UNIVERSITY OF CALIFORNIA SAN DIEGO

Structural and evolutionary relationships within the ATP-binding cassette (ABC) superfamily

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

in

Biology

By

Nuo Tian

Committee in charge:

Professor Milton Saier, Chair
Professor James Golden
Professor Barry Grant

2021

The Thesis of Nuo Tian is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

Dedication

This study is dedicated to my beloved parents, Xiaoxing Zhu and Weijiang Tian who gave me strength when I thought of giving up and continually provided their emotional and financial support.

To my friends, colleagues and mentors who shared their advice and encouragement to finish this study.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to acknowledge my advisor, Dr. Milton Saier, whose knowledge and wisdom guided me along the process of completing my project.

I would like to acknowledge my co-advisor, Dr. Arturo Medrano-Soto, who spent day and night helping me with the project and the editing of this thesis. Without him, the completion of this thesis would have not been possible.

I would like to acknowledge my colleagues, Kevin Hendargo and Yichi Zhang for providing technical support and contributing ideas to this project. Without them, the completion of my thesis would have taken ten times as long.

I would like to acknowledge my husband, Cesar Humberto Nava Gonzales and my parents, Weijiang Tian and Xiaoxing Zhu for their emotional support throughout my project.

ABSTRACT OF THE THESIS

Structural and evolutionary relationships within the
ATP-binding cassette (ABC) superfamily

By

Nuo Tian

Master of Science in Biology

University of California San Diego, 2021

Professor Milton Saier, Chair

ATP-binding Cassette (ABC) transporters use ATP as an energy source and move a variety of substrates concentratively across cellular membranes. Previous studies based on primary protein sequence data suggested that integral membrane ABC exporters evolved independently at least three times, giving rise to three ABC types. Given the increasing availability of ABC structures in the Protein Data Bank (PDB) and the substantially larger number of primary sequences, we could investigate whether the current data support the conclusions obtained based on sequence analyses alone. We conducted sequence and structural analyses on the transmembrane domains (TMDs)

and the nucleotide binding (ATPase) domains (NBDs) of the three proposed ABC types, ABC1 in which a repeat unit of 2 TMSs triplicated, ABC2 in which a repeat unit of 3 TMSs duplicated, and ABC3 in which a repeat unit of 4 TMSs duplicated. The three most divergent families of the 70 ABC exporter families were excluded from our studies. The clustering patterns of both the TMDs and the NBDs showed that ABC1 forms a monophyletic group, whereas ABC2 and ABC3 share a major branch. The topological similarities of the two trees for the TMDs and NBDs strongly support the notion that these two domains have co-evolved. Based on sequence and structural divergence as well as organismal distribution, we suggest that ABC2s evolved first, followed by ABC1, and then ABC3. Our results provide insight into the evolutionary relationships between ABC types and serve as a guide for future studies of the ABC superfamily.

Introduction

ATP-binding cassette (ABC) transporters transport a wide range of substrates such as sugars, lipids, amino acids and macromolecules (Xiong et al. 2015). ABC exporters (such as P-glycoprotein) are also involved in transporting drugs out of the cells and are responsible for multidrug resistance (Glavinas et al. 2004). ABC transporters are featured by a transmembrane domain (TMD) and a nucleotide-binding domain (NBD). The NBD hydrolyzes ATP and provides energy for protein conformational changes between outward- and inward-facing orientations, therefore transporting substrates into and out of cells (Jones and George 2004). ABC transporters are present in both prokaryotes and eukaryotes. Prokaryotic type ABC systems usually have genes encoding the membrane protein and ATP-binding protein organized in operons, while eukaryotic type ABC transporters often have the transmembrane protein and the ATP-binding protein fused (Igarashi et al. 2004).

It is known that the NBD of the ABC transporters have highly conserved structures and sequences and are considered homologous, while the TMDs exhibit many different folds (ter Beek et al. 2014). The Saier lab previously proposed that the transmembrane domains of ABC exporters are polyphyletic, having evolved at least three times independently following different routes of evolution (Wang et al. 2009): the organization of α-helical transmembrane segments (TMSs) in ABC1 originated from a 2-TMS precursor that triplicated to give 6-TMS proteins; ABC2 originated from a 3 TMS precursor that duplicated to give a dissimilar set of 6-TMS proteins; and ABC3 originated from a 4-TMS precursor that duplicated to give 8-TMS, 10-TMS or 12-TMS proteins, where the extra two or four TMSs are in the middle between the two 4 TMS repeat units

(Wang et al. 2009). **Figure 1** illustrates the three types of proposed ABC exporters and the characteristic TMS topologies of their repeat units.

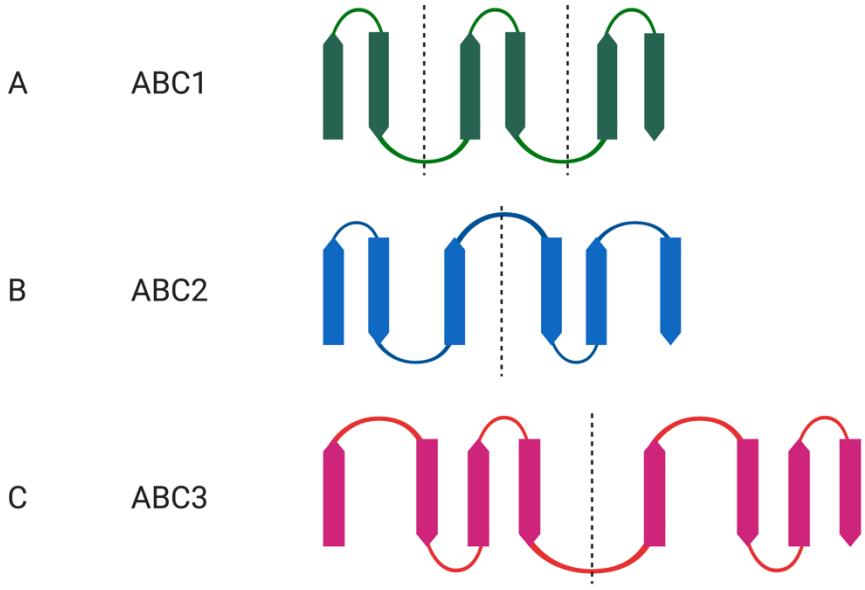

Figure 1. TMS topologies of ABC1, ABC2 and ABC3 exporters. Dashed lines separate the internal repeat units. This figure summarizes the overall topology; loop lengths are not indicative of the actual loop lengths in individual proteins.

There are 64 families of ABC efflux systems in TCDB (Transporter Classification Database, http://www.tcdb.org/) classified into these three types using the criteria mentioned above. This classification was based on sequence similarities. Although structural analyses of ABC porters have been attempted, they were inconclusive due to the lack of sufficient 3D structural data. With the increasing availability of 3D structures in the Protein Data Bank (PDB) (Berman et al. 2000) for members of the ABC superfamily, primary sequence data in public repositories and more advanced software tools, we set out to investigate whether the grouping of ABC transporters and the topologies of their repeat units are supported by the substantially larger amount of data currently available. We were interested in knowing whether analysis of 3D structural data agrees with the conclusions attained by sequence analysis for the three ABC types of membrane porters.

A priori, we expected that proteins within the same ABC type would have similar structures, but in addition, we wanted to test for structural similarities between ABC types. In particular, we were interested in providing structural support for the repeat units predicted by sequence-based analysis within each type. Although negative results for sequence and structural analyses are not sufficient to conclude a lack of homology, our confidence would substantially increase if both types of analyses support the independent origin of the three ABC membrane protein types.

Methods

All programs developed in the Saier laboratory can be downloaded from its public GitHub repository (https://github.com/SaierLaboratory).

Data selection

ABC sequences were downloaded from TCDB using the program extractFamily (Medrano-Soto et al. 2018). We first verified the quality of member assignments within ABC families in TCDB. This was achieved by 1) applying the program getDomainTopology (Medrano-Soto et al. 2020) to confirm that all family members share the characteristic Pfam domain(s) that cover the TMDs. In cases where the Pfam accessions covering the same region are different, they must belong to the same clan. If a protein had no direct hit with the characteristic domain(s) within its family, the program getDomainTopology attempts to "project" the domains of the specific family onto homologs lacking the expected domains; and 2) blasting family members against TCDB and confirming that they hit other families of the same type before bringing up families from other ABC types. Such selection ensures consistency within and among families of the same

type. The following three families were excluded from further analysis because their integral membrane protein constituents are highly divergent from all other members in the ABC superfamily, and failed to satisfy our criteria: the UDP-Glucose/Iron Exporter (U-GlcE) Family (TC# 3.A.1.139; ABC1), the Peroxysomal Fatty Acyl CoA Transporter (P-FAT) Family (TC# 3.A.1.203; ABC1), and the lipopolysaccharide export (LptBFG) Family (TC# 3.A.1.152; ABC2). **Table1** shows the proteins that satisfied our requirements and were used in this analysis.

Table 1. List of ABC families used in this study. Column 1 provides the TC number of the family; column 2 is the family name; column 3 shows the corresponding ABC type as assigned in TCDB. Column 4 indicates the clan of the most common Pfam domain covering the TMDs in that family. The ATPase domains of the three types belong to the same clan: CL0023.

TC Number	Family Name	ABC Type	Pfam Clan
3.A.1.106	The Lipid Exporter (LipidE) Family	1	CL0241
3.A.1.108	The β-Glucan Exporter (GlucanE) Family	1	CL0241
3.A.1.109	The Protein-1 Exporter (Prot1E) Family	1	CL0241
3.A.1.110	The Protein-2 Exporter (Prot2E) Family	1	CL0241
3.A.1.111	The Peptide-1 Exporter (Pep1E) Family	1	CL0241
3.A.1.112	The Peptide-2 Exporter (Pep2E) Family	1	CL0241
3.A.1.113	The Peptide-3 Exporter (Pep3E) Family	1	CL0241
3.A.1.117	The Drug Exporter-2 (DrugE2) Family	1	CL0241
3.A.1.118	The Microcin J25 Exporter (McjD) Family	1	CL0241
3.A.1.119	The Drug/Siderophore Exporter-3 (DrugE3) Family	1	CL0241
3.A.1.123	The Peptide-4 Exporter (Pep4E) Family	1	CL0241
3.A.1.127	The AmfS Peptide Exporter (AmfS-E) Family	1	CL0241
3.A.1.129	The CydDC Cysteine Exporter (CydDC-E) Family	1	CL0241

Table 1 continued

3.A.1.135	The Drug Exporter-4 (DrugE4) Family	1	CL0241
3.A.1.201	The Multidrug Resistance Exporter (MDR) Family (ABCB)	1	CL0241
3.A.1.202	The Cystic Fibrosis Transmembrane Conductance Exporter (CFTR) Family (ABCC)	1	CL0241
3.A.1.206	The a-Factor Sex Pheromone Exporter (STE) Family (ABCB)	1	CL0241
3.A.1.208	The Drug Conjugate Transporter (DCT) Family (ABCC)	1	CL0241
3.A.1.209	The MHC Peptide Transporter (TAP) Family (ABCB)	1	CL0241
3.A.1.210	The Heavy Metal Transporter (HMT) Family (ABCB)	1	CL0241
3.A.1.212	The Mitochondrial Peptide Exporter (MPE) Family (ABCB)	1	CL0241
3.A.1.21	The Siderophore-Fe^{3+} Uptake Transporter (SIUT) Family	1	CL0241
3.A.1.101	The Capsular Polysaccharide Exporter (CPSE) Family	2	CL0181
3.A.1.102	The Lipooligosaccharide Exporter (LOSE) Family	2	CL0181
3.A.1.103	The Lipopolysaccharide Exporter (LPSE) Family	2	CL0181
3.A.1.104	The Teichoic Acid Exporter (TAE) Family	2	CL0181
3.A.1.105	The Drug Exporter-1 (DrugE1) Family	2	CL0181
3.A.1.107	The Putative Heme Exporter (HemeE) Family	2	CL0181
3.A.1.115	The Na^+ Exporter (NatE) Family	2	CL0181
3.A.1.116	The Microcin B17 Exporter (McbE) Family	2	CL0181 *
3.A.1.124	The 3-component Peptide-5 Exporter (Pep5E) Family	2	CL0181
3.A.1.126	The β-Exotoxin I Exporter (βETE) Family	2	CL0181
3.A.1.128	The SkfA Peptide Exporter (SkfA-E) Family	2	CL0181
3.A.1.130	The Multidrug/Hemolysin Exporter (MHE) Family	2	CL0181
3.A.1.131	The Bacitracin Resistance (Bcr) Family	2	CL0181
3.A.1.132	The Gliding Motility ABC Transporter (Gld) Family	2	CL0181
3.A.1.133	The Peptide-6 Exporter (Pep6E) Family	2	CL0181

Table 1 continued

3.A.1.138	The Unknown ABC-2-type (ABC2-1) Family	2	CL0181
3.A.1.141	The Ethyl Viologen Exporter (EVE) Family (DUF990 Family)	2	CL0181
3.A.1.142	The Glycolipid Flippase (G.L.Flippase) Family	2	CL0181
3.A.1.144	Functionally Uncharacterized ABC2-1 (ABC2-1) Family	2	CL0181
3.A.1.145	Peptidase Fused Functionally Uncharacterized ABC2-2 (ABC2-2) Family	2	CL0181
3.A.1.146	The actinorhodin (ACT) and undecylprodigiosin (RED) exporter (ARE) family	2	CL0181
3.A.1.147	Functionally Uncharacterized ABC2-2 (ABC2-2) Family	2	CL0181
3.A.1.148	Functionally Uncharacterized ABC2-3 (ABC2-3) Family	2	CL0181 *
3.A.1.149	Functionally Uncharacterized ABC2-4 (ABC2-4) Family	2	CL0181
3.A.1.150	Functionally Uncharacterized ABC2-5 (ABC2-5) Family	2	CL0181 ^
3.A.1.151	Functionally Uncharacterized ABC2-6 (ABC2-6) Family	2	CL0181 ^
3.A.1.204	The Eye Pigment Precursor Transporter (EPP) Family (ABCG)	2	CL0181
3.A.1.205	The Pleiotropic Drug Resistance (PDR) Family (ABCG)	2	CL0181
3.A.1.211	The Cholesterol/Phospholipid/Retinal (CPR) Flippase Family (ABCA)	2	CL0181
3.A.1.114	The Probable Glycolipid Exporter (DevE) Family	3	CL0404
3.A.1.122	The Macrolide Exporter (MacB) Family	3	CL0404
3.A.1.125	The Lipoprotein Translocase (LPT) Family	3	CL0404
3.A.1.134	The Peptide-7 Exporter (Pep7E) Family	3	CL0404
3.A.1.136	The Uncharacterized ABC-3-type (U-ABC3-1) Family	3	CL0404
3.A.1.137	The Uncharacterized ABC-3-type (U-ABC3-2) Family	3	CL0404
3.A.1.140	The FtsX/FtsE Septation (FtsX/FtsE) Family	3	CL0404
3.A.1.207	The Eukaryotic ABC3 (E-ABC3) Family	3	CL0404

* Pfam matches were below gathering threshold.
^ The Pfam domain and clan were marginally projected (E-value < 10^{-2}) from the closest ABC2 family.

Phylogenetic and sequence similarity trees

To maintain the computational time within reasonable limits, a sample of 50 sequences per major group in the protein tree (ABC1, ABC2a, ABC2b and ABC3) in **Figure 4** were selected to build phylogenies using MrBayes (Ronquist and Huelsenbeck 2003) as well as the Maximum Likelihood, Neighbor Joining and Fitch methods from the Phylip suite (Felsenstein 1989). Multiple alignments were generated with MAFFT (Katoh and Standley 2013) using the L-INS-i algorithm. Uninformative positions in the multiple alignment were removed with the program Trimal (Capella-Gutierrez et al. 2009) to keep positions with less than 30% gaps. Using MrBayes, we assumed different substitution rates among sites and followed a gamma distribution with 4 rate categories. Posterior probabilities were estimated with Metropolis coupling (1 cold and 3 heated chains), and 2,000,000 generations were used to lower the average standard deviation of split frequencies below 0.01. Phylogenetic trees created using the Phylip suite were built using the programs NEIGHBOR, FITCH and PROML with 100 bootstrap replicas. Due to the amount of sequence diversity in the ABC superfamily, we were unable to build reliable trees. MrBayes did not converge, and the average standard deviation of split frequencies was greater than 0.25 (well above the recommended threshold of 0.01in the manual). We even obtained the same result for trees generated using the ATPase domains, which are known to be homologous (see Figure 5 and discussion in the text). Phylip trees did not generate significant bootstrap support for key branches. Therefore, we continued the analysis with the protein trees generated with the program mkProteinClusters.

To study the relationships of both TMDs and NBDs within and among ABC families, we used our in-house program mkProteinClusters (Medrano-Soto et al. 2018) to cluster representative protein sequences based on pairwise sequence similarity scores. Membrane proteins and ATPases

were treated separately. For fusion proteins containing both TMDs and the ATPase domain, the TMDs were manually cut with the aid of the in-house program phoboshop. This program provides a graphical user interface that facilitates cutting segments of proteins. Regions matching the Pfam accessions of the ATP-binding domains of ABC transporters (PF0005) were also extracted for clustering. The program mkProteinClusters compares bit scores of pairwise Smith-Waterman alignments as generated by SSEARCH (Pearson 1991) and uses the statistical environment R (https://www.r-project.org) to produce a hierarchical clustering tree of the input sequences. We applied the Ward agglomerative method for both membrane proteins and ATPases because it generated the trees that best separated ABC types compared to other methods (i.e., Average, Weighted, Single, Complete). To identify the tree with the most robust topology, we tested different SSEARCH parameters (z = 1, 11, 21; k = 500, 1000; s = BL50, BL62) and selected the tree topology supported by at least 85% of the trees generated. Trees were drawn with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Repeat unit analyses

We first expanded the number of sequences of each ABC type by blasting member proteins in TCDB against the NCBI non-redundant protein database using the program famXpander (Medrano-Soto et al. 2018). Then we generated a multiple alignment for each ABC type using the L-INS-i algorithm in MAFFT (Katoh and Standley 2013) and trimmed the alignment by removing positions with more than 30% gaps using trimAL (Capella-Gutierrez et al. 2009). We used AveHAS (Zhai and Saier 2001) to visualize the average hydropathy of the multiple alignment and used AncientRep (Reddy and Saier 2012) to search for repeats within each ABC type. When we

used an AncientRep cut after the fourth TMS in ABC3 proteins, we found high-scoring (GSAT: 25, E-value: 5.7e-14) and clean hydropathy alignments supporting 4+4 and 4+2+4 topologies. However, because repeat units within single proteins are poorly conserved in ABC types 1 and 2, we took advantage of the ability of AncientRep to search for repeats in different sequence regions of members of the same family. For example, if TMSs 1-3 of one protein are significantly similar to TMSs 4-6 of another protein, and the two proteins align throughout their lengths, an internal repeat of three TMSs can be inferred (Reddy and Saier 2012). For ABC1 and ABC2, we compared the first two TMSs and first three TMSs, respectively, with the rest of the proteins to search for repeat units. If the first two TMSs align with the third and the fourth TMSs (or with the fifth and the sixth TMSs), it counts as evidence for 2+2+2 topology. We only considered alignments of two or three TMSs. The criteria for identifying alignments that support alternative repeat unit topologies are shown in **Table 2**. Given that the program HMMTOP (Tusnady and Simon 2001) frequently mispredicts TMSs in ABC members, we used the in-house program tmweaver to obtain the coordinates of hydrophobic peaks (inferred TMSs) in individual proteins based on an input multiple alignment. In the AncientRep output, we considered 2 TMSs aligned if at least 10 to 15 residues of both TMSs were aligned. The number of alignments supporting topologies 2+2+2 and 3+3 was counted, and their average GSAT scores were calculated.

Table 2. Criteria for inferring the internal repeat unit of ABC1 and ABC2 proteins. The first column represents the TMS aligned before the AncientRep cut, and the first row represents the TMS aligned after the AncientRep cut. **A.** AncientRep cut after the second TMS. The first two TMS were compared with the rest of the protein. **B.** AncientRep cut after the third TMS. The first three TMSs were compared to the rest of the protein.

A.

Right of AncientRep cut / Left of AncientRep cut	TMS3-4	TMS4-5	TMS5-6
TMS1-2	2+2+2	3+3	2+2+2

B.

Right of AncientRep cut / Left of AncientRep cut	TMS4-5	TMS5-6	TMS4-6
TMS1-2	3+3	2+2+2	conflict
TMS2-3	conflict	3+3	conflict
TMS1-3	conflict	conflict	3+3

It is well known that sequence alignments involving unrelated integral membrane proteins can artificially yield scores beyond thresholds of significance, due to biases toward hydrophobic residues introduced by physicochemical constraints in the membrane environment (Wong et al. 2010; Wong et al. 2011). In order to control for this possibility, we used the in-house program quicklsat to construct a negative control of randomized proteins that preserve the amino acid composition and TMS topology of reference ABC proteins. The program achieves this by shuffling residues in TMSs and loops separately while preserving their original positions. We aligned the full ABC protein with the shuffled protein and confirmed that the two proteins have similar TMS topologies but produce insignificant E-values. Because of the poor sequence similarity among

shuffled proteins, we were unable to generate meaningful multiple alignments. However, given that we know a priori the position of the TMSs, we made cuts after the second and third TMS for all ABC1 and ABC2 shuffled proteins, and aligned them with the rest of the sequences to seek internal repeats. Evidence supporting repeat units is identified when alignments between different segments of real proteins show higher scores than alignments between different segments of shuffled proteins.

We also used the program HHrepID (Biegert and Soding 2008) to search for repeat units within the three ABC types using maximum 3 PSI-BLAST (Altschul et al. 1997) iterations to build multiple sequence alignments, considering secondary structure inferences with PSIPRED (Jones 1999), 3 merge rounds before repeats are inferred from posterior probabilities, and a repeat P-value threshold $< 10^{-2}$.

Structural analysis

All available 3D structures for proteins in **Table 1** were fetched from the Protein Data Bank (PDB). All structures were then cut into 4-helix bundles (4HB) (see **Figure 2**) and aligned using our in-house program Deuterocol (Medrano-Soto et al. 2020), which performs structural superpositions with the programs Superpose (Krissinel and Henrick 2004) and TM-align (Zhang and Skolnick 2005). We chose alignments of 4HBs because 1) we frequently observed significant superpositions of 3HBs between unrelated structures, and 2) the largest repeat unit observed in the ABC type 3 superfamily is 4 TMSs.

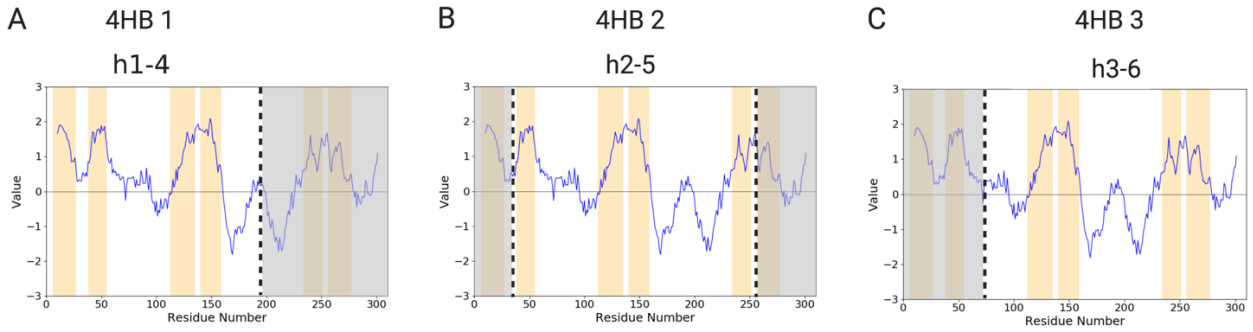

Figure 2. Cutting ABC structures into 4HBs. For simplicity, sequences were used to illustrate the process, but actual cuttings were performed on 3D structures. For a protein with 6 transmembranal α-helices (TMSs; shown as hydrophobic peaks and highlighted with tan bars), A. the first 4HB corresponds to TMSs 1-4. B. the second 4HB corresponds to TMSs 2-5. C. the third 4HB corresponds of TMSs 3-6. Dotted vertical lines represent cutting points. Grayed areas represent removed regions.

For each alignment, the root-mean-square deviation (RMSD) score, TM-score, and coverage were calculated to determine the significance of the alignment. RMSD is the root-mean-square distance between corresponding atoms after an optimal rotation of one structure relative to another. To assess the overall quality of alignments, the number of residues in the alignment relative to the size of the bundles (coverage) was divided by the RMSD. The higher the coverage and the smaller the RMSD value, the better the alignment. The TM-score, on the other hand, weights the residue pairs at smaller distances relatively more strongly than those at larger distances (Zhang and Skolnick 2005). We used RMSD < 4 Å, TM-Score > 0.55 and coverage > 75% as cutoffs to consider an alignment significant. Top alignments of each comparison were filtered and visually inspected with PyMOL (https://pymol.org).

Each pair of aligned structures was represented by the pair of 4HBs that yielded the highest coverage and lower RMSD. The significance of the alignment between structures *i* and *j* was calculated as the similarity score

$$S_{i,j} = \frac{Cov_{i,j}}{RMSD_{i,j}} : i \neq j,$$

where coverage represents the region of the 4HB that is involved in the alignment relative to the shorter 4HB. Only the regions of the TMSs within the membrane plane were considered for the calculation of coverage. The plane of the membrane relative to the structures were extracted for the OPM (Orientations of Proteins in Membranes) database (Lomize et al. 2012). $S_{i,j}$ increases proportionally with the coverage and is inversely proportional to the RMSD. To cluster the structures, we first calculated the normalized similarity score $N_{i,j}$ such that

$$N_{i,j} = \frac{S_{i,j}}{\max\left(S_{x,y}: x,y=1..n\right)} \text{ and } N_{i,j} = 1 : i=j,$$

where *n* is the total number of structures in the analysis that had at least one significant alignment with other structures. Finally, we estimated the dissimilarity $D_{i,j}$ between each pair of structures as

$$D_{i,j} = 1 - N_{i,j}.$$

We applied hierarchical clustering to the structural data based on the dissimilarity metric $D_{i,j}$ using the Statistical computing environment R (https://www.r-project.org) and the Ward agglomerative method to minimize the within-cluster variance.

Similarly, we also fetched all the available ATPase structures in the families listed in **Table 1** from PDB. Instead of cutting them into 4HB, we used only the regions coding for the ATPases because they are considered monophyletic and structurally highly similar. RMSD and coverage were calculated, and hierarchical clustering was performed as described above.

We also searched for a structural repeat unit for ABC1 and ABC2 proteins; ABC3 proteins have no available structures with 8 or 10 TMSs and therefore were not examined. Based on **Figure**

2, if the topology of the repeat unit is 2+2+2, a good 2-helix alignment between 4HBs 1 and 3 would be observed because helices 1 and 2 should align with helices 3 and 4, and helices 3 and 4 should align with helices 5 and 6. To search for 3+3 topologies, we decided to cut structures in 3-helix bundles (3HBs) and search for high-scoring 3-TMS alignments. In this case, we scanned for alignments between helices 1-3 and helices 4-6.

Results

Protein tree of ABC sequences

All protein sequences in the three ABC types were extracted from TCDB and split into transmembranal and ATPase domains. To comprehensibly study the collective evolutionary histories of TMDs in ABC exporters, we first assumed that the three ABC types are homologous and attempted to build phylogenetic trees including all three types (see **Methods**). However, our attempts to construct phylogenies were unsuccessful as they did not converge properly due to the high level of sequence diversity within and among types. Therefore, we generated cross-type trees for each individual domain based on the bit scores of Smith-Waterman pairwise alignments as in previous reports (Medrano-Soto et al. 2018; Medrano-Soto et al. 2020; Wang et al. 2020). The individual clustering trees for the transmembrane and ATPase domains are shown in **Figure 3** and **Figure 4**, respectively.

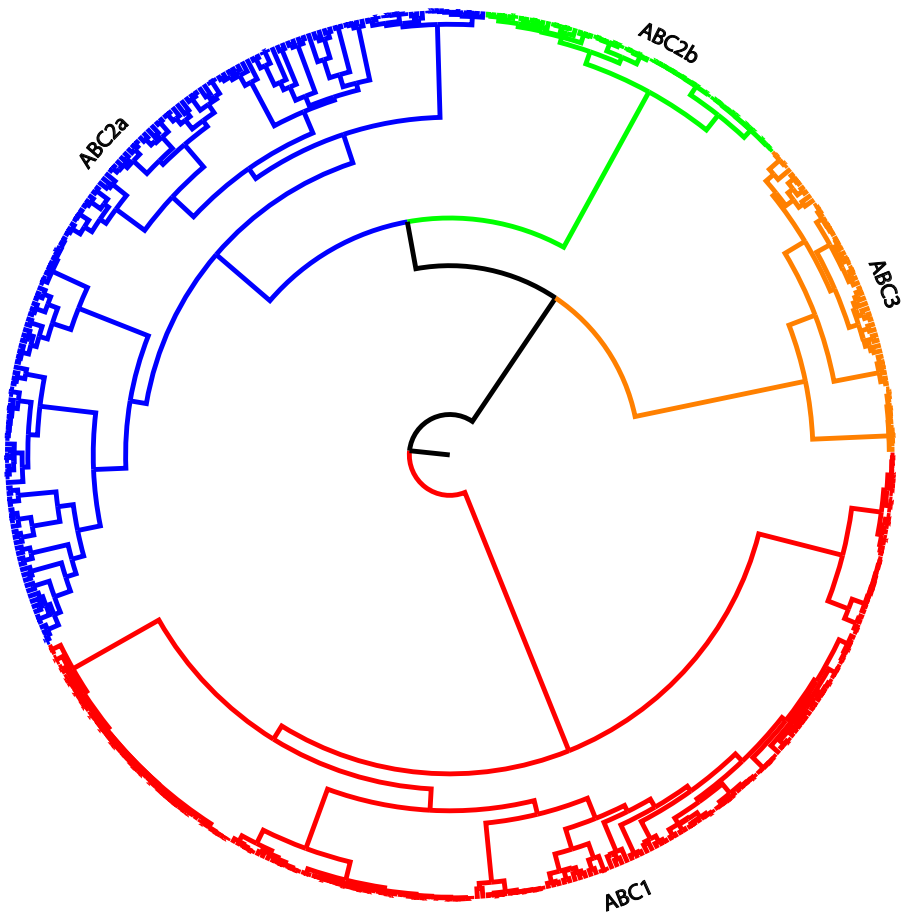

Figure 3. Tree for the transmembrane domains of the three ABC types. Only the hydrophobic regions containing TMSs were used to construct this tree (see **Methods**). The tree has strong clustering structure (Agglomerative coefficient 0.992). No scale bar is provided because only the topology is meaningful. The groups ABC2a and ABC2b are highlighted to facilitate comparison with the tree for ATPase domains in Figure 4. See text for discussion.

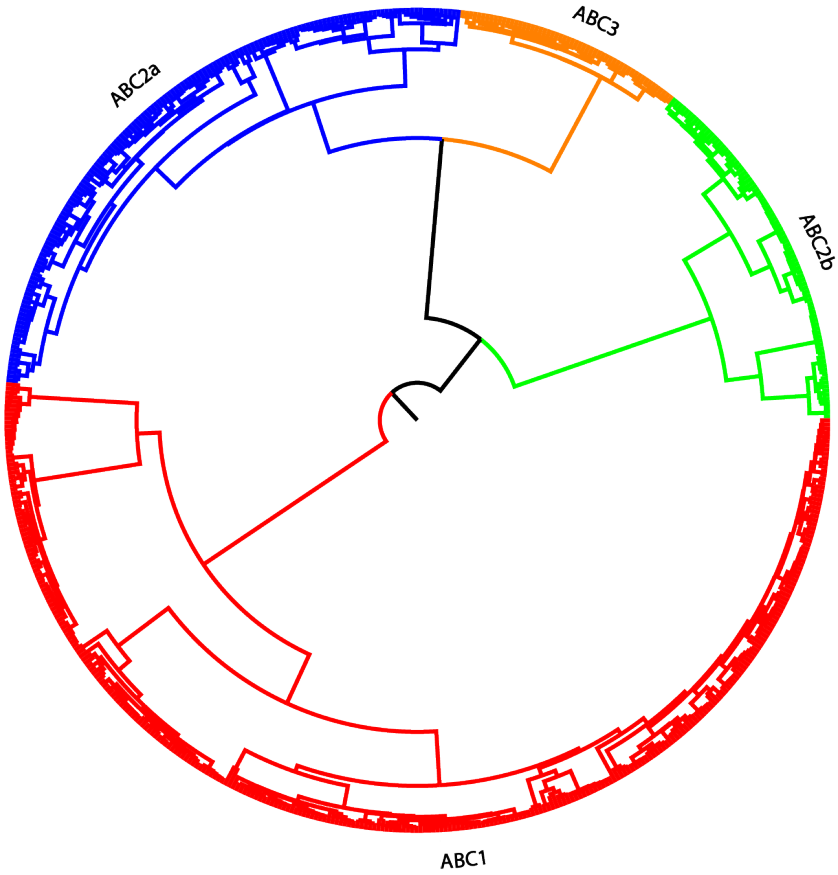

Figure 4. Tree for ATPase domains of the three ABC types. Only the ATPase domains were used to construct this tree (see **Methods**). Notice how ABC2 proteins are split into groups ABC2a and ABC2b, which can also be identified in Figure 3. The tree has a strong clustering structure (Agglomerative coefficient 0.980). No scale bar is provided because only the topology is meaningful. See text for discussion.

The strong clustering structures of the trees in **Figures 3** and **4** (Agglomerative coefficients 0.992 and 0.980, respectively), shows that ABC1 forms a monophyletic group, while ABC2 and ABC3 share a major branch. In both figures, ABC2 is internally separated into groups ABC2a and ABC2b. This can be explained for the most part by the composition of prokaryotic/eukaryotic proteins within this type. It has been shown that eukaryotic ABC transporters and prokaryotic ABC transporters differ in terms of their genes and domain organizations (Igarashi et al. 2004). Group

ABC2b consists exclusively of two eukaryotic families, the Eye Pigment Precursor Transporter (EPP) Family (TC# 3.A.1.204) and the Pleiotropic Drug Resistance (PDR) Family (TC# 3.A.1.205), while ABC2a consists mainly of prokaryotic proteins except for one eukaryotic family, The Cholesterol/Phospholipid/Retinal (CPR) Flippase Family (TC# 3.A.1.211), which is in fact the most distant within that clade. ABC3 contains mostly prokaryotic proteins, with one exception, the Eukaryotic ABC3 (E-ABC3) Family (TC# 3.A.1.207) having a mixture of archaeal and eukaryotic proteins. As mentioned in the **Introduction,** prokaryotic type ABC systems are usually encoded by genes encoding the integral membrane protein and the ATP-binding protein organized in operons, while eukaryotic type ABC transporter genes often have the transmembrane protein and the ATP-binding domains either fused in a single gene product, or the two genes map separately in the genome. In **Figure 3**, ABC2 is monophyletic and ABC2a and ABC2b share the same branch. This can be explained by the fact that the best alignment of TMDs between ABC2a and ABC2b has a better, albeit comparable, score (E-value = 3.2×10^{-6}) than the best alignments for ABC2a-ABC3 (E-value = 3.8×10^{-5}) or ABC2b-ABC3 (E-value = 2.8×10^{-3}). In contrast, the ATPase tree in **Figure 4** shows that the ABC2a group shares a major branch with ABC3, separating it from the ABC2b group. In this case, the ATPase domains in ABC2a show marginally better alignment scores with ABC3 (E-value = 6.5×10^{-23}) compared to ABC2b (E-value = 3.0×10^{-21}). Furthermore, we observed that the top alignment (E-value = 1.7×10^{40}) between prokaryotic and eukaryotic ATPases for ABC1 and the top alignment (E-value = 1.5×10^{35}) between prokaryotic eukaryotic ATPases for ABC3 are significantly better than the top score between prokaryotic and eukaryotic ATPases for ABC2 (E-value = 3.0×10^{-21}). Therefore, the ATPase domains are significantly more different between prokaryotic and eukaryotic ABC2 transporters relative to the prokaryotic/eukaryotic differences within ABC1 and ABC3. It is not clear why

ABC2a and ABC3 ATPases share the same branch, but we noticed that like ABC2a, ABC3 is mostly prokaryotic. ATPases are homologous between different types of ABC transporters (Wang et al. 2009).

Although there is a discrepancy in the clustering of group ABC2a between **Figures 3** and **4**, overall the transmembrane tree and the ATPase tree have similar topologies. This conclusion is also supported by the observation that 9% of the trees generated for TMDs, when testing different SSEARCH36 parameters to calculate sequence similarities (i.e., k = 1000; z = 3, 13, 23; BL62), separate groups ABC2a and ABC2b in the same way as the ATPase tree in **Figure 4**. It is noteworthy that 100% of the trees for the ATPase domains consistently split groups ABC2a and ABC2b. In summary, both trees consistently split ABC2 into two groups, while ABC1 and ABC3 form distinct clusters, thus providing strong support for the previously published suggestion that the TMD and the ATPase domains have co-evolved with very few exceptions (Kuan et al. 1995).

Given that ATPases among the three ABC types are homologous, we selected 25 sequences from each group (ABC1, ABC2a, ABC2b and ABC3) in **Figure 4** and attempted to generate phylogenies (see **Methods**). Although the phylogeny built with MrBayes did not converge (average standard deviation of split frequencies is 0.26, which is much larger than the recommended convergence thresholds of 0.01), the overall topology agrees with our current classification of the three ABC types (**Figure 5**). The tree is drawn as a cladogram because the length of the branches is unreliable given that MrBayes did not converge. Although the major nodes supporting each individual ABC group have maximal support (posterior probability = 1.0), the support for the node connecting ABC2a-ABC3 node is significantly weaker (0.7). Therefore, the tree was unable to place ABC2a with ABC3 with high confidence.

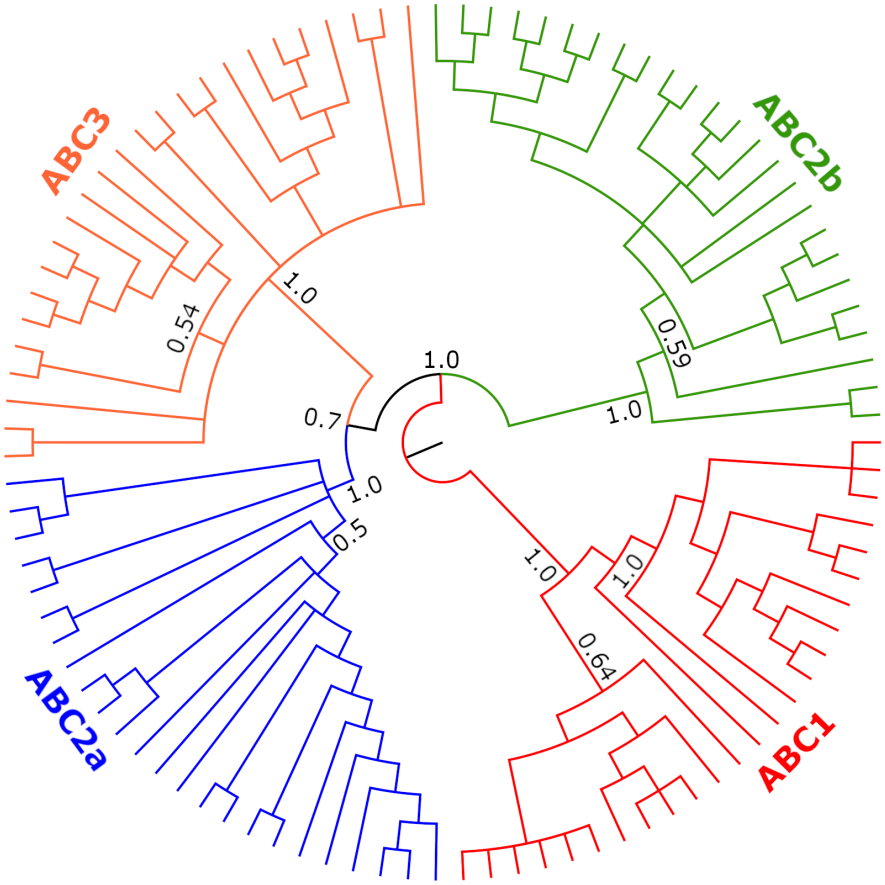

Figure 5. Cladogram representation of the phylogeny generated with MrBayes using 25 sequences per ABC group (ABC1, ABC2a, ABC2b, ABC3). One ABC2a protein, Q8IUA7 (TC# 3.A.1.211.16), was removed from the analysis as it did not cluster with any of the ABC groups. For the sake of clarity, posterior probabilities are shown only for major nodes. Note that the node connecting ABC3 and ABC2a is not strongly supported (0.7), but the nodes supporting the integrity of each individual ABC group have maximal support (1.0). Because MrBayes did not converge, only the overall topology is meaningful (see **Methods**).

Repeat unit analysis

Sequence-based approach

Proteins in the three ABC types were scanned for repeat units as described in **Methods**. For the ABC3 type, we found significant scores between TMSs 1-4 and TMSs 5-8 within the same protein (**Figure 6**). We were also able to identify four TMS repeat for ABC3 with a significant E-

value of 2.5×10^{-22} (TC# 3.A.1.207.1) using HHRepID. This is consistent with our previous observation that ABC3 has a 4+4 topology (Wang et al. 2009). The similarity between the two repeat units suggests that the duplication is recent enough to be detected in any one sequence. We regard as less likely the possibility that the duplication is ancient, but selective pressures could have acted to prevent the two repeat units from diverging significantly. It is worth noting that many 4 TMS membrane proteins of the ABC3 superfamily form homo- or hetero-dimers in the complete export system. Both ABC1 and ABC2 systems also form dimers which sometimes are fused, forming proteins with 12 TMSs. Thus, the 4 TMS repeat unit in ABC3 systems appears to be the functional equvalent of the 6 TMS unis in ABC1 and ABC2 systems.

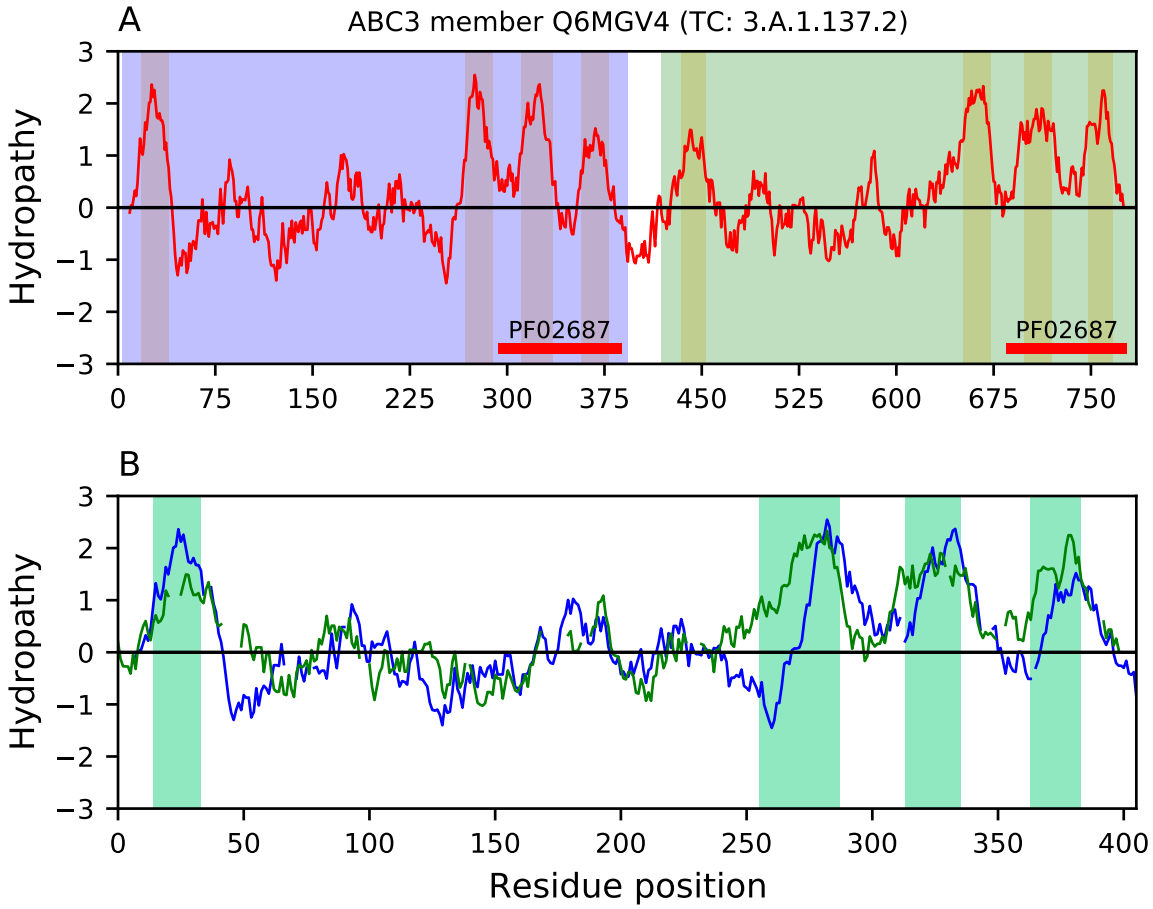

Figure 6. Hydropathy alignment between the first and the second halves of ABC3 member Q6MGV4 (TC# 3.A.1.137.2). The alignment has a GSAT score of 18 and E-value of 1.9×10^{-10}.

Figure 7 shows the results of the repeat unit analysis for ABC1 proteins. We observed a predominance of 2+2+2 TMS topologies, with no single alignment supporting the 3+3 topology when using AncientRep, cut either after the second TMS or after the third TMS (**Figures 7A** and **7C**). The average GSAT score for cases supporting 2+2+2 is below 9, which may be considered marginally significant given the short length of the alignments (**Figures 7B** and **7D**).

For ABC2 proteins, when the AncientRep program cut was after the second TMS, there were as many alignments supporting a 3+3 topology as supporting a 2+2+2 topology (**Figure 8A**).

However, when the AncientRep program cut was after the third TMS, the number of alignments supporting the 3+3 topology was in great excess of those supporting a 2+2+2 topology (**Figure 8C**). Although the average GSAT scores supporting the 3+3 topology were consistently higher than the average GSAT scores supporting the 2+2+2 topology, the difference between the scores is again marginal (**Figure 8B** and **8D**). If the TMS topology for ABC1 is 2+2+2 and the TMS topology for ABC2 is 3+3, these results indicate that fusions of the repeat units to form 2x larger proteins were ancient events, an observation supported by the fact that no recognizable homologous 2 (for ABC1) or 3 (for ABC2) TMS proteins have ever been found.

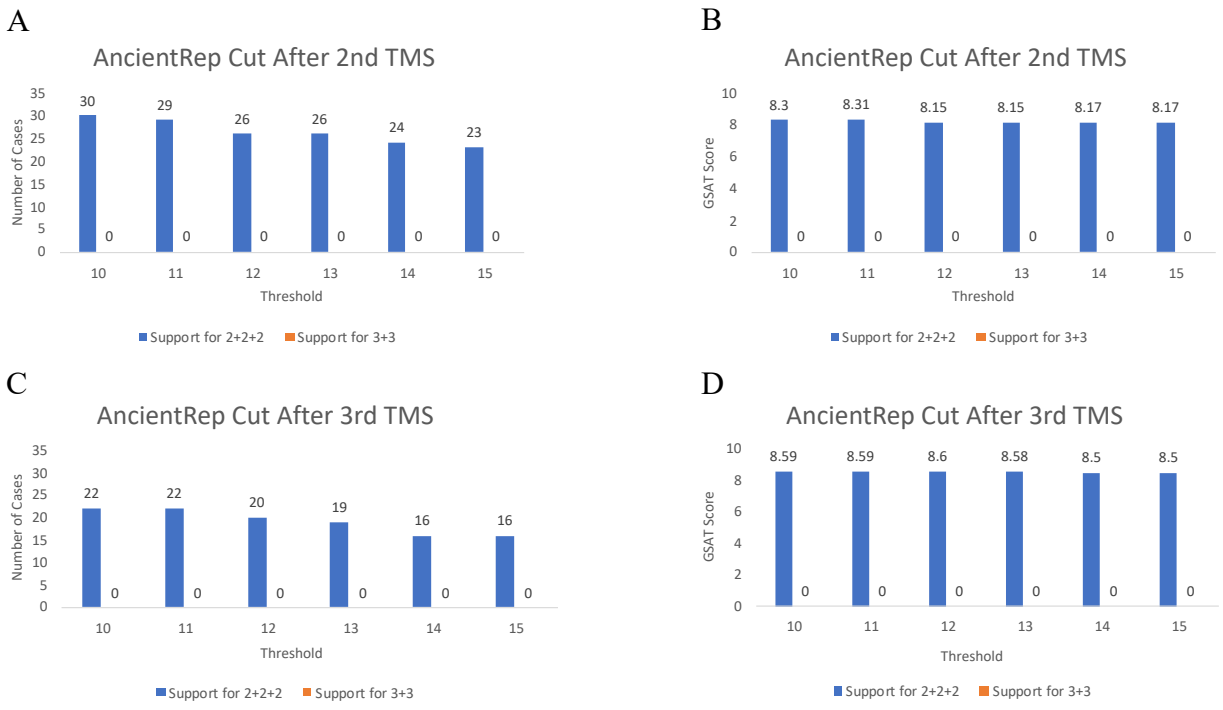

Figure 7. Repeat unit analysis for ABC1 proteins. A. Number of alignments supporting 2+2+2 (blue bars) and 3+3 (orange bars) topologies when the first 2 TMSs are compared to the rest of the transmembrane domain (AncientRep cut after the second TMS). B. Average GSAT scores of the alignments in A. C. Number of alignments supporting 2+2+2 and 3+3 topologies when the first 3 TMSs are compared to the rest of the transmembrane domains (AncientRep cut after the third TMS). D. Average GSAT scores of the alignments in C. Notice that there are no alignments supporting the 3+3 topology.

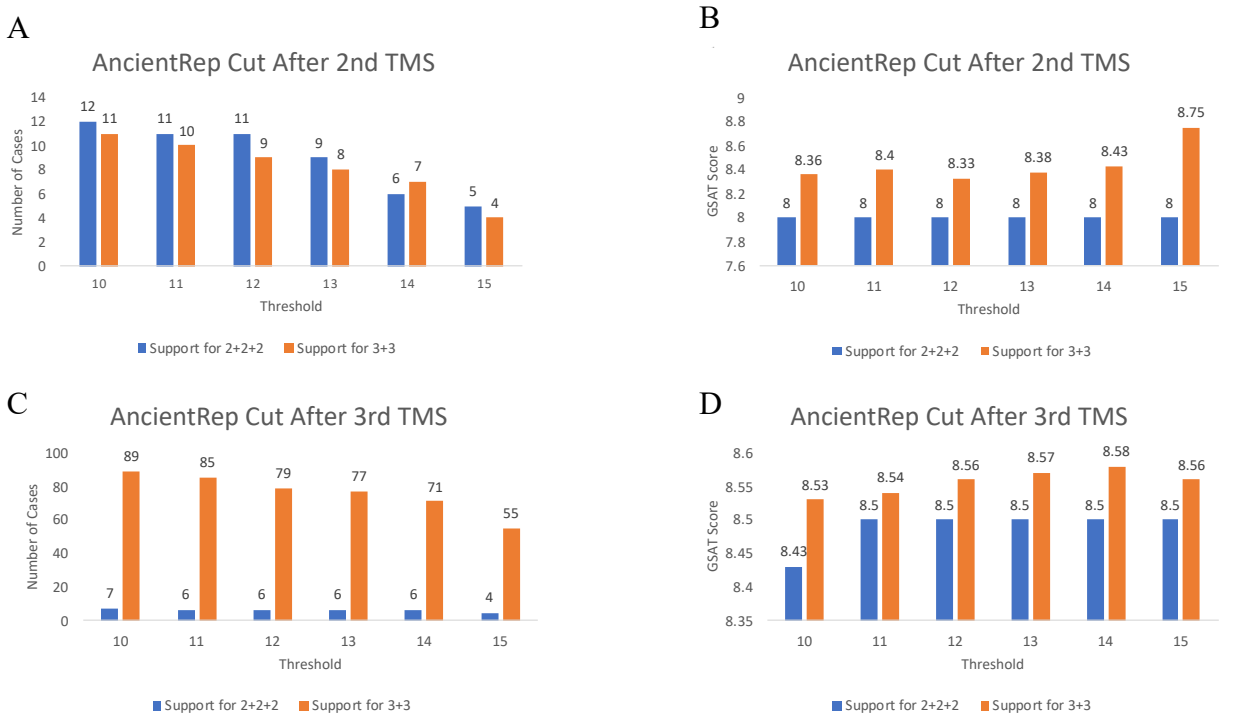

Figure 8. Repeat unit analysis for ABC2 proteins. The format is the same as in Figure 7. A. Number of alignments supporting 2+2+2 and 3+3 topologies when the first 2 TMSs are compared to the rest of the transmembrane domain. B. Average GSAT scores corresponding to the alignments in panel A. C. Number of alignments supporting 2+2+2 and 3+3 topologies when the first 3 TMSs are compared to the rest of the sequences. D. Average GSAT scores of the alignments in panel C. In panel A, the similar numbers of alignments supporting the two potential topologies contrasts with the predominance of alignments supporting topology 3+3 in panel C. The GSAT scores in panels B and D are similar, and alignments supporting topology 3+3 had consistently higher values (see text for discussion).

In order to determine the significance of the scores supporting the two possible topologies, a negative control of randomized sequences was generated as described in **Methods**. The purpose was to compare alignments between real biological sequences with alignments between shuffled membrane proteins that preserve the same amino acid composition and TMS topology of the sequences. The purpose is thus to test the posit that alignments between homologous membrane proteins will have better scores than alignments between shuffled sequences. **Figure 9** shows the

alignment between a real ABC1 protein and its shuffled version. The high resemblance of the TMS topologies in both proteins and their poor alignment scores confirm the efficacy of the shuffling strategy (see **Methods**). This indicates that two unrelated transporters with sequences showing similar spacing of TMSs can produce good-looking, but misleading, hydropathy alignments. In addition, given that the alignment of hydrophobic residues will produce acceptable scores according to standard substitution matrices (e.g. BLOSUM, PAM, etc.), the significance of the overall alignment may artificially improve beyond thresholds of significance between unrelated integral membrane proteins (Wong et al. 2010; Wong et al. 2011). The higher the density of TMSs in the sequences to be compared and the shorter the loops connecting the TMSs, the better the score of the alignment can be expected as an artifact of the higher number of hydrophobic residues in similar positions within the alignment. This artificial effect is magnified in highly hydrophobic TMSs rich in aliphatic residues, also referred to as simple TMSs according to the TMSOC classification (Wong et al. 2012).

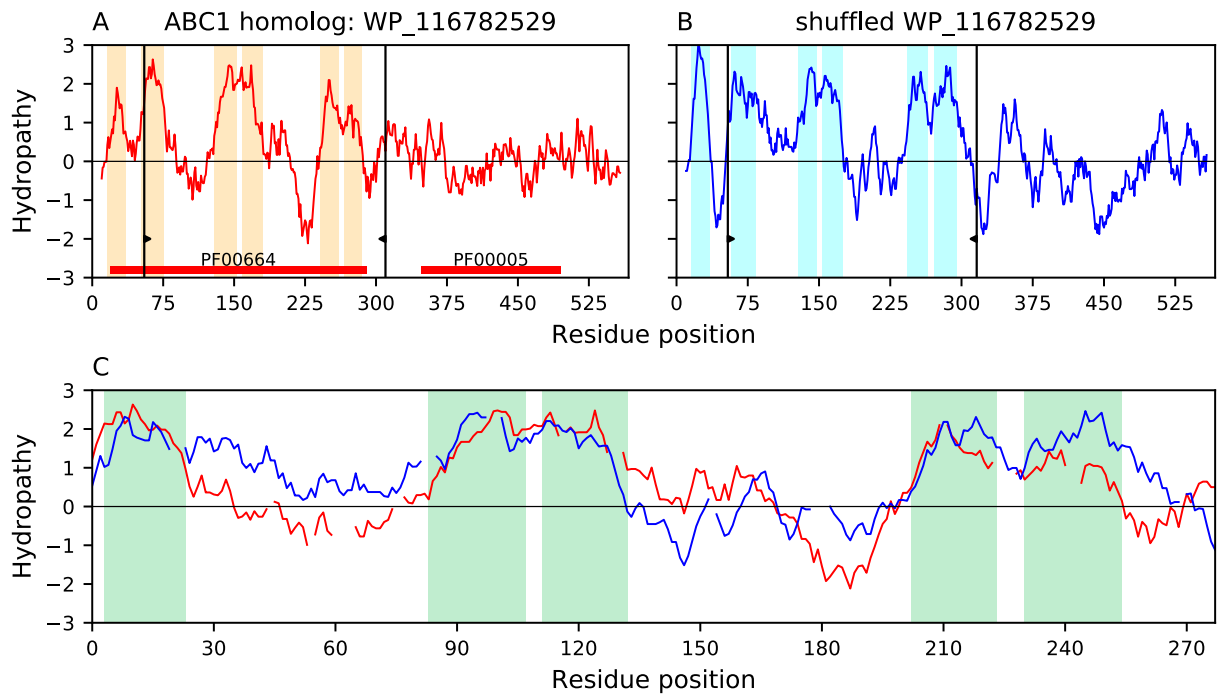

Figure 9. Hydropathy alignment between the real transmembrane domain of the ABC1 homolog (WP_116782529; A) and its shuffled version (B). The two sequences have the same TMS topologies and a seemingly good hydropathy alignment (C), but an insignificant E-value of 0.18.

Table 3 shows the top alignment scores obtained for ABC1 and ABC2 proteins when searching for their repeat units. For purposes of comparison, the table also shows the score produced by aligning randomized versions of real ABC1 and ABC2 proteins. We observed similar GSAT scores and E-values supporting internal repeat units for the real and randomized proteins. This indicates that the alignment scores of real sequences are not high enough to reliably discriminate them from randomized sequences. Note that the E-values obtained would be regarded as acceptable if we were aligning globular proteins. As expected from **Figure 9**, hydropathy alignments involving randomized sequences look as good as alignments involving real sequences. The scores observed in ABC1 supporting 2+2+2 and in ABC2 supporting 3+3 and 2+2+2 can be explained by 1) the similar short spacing between the TMSs, and 2) as discussed above, by

artificially inflated scores in short high-coverage alignments because equally spaced TMSs with short loops contain a high density of hydrophobic residues. For ABC1 proteins, when the cut is after the third TMS, the longer hydrophilic loops after the second and fourth TMSs make a 3-TMS alignment unlikely. For ABC2 proteins, when the cut is made after the second TMS, the first two TMSs can align with TMSs 4-5 or TMSs 5-6 because they have similar spacing (**Figure 1)**. Therefore, we observed similar numbers of cases, thus lacking support for either one of the two possible topologies (**Figure 8A**). Due to similar top alignment scores between real ABC proteins and the negative control, we could not identify a repeat unit in ABC1 and ABC2 with high confidence based on sequence analysis alone. Analyses using HHRepID did not identify repeat units in ABC1 or ABC2-type membrane proteins. However, given that the lengths of the loops connecting TMSs are well conserved within each ABC type, they must play a role in shaping the topology of the repeat units.

Table 3. Top alignment scores of repeat unit analyses for the ABC1 and ABC2 families. The first two TMSs of ABC1 proteins and the first three TMSs of ABC2 proteins, as well as their shuffled versions, were compared to the rest of their corresponding transmembrane domains. The GSAT scores and E-values were calculated as described in **Methods**.

	Protein 1	Protein 1 aligned TMS	Protein 2	Protein2 aligned TMS	GSAT score	E-value
ABC1	WP_108667024	1-2	WP_091821458	5-6	10	3.70×10^{-06}
ABC1 shuffled protein	Shuffled_seq_1	1-2	Shuffled_seq_2	3-4	10	3.50×10^{-07}
ABC2	WP_031465087	1-3	KQC10936	4-6	10	7.40×10^{-06}
ABC2 shuffled protein	Shuffled_seq_3	1-3	Shuffled_seq_4	4-6	13	5.30×10^{-08}

3D Structure-based approach

As expected, within-type structural superpositions have better scores (lower RMSD, higher TM-Score and higher coverage) than cross-type comparisons (**Table 4**). Within-ABC2 comparisons, at both sequence and structural levels, yielded the poorest scores, indicating that this type has more divergent structures. Under the assumption that ABC types evolve at comparable rates, we hypothesize that ABC2 is the most ancient type because it is the most divergent. ABC1 vs. ABC2 and ABC1 vs. ABC3 comparisons have marginally significant RMSDs and TM-scores. However, they involve low alignment coverages of 77.37% and 75.81%, respectively, which means that the aligned regions only contain 3 TMSs and are likely not significant (see the benchmark with a negative control below). ABC2 vs. ABC3 also has marginally significant RMSD and TM-scores but it involves a much higher coverage of 97.08%, indicating that the alignment covers almost the full 4HB. This emphasizes the structural similarity between the two types and suggests the possibility of homology. **Figure 10** presents the top structural superpositions of ABC2 vs. ABC2 and ABC2 vs. ABC3. At the sequence level, an ABC2-ABC3 relationship was identified, but the signal was weak (see Section "Protein tree of ABC families"). **Figure 11** presents the hydropathy curve of the alignment corresponding to the 4HBs involved in the structural superposition shown in **Figure 10B**. Given that the sequence alignment score is poor (E-value: 0.29), and the structural superposition has marginal significance, it is not possible to conclude or discard the possibility of limited homology for regions of the ABC2 and ABC3 proteins.

Table 4. Top structural alignment scores of 4HBs within and between ABC types. RMSD is the average root-mean-square distance between corresponding atoms after an optimal rotation of one structure relative to the other. The coverage relative to the size of the bundles is necessary to assess the significance of the RMSD scores. The TM-score, on the other hand, weights residue pairs at smaller distances relatively more strongly than those at larger distances. RMSD < 4 Å, TM-Score > 0.55, and coverage > 75% indicate an acceptable structural alignment.

	ABC1 VS ABC1	**ABC2 VS ABC2**	**ABC3 VS ABC3**	**ABC1 VS ABC2**	**ABC1 VS ABC3**	**ABC2 VS ABC3**
RMSD (Å)	1.51	2.25	0.17	2.86	3.77	3.53
TM-Score	0.91766	0.80376	0.99931	0.58373	0.64528	0.68085
Coverage (%)	100	96.32	100	77.37	75.81	97.08

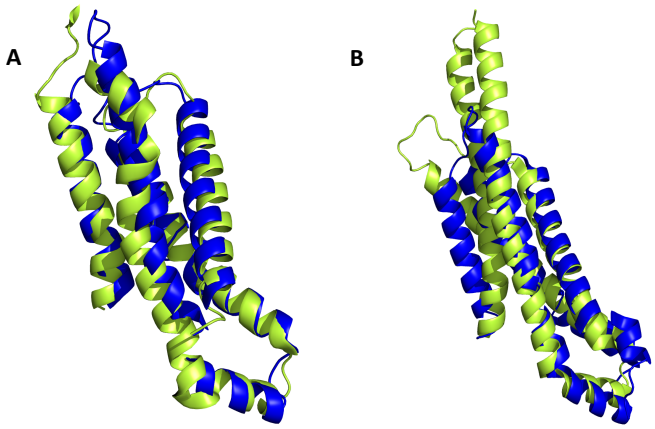

Figure 10. Top 4HB bundle alignments. A. ABC2 vs ABC2 (6HBU_A, helices 1-4 vs 5XJY_A, helices 1-4; RMSD: 2.25 Å; coverage: 96.32% and TM-Score: 0.8037). B. ABC2 vs ABC3 (6HCO_A, helices 1-4 vs 5WS4_B, helices 1-4; RMSD: 3.53 Å; coverage: 97.08% and TM-Score: 0.68085).

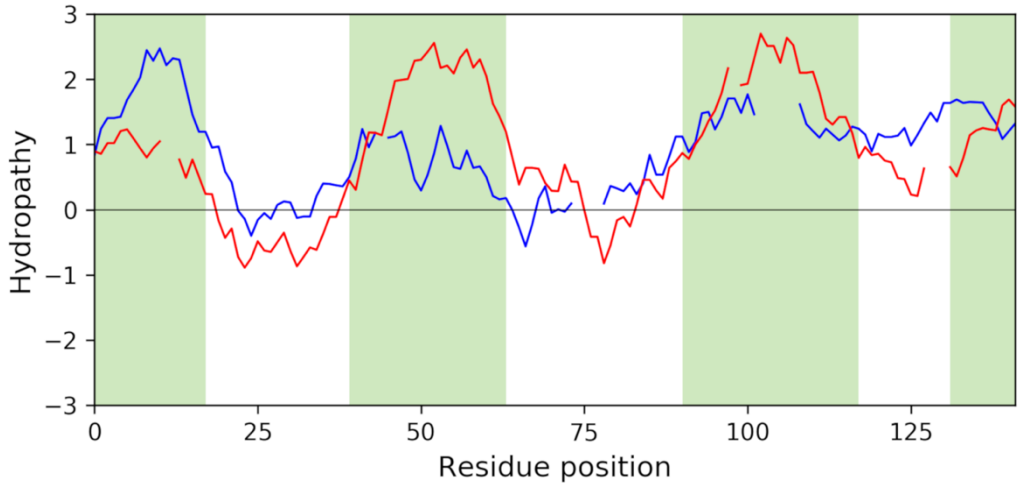

Figure 11. Hydropathy alignment of the top 4HB structural alignment of ABC2 vs ABC3 shown in Figure 9B. A large hydrophilic loop was removed from the ABC3 protein (blue curve) between TMSs 1 and 2. The poor quality of the alignment (E-value: 0.29) and the marginal significance of the structural alignment (RMSD: 3.53 Å; coverage: 97.08% and TM-Score: 0.68085) are not enough to conclude homology between ABC2 and ABC3.

For the structural internal repeat unit of ABC1, we could not find an alignment between helices 1-4 and helices 3-6, supporting the 2+2+2 topology. Surprisingly, we found a 3-helix alignment between helices 1-3 and helices 4-6 supporting a 3+3 TMS topology (**Figure 12**). To assess the relevance of this finding, we aligned ABC structures against the structures of a negative control set comprised of the following unrelated families: The Mouse Virulence Factor (MVF) Family (TC# 2.A.66.4), The Presenilin ER Ca^{2+} Leak Channel (Presenilin) Family (TC# 1.A.54), The Gap Junction-forming Innexin (Innexin) Family (TC# 1.A.25), The (Largely Archaeal Putative) Hydrophobe/Amphiphile Efflux-3 (HAE3) Family (TC# 2.A.6.7), The Gap Junction-forming Connexin (Connexin) Family (TC# 1.A.24), The Transmembrane Channel (TMC) Family (TC# 1.A.17.4), The Polycystin Cation Channel (PCC) Family (TC# 1.A.5), The Major Intrinsic Protein (MIP) Family (TC# 1.A.8), The Transient Receptor Potential Ca^{2+} Channel (TRP-CC) Family (TC# 1.A.4), The (Gram-positive Bacterial Putative) Hydrophobe/Amphiphile Efflux-2

(HAE2) Family (TC# 2.A.6.5), The Mg2+ Transporter-E (MgtE) Family (TC# 1.A.26), The Sugar Porter (SP) Family (TC# 2.A.1.1), The $gp91^{phox}$ Phagocyte NADPH Oxidase-associated Cytochrome b_{558} (Phox) Family (TC# 5.B.1), The Neurotransmitter:Sodium Symporter (NSS) Family (TC# 2.A.22), The Voltage-gated Ion Channel (VIC) Superfamily (TC# 1.A.1), The Melittin (Melittin) Family (TC# 1.C.18), and The Outer Membrane Beta-barrel Endo-protease, Omptin (Omptin) Family (TC# 9.B.50).

We found that good structural alignments of 3 α-helix bundles (3HBs) can frequently be observed between unrelated families. For example, the top 3HB alignment between ABC1 system 3.A.1.208.4 (PDB: 5YWA_H; helices 3-5) and gap junction system 1.A.25.3.1 (PDB: 6G9O_A; helices 2-4) has RMSD: 2.24 Å and TM-score: 0.77885, but there was no significant sequence alignment between the structurally aligned segment or the entire proteins. Therefore, we cannot claim that the structural alignment in **Figure 12** provides a reliable prediction of topology. In a similar way, we were unable to find evidence of any structural repeat unit for ABC2. Unfortunately, there are no available ABC3 structures with 8 or 10 TMSs (all have 4 TMSs), and we were therefore unable to perform structural repeat unit analysis.

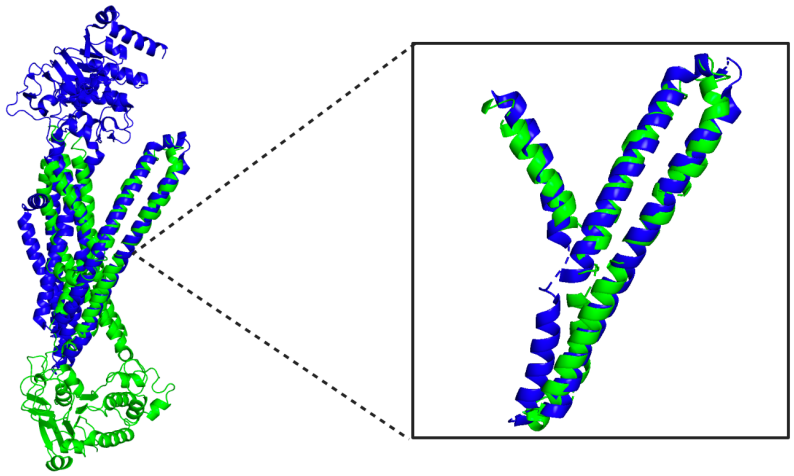

Figure 12. Structural alignment between two ABC1 proteins: 3B60 (α-helices 1-3) vs 5MKK (α-helices 4-6). Here, TMS 1 aligns with TMS 4, TMS 2 aligns with TMS5 and TMS 3 aligns with TMS 6. The alignment has an RMSD of 3.15 Å, coverage of 85%, and a TM-Score of 0.64792.

Clustering analysis of structural similarities

We used the ratio Coverage/RMSD to quantify the level of similarity between 4HBs and performed hierarchical clustering analysis as described in **Methods.** The resulting tree is shown in **Figure 13**. Consistent with the clustering tree of the transmembranal domain sequences in **Figure 3**, ABC1 forms a monophyletic cluster, while ABC2 and ABC3 share a branch, and ABC2 is separated internally into groups ABC2a and ABC2b. This supports the relationship between ABC2 and ABC3 suggested by sequence-based analyses.

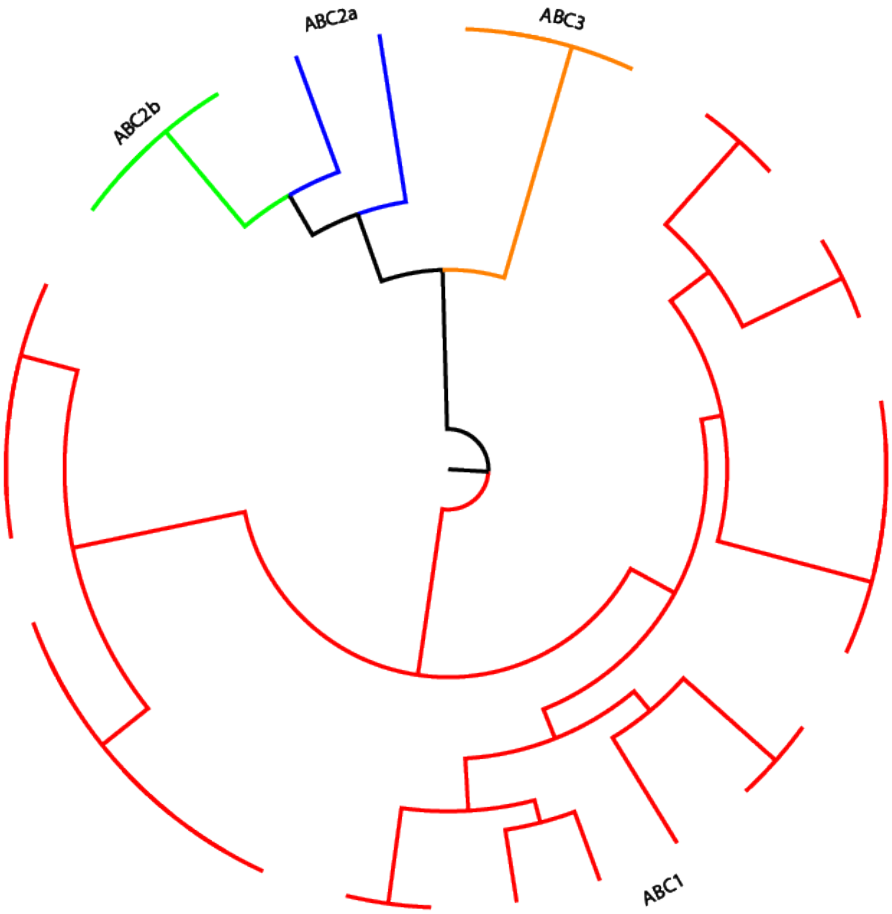

Figure 13. Hierarchical clustering of 3D structural similarities of transmembrane domains across ABC types. See **Methods** for details on how the tree was generated. No scale bar is provided because only the topology of the tree is meaningful. For simplicity, the labels of individual leaves were omitted. Notice the split between ABC2a and ABC2b with highly similar overall topology to the sequence-based tree in Figure 3.

Figure 14 shows the clustering tree of ATPase structural similarities (see **Methods**). In this tree, there is no clear separation of the three ABC types compared to the other trees, ABC1 is not a monophyletic group, and the upper branch has a mixture of all three types. Notwithstanding, most ABC1 structures are clustered together and all ABC2 and ABC3 structures share a major branch: a tendency that agrees with the other trees. The anomalies in **Figure 14** can be explained, at least in part, by the following considerations: 1) ATP hydrolyzing domains have highly

conserved structures (ter Beek et al. 2014) and can generate confounding Coverage/RMSD ratios. 2) An artifact of the tree building algorithm could be responsible; that is, although sequence-wise, the misgrouped proteins are indeed most similar to their respective ABC types, their actual Coverage/RMSD ratios are beyond the values necessary to minimize the variance of their correct ABC type by the Ward method. 3) A lack of representative ABC2 and/or ABC3 structures may have prevented a balanced representation of each type in the analysis.

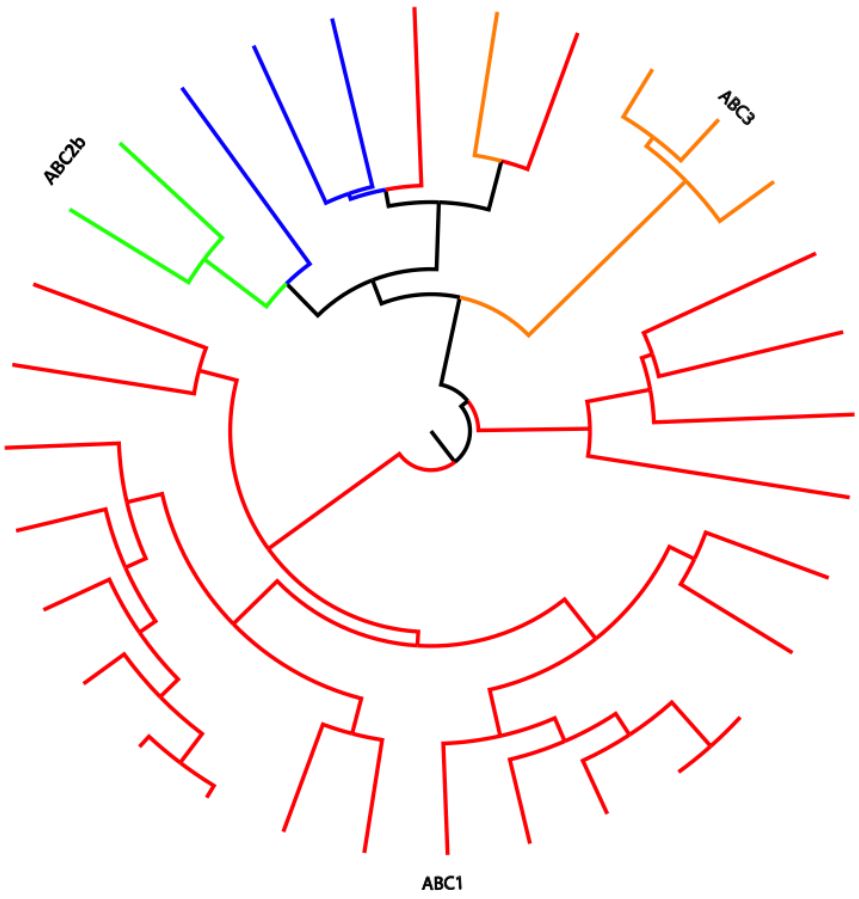

Figure 14. Hierarchical clustering of 3D structural similarities of the ATPase domains across ABC types. See **Methods** for details on how the tree was generated. No scale bar is provided because only the topology of the tree is meaningful. For simplicity the labels of individual leaves were omitted. Notice how some ABC1 and ABC3 structures are clustered with ABC2 structures (see text for discussion).

Conclusions

An important goal of this thesis was to test the robustness of the current classification of ABC exporters into three types in light of the substantially increased amount of sequence and 3D structural data available in public repositories since our initial study was conducted (Wang et al. 2009). The protein trees in **Figures 3** and **4**, the distribution of Pfam domains, and the structural similarities in **Figure 14** all support the three-type classification of major types of ABC exporters. Trees built based on the sequences and 3D structures of transmembrane and ATPase domains present ABC1 as a monophyletic group, while ABC2 and ABC3 share a major branch. In 9% of the sequence-based trees generated for the TMDs, as well as all trees for the ATPase domains, we observed a split of ABC2 proteins by the ABC3 cluster that correlates with the prokaryote/eukaryote taxonomic distribution of ABC2 proteins. That is, proteins in group ABC2a are mostly prokaryotic while proteins in group ABC2b are entirely eukaryotic. Interestingly, both ABC1 and ABC3 include prokaryotic and eukaryotic systems, but despite their very considerable sequence divergence, they still consistently group properly within their respective type. Because ABC2 proteins are the most divergent with respect to sequence, structure and organismal distribution, we propose that this is the most ancient type. In addition, although the evidence is not conclusive, the consistent sharing of a major branch between ABC2 and ABC3, at both the sequence and structural levels, suggests a possible evolutionary relationship between these two types. Although we are reluctant to try to delineate what this relationship might be, one possibility is that ABC3 TMSs 1-3 derived from ABC2 TMSs 1-3.

Regarding the repeat units for the three ABC types, we were able to find significant 4-TMS alignments within single proteins for ABC3 using both HHrepID and ancientRep, which is consistent with previous observations that ABC3 has a 4-TMS repeat unit (Wang et al. 2009).

However, we found evidence, though not as strong, supporting the proposed 2-TMS and 3-TMS repeat units for ABC1 and ABC2, respectively, using our methods. Although we did identify sequence alignments that supported the proposed topologies for ABC1(2+2+2) and ABC2 (3+3), the GSAT scores and E-values of the alignments were not sufficiently different from scores obtained with a negative control of randomly shuffled proteins that preserve the amino acid composition and TMS topologies of native ABC proteins. In addition, we were unable to find reliable evidence of structural repeat units, partly due to the small size of the likely repeat units (2 or 3 TMSs). In structural superpositions with the negative control, we observed that two- or three-helix bundles can frequently be aligned by chance and yield scores beyond cutoffs of significance. However, we cannot conclude that there are no repeat units in ABC1 and ABC2, based on the lack of high-quality alignments. The repeat units of ABC1 and ABC2 may have diverged to the point that the signal of the repeat is so weak that current methods cannot reliably discriminate it from alignments with unrelated sequences that have similar TMS topologies. However, it is possible that future genomic sequencing projects will allow this issue to be settled by revealing ABC1 and ABC2 sequences with stronger signals supporting their internal repeat units. In the meantime, we will continue to work on the design and development of more sensitive software tools to detect highly divergent repeat units.

Based on the energy-coupling ATPase proteins, the ABC superfamily is one of the largest superfamilies found in nature. However, it should be noted that this is not necessarily true for the membrane proteins, which on the basis of both sequence and 3D structural data are most likely polyphyletic.

During the stage of data consistency verification, we identified and removed from our analyses three families with characteristics highly divergent from those of other ABC families.

First, members of the putative ABC1-type family, U-GlcE (TC# 3.A.1.139), have typical ABC1-type ATPases, but, the TMDs are very different from other ABC1 TMDs, as observed when blasting them against TCDB. Additionally, they have different Pfam domain and clan designations compared to other ABC1 TMDs. Second, the characteristic TMD Pfam domain designation (PF06472) of putative ABC1 family, P-FAT (TC# 3.A.1.203), differs from other ABC1 TMDs (PF00664) although they share the same clan (CL0241). Additionally, TC Blast searches of the TMDs of family P-FAT do not retrieve other ABC1 proteins, and the Pfam domains cannot be projected from other ABC1 families, casting doubt on the membership of family P-FAT to clan CL0241. Third, members of the putative ABC2 family, LptBFG (TC# 3.A.1.152), clearly have a 3+3 TMS topology, but these proteins have a large hydrophilic loop between the two 3-TMS hydrophobic (putative) repeat units, a unique feature of this family. Their TMDs match a different Pfam domain (PF03739) and belong to a different clan (CL0404) compared to the rest of the ABC2 TMDs. TCBlast searches of the TMDs of LptBFG family members do not retrieve other ABC2 proteins, and the corresponding Pfam domains cannot be projected. Finally, TMDs of the LptBFG family also show a different 3D structure. It is possible that these divergent families represent additional ABC types, especially in the case of the LptBFG family. These observations suggest that future evolutionally analyses of these three families will prove to be most interesting.

The ABC functional superfamily is involved in medically relevant functions such as the influx/efflux of drugs, toxins and macromoleculres, and in maintaining cellular homeostasis (Vasiliou et al. 2009). Therefore, increasing our understanding of the structural and evolutionary relationships among ABC transporter types has important implications with respect to targeting ABC transporters for drug discovery and experimental protocol design. We are confident that the work reported here will benefit the many researchers studying these important transport systems.

Bibliography

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.

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

Biegert A, Soding J. 2008. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* **24**: 807-814.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**: 164-166.

Glavinas H, Krajcsi P, Cserepes J, Sarkadi B. 2004. The role of ABC transporters in drug resistance, metabolism and toxicity. *Curr Drug Deliv* **1**: 27-42.

Igarashi Y, Aoki KF, Mamitsuka H, Kuma K, Kanehisa M. 2004. The evolutionary repertoires of the eukaryotic-type ABC transporters in terms of the phylogeny of ATP-binding domains in eukaryotes and prokaryotes. *Mol Biol Evol* **21**: 2149-2160.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**: 195-202.

Jones PM, George AM. 2004. The ABC transporter structure and mechanism: perspectives on recent research. *Cell Mol Life Sci* **61**: 682-699.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.

Krissinel E, Henrick K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* **60**: 2256-2268.

Kuan G, Dassa E, Saurin W, Hofnung M, Saier MH, Jr. 1995. Phylogenetic analyses of the ATP-binding constituents of bacterial extracytoplasmic receptor-dependent ABC-type nutrient uptake permeases. *Res Microbiol* **146**: 271-278.

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. 2012. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* **40**: D370-376.

Medrano-Soto A, Ghazi F, Hendargo KJ, Moreno-Hagelsieb G, Myers S, Saier MH, Jr. 2020. Expansion of the Transporter-Opsin-G protein-coupled receptor superfamily with five new protein families. *PLoS One* **15**: e0231085.

Medrano-Soto A, Moreno-Hagelsieb G, McLaughlin D, Ye ZS, Hendargo KJ, Saier MH, Jr. 2018. Bioinformatic characterization of the Anoctamin Superfamily of Ca2+-activated ion channels and lipid scramblases. *PLoS One* **13**: e0192851.

Pearson WR. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**: 635-650.

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

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572-1574.

ter Beek J, Guskov A, Slotboom DJ. 2014. Structural diversity of ABC transporters. *J Gen Physiol* **143**: 419-435.

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

Vasiliou V, Vasiliou K, Nebert DW. 2009. Human ATP-binding cassette (ABC) transporter family. *Hum Genomics* **3**: 281-290.

Wang B, Dukarevich M, Sun EI, Yen MR, Saier MH, Jr. 2009. Membrane porters of ATP-binding cassette transport systems are polyphyletic. *J Membr Biol* **231**: 1-10.

Wang SC, Davejan P, Hendargo KJ, Javadi-Razaz I, Chou A, Yee DC, Ghazi F, Lam KJK, Conn AM, Madrigal A, Medrano-Soto A, Saier MH. 2020. Expansion of the Major Facilitator Superfamily (MFS) to include novel transporters as well as transmembrane-acting enzymes. *Biochim Biophys Acta Biomembr* **1862**: 183277.

Wong WC, Maurer-Stroh S, Eisenhaber F. 2010. More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput Biol* **6**: e1000867.

Wong WC, Maurer-Stroh S, Eisenhaber F. 2011. Not all transmembrane helices are born equal: Towards the extension of the sequence homology concept to membrane proteins. *Biol Direct* **6**: 57.

Wong WC, Maurer-Stroh S, Schneider G, Eisenhaber F. 2012. Transmembrane helix: simple or complex. *Nucleic Acids Res* **40**: W370-375.

Xiong J, Feng J, Yuan D, Zhou J, Miao W. 2015. Tracing the structural evolution of eukaryotic ATP binding cassette transporter superfamily. *Sci Rep* **5**: 16724.

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.

Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**: 2302-2309.