A novel neutral amino acid transporter from the hyperthermophilic archaeon Thermococcus sp. KS-1. J Biochem 133: 173-180.

45. McDermott PF, McMurry LM, Podglajen I, Dzink-Fox JL, Schneiders T, et al. (2008) The marC gene of Escherichia coli is not involved in multiple antibiotic resistance. Antimicrob Agents Chemother 52: 382-383.

46. Marrero J, Auling G, Coto O, Nies DH (2007) High-level resistance to cobalt and nickel but probably no transenvelope efflux: Metal resistance in the Cuban Serratia marcescens strain C-1. Microb Ecol 53: 123-133.

47. Rodrigue A, Effantin G, Mandrand-Berthelot MA (2005) Identification of rcnA (yohM), a nickel and cobalt resistance gene in Escherichia coli. Journal of bacteriology 187: 2912-2916.

48. Ray P, Girard V, Gault M, Job C, Bonneu M, Mandrand-Berthelot MA, Singh SS, Job D, Rodrigue A (2013) Pseudomonas putida KT2440 response to nickel or cobalt induced stress by quantitative proteomics. Metallomics : integrated biometal science 5: 68-79.

49. Sonden B, Kocincova D, Deshayes C, Euphrasie D, Rhayat L, Laval F, Frehel C, Daffé M, Etienne G, Reyrat JM (2005) Gap, a mycobacterial specific integral membrane protein, is required for glycolipid transport to the cell surface. Mol Microbiol 58: 426-440.

50. Sonden B, Kocincova D, Deshayes C, Euphrasie D, Rhayat L, Laval F, Frehel C, Daffé M, Etienne G, Reyrat JM (2005) Gap, a mycobacterial specific integral membrane protein, is required for glycolipid transport to the cell surface. Molecular microbiology 58: 426-440.

51. Irani VR, Maslow JN (2005) Induction of murine macrophage TNF-alpha synthesis by Mycobacterium avium is modulated through complement-dependent interaction via complement receptors 3 and 4 in relation to M. avium glycopeptidolipid. FEMS microbiology letters 246: 221-228.

52. Villeneuve C, Etienne G, Abadie V, Montrozier H, Bordier C, Laval F, Daffe M, Maridonneau-Parini I, Astarie-Dequeker C (2003) Surface-exposed glycopeptidolipids of Mycobacterium smegmatis specifically inhibit the phagocytosis of mycobacteria by human macrophages. Identification of a novel family of glycopeptidolipids. The Journal of biological chemistry 278: 51291-51300.

53. Pourshafie M, Ayub Q, Barrow WW (1993) Comparative effects of Mycobacterium avium glycopeptidolipid and lipopeptide fragment on the function and ultrastructure of mononuclear cells. Clinical and experimental immunology 93: 72-79.

54. Sut A, Sirugue S, Sixou S, Lakhdar-Ghazal F, Tocanne JF, Lanéelle G (1990) Mycobacteria glycolipids as potential pathogenicity effectors: alteration of model and natural membranes. Biochemistry 29: 8498-8502.

55. Kimball RA, Martin L, Saier MH, Jr. (2003) Reversing transmembrane electron flow: the DsbD and DsbB protein families. J Mol Microbiol Biotechnol 5: 133-149.

56. Saier MH, Jr., Reddy VS, Tamang DG, Vastermark A (2014) The transporter classification database. Nucleic Acids Res 42: D251-258.

57. Zhai Y, Saier MH, Jr. (2001) 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.

58. Reddy VS, Saier MH, Jr. (2012) BioV Suite--a collection of programs for the study of transport protein evolution. FEBS J 279: 2036-2046.

59. 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.

60. Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17: 849-850.

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

62. Brenner S (1987) Of Urfs and Orfs - a Primer on How to Analyze Derived Amino-Acid-Sequences - Doolittle,Rf. Nature 329: 496-497.

63. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.

64. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution 30: 772-780.

65. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome research 15: 330-340.

66. Zhai Y, Saier MH, Jr. (2001) 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.

67. 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.

68. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202-208.

69. Silvestro D, Michalak I (2012) raxmlGUI: a graphical front-end for RAxML. Organisms Diversity & Evolution 12: 335-337.

70. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312-1313.

71. Chen JS, Reddy V, Chen JH, Shlykov MA, Zheng WH, Cho J, Yen MR, Saier MH Jr. (2011) Phylogenetic characterization of transport protein superfamilies: superiority of SuperfamilyTree programs over those based on multiple alignments. J Mol Microbiol Biotechnol 21: 83-96.

72. 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.

73. Yen MR, Chen JS, Marquez JL, Sun EI, Saier MH (2010) Multidrug resistance: phylogenetic characterization of superfamilies of secondary carriers that include drug exporters. Methods Mol Biol 637: 47-64.

74. Vastermark A, Wollwage S, Houle ME, Rio R, Saier MH, Jr. (2014) Expansion of the APC superfamily of secondary carriers. Proteins 82: 2797-2811.

75. Lee A, Vastermark A, Saier MH, Jr. (2014) Establishing homology between mitochondrial calcium uniporters, prokaryotic magnesium channels and chlamydial IncA proteins. Microbiology 160: 1679-1689.

76. Wong FH, Chen JS, Reddy V, Day JL, Shlykov MA, Wakabayashi ST, Saier MH Jr. (2012) The amino acid-polyamine-organocation superfamily. Journal of molecular microbiology and biotechnology 22: 105-113.

77. Reddy VS, Shlykov MA, Castillo R, Sun EI, Saier MH, Jr. (2012) The major facilitator superfamily (MFS) revisited. The FEBS journal 279: 2022-2035.

78. Reddy BL, Saier MH, Jr. (2013) Topological and phylogenetic analyses of bacterial holin families and superfamilies. Biochimica et biophysica acta 1828: 2654-2671.

79. Page RD (2002) Visualizing phylogenetic trees using TreeView. Curr Protoc Bioinformatics Chapter 6: Unit 6 2.

80. Palmieri F (2013) The mitochondrial transporter family SLC25: identification, properties and physiopathology. Mol Aspects Med 34: 465-484.

81. Denoncin K, Collet JF (2013) Disulfide bond formation in the bacterial periplasm: major achievements and challenges ahead. Antioxid Redox Signal 19: 63-71.

82. Kuan J, Saier MH, Jr. (1993) The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. Crit Rev Biochem Mol Biol 28: 209-233.

83. Kunji ER, Robinson AJ (2010) Coupling of proton and substrate translocation in the transport cycle of mitochondrial carriers. Curr Opin Struct Biol 20: 440-447.

84. Pebay-Peyroula E, Dahout-Gonzalez C, Kahn R, Trezeguet V, Lauquin GJ, Brandolin G (2003) Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. Nature 426: 39-44.

85. Klingenberg M (2008) The ADP and ATP transport in mitochondria and its carrier. Biochim Biophys Acta 1778: 1978-2021.

86. Haier B (2001) Funktionelle Analyse des Lysin-Exportcarriers aus Corynebacterium Glutamicum: Universität zu Köln. 83 p.