Systematic *in silico* discovery of novel solute carrier-like proteins from proteomes

Gergely Gyimesi* and Matthias A. Hediger*

Membrane Transport Discovery Lab, Department of Nephrology and Hypertension and Department for BioMedical Research, Inselspital, University of Bern, Freiburgstrasse 15, CH-3010 Bern, Switzerland.

*Corresponding authors.
E-mail: gergely.gyimesi@dbmr.unibe.ch (GG), matthias.hediger@ibmm.unibe.ch (MH).

Abstract

Solute carrier (SLC) proteins represent the largest superfamily of transmembrane transporters. While many of them play key biological roles, their systematic analysis has been hampered by their functional and structural heterogeneity. Based on available nomenclature systems, we hypothesized that many as yet unidentified SLC transporters exist in the human genome, which await further systematic analysis. Here, we present criteria for defining "SLC-likeness" to curate a set of "SLC-like" protein families from the Transporter Classification Database (TCDB) and Protein families (Pfam) databases. Computational sequence similarity searches surprisingly identified ~130 more proteins in human with potential SLC-like properties compared to previous annotations. Interestingly, several of these have documented transport activity in the scientific literature. To complete the overview of the "SLC-ome", we present an algorithm to classify SLC-like proteins into protein families, investigating their known functions and evolutionary relationships to similar proteins from 6 other clinically relevant experimental organisms, and pinpoint structural orphans. We envision that our work will serve as a stepping stone for future studies of the biological function and the identification of the natural substrates of the many under-explored SLC transporters, as well as for the development of new therapeutic applications, including strategies for personalized medicine and drug delivery.

Introduction

Membrane transporters and channels are the main entry routes for nutrients, ions, xenobiotics and serve as major exit routes for waste products and metabolites. The solute carrier (SLC) protein superfamily accounts

for over 50% of all transport-related proteins and about 10% of all membrane proteins encoded by the human genome. With more than 400 annotated members, it is the largest superfamily of membrane transporter proteins [1,2]. The roles of SLC transporters as cellular gatekeepers, determinants of nutrient homeostasis and facilitators of drug metabolism and drug targeting has recently been revisited [3]. Around ~50% of currently annotated SLCs are predicted to be associated with human disease phenotypes and many SLCs are considered to represent promising drug targets or drug delivery systems or to affect drug ADMET (absorption, distribution, metabolism, extrusion, toxicity). It has recently become evident that the SLC superfamily offers an enormous unexplored therapeutic treasure. But while the list of approved drugs that target transporter proteins is increasing, many promising SLCs still remain unexplored, uncharacterized and underrepresented in the literature.

The SLC nomenclature system has traditionally been used to classify mammalian secondary active and facilitative transporters, including exchangers and antiporters, into families based on sequence identity per enquiry by the Human Gene Nomenclature Committee (HGNC) starting in the early 1990s [1]. Originally, SLCs assignments have been made for all membrane transport proteins that are not channels, ATP-driven pumps, aquaporins, porins of the outer mitochondrial membrane or ATP-binding cassette (ABC) transporters, while usually having multiple transmembrane spanning segments and exhibiting transmembrane transport of a solute, or showing homology to membrane proteins having such features. Due to this construction, the SLC superfamily is structurally and functionally highly heterogeneous and thus most likely evolutionarily polyphyletic in origin. Because of these properties and the lack of common sequence patterns in the different SLC transporters, it has been difficult to assess how many SLC transporters exist in the human genome or which proteins could be predicted as SLC transporters. Proteins were typically added to the SLC system on a case-by-case basis. However, in view of recent requests to add new members into the SLC nomenclature, we suspected that the SLC system might be incomplete.

Despite their heterogeneity, SLC transporters seem to share common properties that probably were evolutionarily selected based on their suitability as facilitative transporters, secondary active transporters or exchangers. A remarkable property is that they generally have a symmetric inverted repeat architecture,

which can be observed based on the available structures [4] and can sometimes also be detected at the sequence level [5,6]. Another important aspect is that the currently well-studied SLC transporters seem to follow an alternating access mechanism, meaning that the substrate-binding site is exposed on either one or the other side of the membrane, but not on both sides simultaneously [7,8]. Consequences of these properties are that SLC transporters typically contain many transmembrane helices (TMHs), and functionally exhibit saturable transport activity. In addition, most but not all SLC carriers transport water-soluble small molecules. These properties could be used as criteria to identify additional SLC transporter proteins.

In fact, there have been earlier attempts to gather additional SLCs from the human genome [9–11], as well as to classify them using automatic methods [12,13]. In this regard, one study [9] used BLAST searches to find SLC transporters that have local sequence similarities and found that 15 of the known SLC families fall into 4 phylogenetic clusters, which were termed α, β, γ, and δ groups. In addition, they have found 19 sequences that have previously not been described as SLCs. A later study [12] used a more sensitive HMM-HMM (hidden Markov-model) comparison-based method to identify locally similar regions in known SLC proteins. Visualization of the similarity network revealed visible protein clusters that correlated with existing SLC families. In addition, they identified two unannotated protein sequences that showed similarity to existing SLC proteins. A common limitation of these studies, however, is that they only searched for proteins that were similar to proteins already annotated as human SLC transporters. Nevertheless, these efforts using sequence similarity-based approaches have highlighted that there are additional as yet unannotated SLC transporters in human protein databases.

In our current study, we aimed to identify missing SLC transporters that may differ from those currently annotated in human. To do this, we turned to sequence databases and annotation systems that are phylogenetically broader and not limited to human proteins, and we developed criteria to define “SLC-like” proteins. Our method is thus more complete and general than previous approaches and tackles the task of identifying SLC transporters on a semantic level, i.e. starting from explicit first principles. For this reason, we turned to the Transporter Classification DataBase (TCDB) to enable the extraction of missing “SLC-like” proteins.

The Transporter Classification DataBase (TCDB) is an alternative classification system that was created in the 1990s in parallel to the SLC nomenclature series [14,15]. It collects transport-related membrane proteins, including membrane receptors, transporters, ion channels, and membrane-anchored enzymes from all kingdoms of life, with a particular focus on proteins from lower organisms. The proteins in the TCDB are organized hierarchically into subfamilies, families and superfamilies based on phylogenetic and functional considerations, and each member in the database is given a five-segmented TC# similar to the EC# that is used for enzyme classification. In addition, a brief description is provided for each family that introduces identified members and contains links to the most important relevant papers.

However, the TCDB dataset is not directly applicable for creating an overview of the human SLC-ome (the collection of SLC transporters encoded in the human genome). One of the reasons is that the TCDB is set up as a "representative database", which means that it only contains certain representative sequences from each family. In addition, there is no particular focus on human proteins, and in fact several annotated human SLC transporters are not present in the database.

It also seems problematic to consider all proteins in the TCDB that are annotated as part of the secondary transporter superfamily TC# 2.A as being SLC-like. Indeed, many of the TCDB families annotated as part of TC# 2.A exhibit structural or sequence features that do not match the characteristics of existing SLCs. Examples of this are the Trk K^+ transporters (#2.A.38), which display an ion-channel like structural fold [16], the GUP glycerol uptake proteins (#2.A.50), which exhibit enzymatic activity based on follow-up studies [17], and the Twin Arginine Targeting (Tat) family (#2.A.64), which are actually protein secretion complexes [18,19]. Since none of these families correspond to structural or functional characteristics of currently known SLC transporters, it is likely that not all proteins annotated under the TC# 2.A superfamily are "SLC-like". Thus, while the TCDB could be a rich source of information for finding new transporters, it is clear that the perception of what a "transporter" should be according to TCDB does not always correspond to the typical properties of well-characterized SLC proteins. Additional filtering of TCDB data is therefore necessary.

As part of the TransportDB project, there were parallel efforts to collect secondary transporters from human and several other organisms [20–22]. Within the TransportDB project, the authors have built an automatic transporter annotation pipeline (TransAAP), which relies on BLAST searches, the Clusters of Orthologous Groups (COG) database [23], "selected HMMs for transporter protein families" [22] from the TIGRfam and Pfam databases [24,25] and hydropathy predictions of TMHMM [26]. This pipeline was used as a semiautomatic tool to annotate transporter-like proteins from the NCBI RefSeq database. However, based on the currently available TransportDB website, the resulting protein hits are neither linked to protein annotation databases, such as UniProt [27], nor are their official gene symbols or SLC names displayed. Therefore, no correlation with the existing SLC nomenclature is provided, and it is not trivial to say whether or not an existing SLC protein is included in the database.

We would also like to mention the Protein families (Pfam) database [28], which aims to maintain a curated set of protein families, often represented by functional domains. Notably, Pfam provides curated HMMs for each Pfam family to facilitate sequence similarity searches for the occurrence of those domains. In addition, Pfam groups protein families into higher-order groups called clans. Pfam clans contain evolutionarily related families whose relationships are supported either by sequence similarity, structural similarity or other orthogonal biological evidence [29]. While many known Pfam models correspond to the functional regions of known SLC transporters, Pfam neither attaches special importance to transport-related domains nor to membrane-spanning domains. For this reason, while the Pfam database could be a rich source of information about transporters, extracting Pfam families that encode transporter-like domains is non-trivial.

Therefore, there is a clear need in the field to define the criteria of "SLC-likeness" and to identify and classify all proteins in humans and other species that exhibit "SLC-likeness". Thus, in our work, we have added semantic content in the linguistic sense to the term "SLC-likeness" by defining the essential criteria for it, and we have carried out an exhaustive search for proteins that potentially meet these criteria, both with manual curation of datasets and with automatic sequence similarity-based approaches.

Results

Elaboration of criteria for "SLC-likeness"

Since the TCDB takes a very inclusive approach to collecting membrane transport-related proteins from a broad range of biological organisms, we have selected it as the source database for our endeavors. However, as outlined in the introduction, selecting SLC-like proteins from the TCDB is non-trivial. Therefore, we have introduced a set of criteria based on current knowledge of SLC transporters in order to select SLC-like protein families from the TCDB. We believe that these criteria represent the most important properties of currently known SLC transporters. The criteria used to infer SLC-likeness were as follows.

(1) Structure of the protein should be α-helical, with at least three transmembrane helices (TMHs). In exceptional cases, 2-TMH families were also included, but β-barrel proteins, mostly β-structure proteins, membrane-anchored proteins, cyclic peptides and proteins consisting only of soluble domains (based on predictions or structural data) were excluded.

(2) The size of the transported substrate should fall within the small-molecule range (i.e. oligopeptides might be accepted as substrates but protein secretion systems are excluded). Also excluded are DNA-, RNA- and polysaccharide-transporting systems.

(3) Proteins where transport activity was used as a synonym for trafficking (i.e. protein or vesicle translocation within the cell) but otherwise seemingly having no small-molecule transmembrane transport activity were excluded. On this basis, chaperones and other proteins helping the insertion of nascent proteins into a cellular membrane were also excluded.

(4) Proteins with a channel-like mechanism were excluded, except in some rare cases. In particular, holins, toxins and other pore-forming proteins, and proteins bearing similarity to them, have been excluded.

(5) Proteins with enzymatic activity, or similarity to known enzymes were excluded. In some cases, where the protein was believed to contain both a transport domain and a soluble enzyme domain, the proteins have been included.

(6) Receptors that trigger endocytosis upon substrate binding were excluded. Only receptors were included where the receptor protein itself mediates the translocation of the substrate through the membrane, or the insertion of the substrate in the membrane, if that is its final location.

(7) ATP- and GTP-dependent transporters (e.g. ABC, ECF) were excluded.

(8) For some (mostly putative) transporter families, TCDB does not give an explanation why the proteins would be considered as transporters. Families with no resemblance to known transporters and no indication or argument as to why they would be transporters were excluded.

We must emphasize that for specific proteins or protein families, the verification of some of the criteria for SLC-likeness as described above requires extensive and detailed information about the nature of the transport phenomenon. For example, deciding whether the substrate is a small molecule requires the identification of the transported solute, while ruling out a channel-like mechanism requires extensive information about the transport mechanism itself. However, due to missing information, these criteria could not always be verified while filtering the TCDB for SLC-like protein families. In certain cases, even the identity of the protein that performs the actual membrane translocation within a transporter complex can be unclear. For similar reasons, the historical discrepancy that the naming and classification of a gene/protein generally precludes the detailed analysis of its structure and function has also caused the currently known ("classical") SLC nomenclature to contain transport proteins with a channel-like mechanism (e.g., the SLC41 family), single TMH (SLC27 family) as well as auxiliary proteins to actual transporters (SLC3 family). Some of these classical SLC families have been included in our selection in order to provide consistency with the existing SLC nomenclature despite the fact that they do not meet some of our criteria for SLC-likeness. Nevertheless, we believe that our criteria are broad enough to allow the identification of all putative SLC-like transporters, while also being specific enough to distinguish them from other well-known, non-transport related transmembrane protein families. Our criteria enabled us to initiate, for the first time, an attempt to set up semantic guidelines defining SLC-likeness, in the sense that explicit criteria are established to disambiguate "transmembrane solute transport" and distinguish from other related membrane protein

activities such as channel-like transport, receptor-mediated endocytosis or protein secretion. In our current study, we designate proteins and protein families that potentially meet the above criteria as "SLC-like".

Search for novel "SLC-like" proteins

The above-mentioned criteria were applied to manually select protein families in the TCDB that either fulfill or can potentially fulfill the criteria for SLC-likeness based on the description of each third-level family from the TCDB database. Throughout this manuscript, we use the term "TCDB family" to refer to third-level groupings in the classification hierarchy (i.e. TC# x.y.z), while "subfamily" and "superfamily" refer to fourth-level (TC# x.y.z.w) and second-level (TC# x.y) classes, respectively. In the work presented herein, we have analyzed superfamilies TC# 1.A, 2.A, 9.A and 9.B and the families, and in certain cases the subfamilies within them (see Table 1).

Table 1. Families and subfamilies examined from the TCDB for SLC-like proteins. Superfamilies marked in boldface have been analyzed. The superfamily #1.A was examined since several existing SLC families are classified here, while superfamily #2.A was expected to contain most known SLC proteins. Superfamilies #9.A and #9.B were also examined. Numbers show the total number of level 3 families and level 4 subfamilies in each superfamily, as well as those found to be "SLC-like" (see text).

TCDB superfamilies	**Level 3 families (TC# x.y.z)**			**Level 4 subfamilies (TC# x.y.z.w)**		
	Total	**Total examined**	**Found to be "SLC-like"**	**Total**	**Total examined**	**Found to be "SLC-like"**
1 – Channels/Pores	(556)			(997)		
1.A – α-Type Channels		**112**	**7**		**321**	**17**
2 – Electrochemical Potential-driven Transporters	(178)			(561)		
2.A – Porters (uniporters, symporters, antiporters)		**129**	**114**		**556**	**501**
3 – Primary Active Transporters	(47)			(250)		
4 – Group Translocators	(18)			(36)		
5 – Transmembrane Electron Carriers	(17)			(42)		
8 – Accessory Factors Involved in Transport	(163)			(252)		
9 – Incompletely Characterized Transport Systems	(388)			(657)		

	9.A – Recognized Transporters of Unknown Biochemical Mechanism		**69**	**20**		**122**	**37**
	9.B – Putative Transport Proteins		**306**	**25**		**535**	**45**
	Total:	(1367)			(2795)		
			616	**167**		**1534**	**602**

It was a significant curation effort to manually assess the 1534 subfamilies within 616 families in the above-mentioned superfamilies of the TCDB, which were expected to contain SLC transporter-like proteins (Table 1). To streamline our curation efforts, we have also expanded our analysis to Pfam protein families. To this end, we used HMMER [30] to search for all Pfam models in the sequences of all TCDB members within the four superfamilies analyzed, as shown in Table 1. The resulting Pfam families found in TCDB sequences were manually analyzed in tandem with the TCDB families and subfamilies in which they occur whether they also meet the criteria for SLC-likeness as defined above. The SLC-likeness criteria have been assessed according to the following guidelines:

- As an inclusive approach, we mainly searched for information that can be used to exclude (sub)families based on one or more criteria. Missing information (e.g. about transport mechanism, structure, potential other functions, exact chemical identity of substrate) was not interpreted as a reason for exclusion, unless it could be replaced by predictions (see below). Therefore, we strived to include (sub)families in our list that potentially fulfill the criteria for being SLC-like.
- When a (sub)family could be excluded from being SLC-like based on one criterion, the remaining criteria were not assessed.
- The functional criteria were assessed based on information presented on the TCDB web site in the family description text and in the description of certain individual members, where this was present, as well as in the textual descriptions of Pfam families. Families that claim their members to be "channels" in their descriptions were assumed to function with a channel-like mechanism and were therefore excluded, unless there was a evidence for saturable transport or alternating access mechanism for certain members. TCDB and Pfam families showing similarity to other transporter

families, either as mentioned in the family descriptions in text or *via* SCOOP-similarity [31], have been chosen to potentially fulfill the functional criteria, unless specific functional annotations indicated otherwise.

- The criterion about the size of the substrate was judged based on the known or putative substrate according to TCDB family descriptions. The assumption that certain proteins are transporters, whether based on experimental observations or purely on prediction, always entails either a tested or an assumed class of substrates. Families where no indication was available why they would function as transporters were thus excluded for the moment until further information on their function becomes available.
- The structural criterion was assessed based on the "number of transmembrane segments" annotations in the TCDB for each protein. (Sub)families where their TCDB descriptions mentioned a β-barrel structure or relationship with such proteins were excluded. Care was taken to identify outliers, i.e. proteins with less TMHs as the functional unit, which could potentially be fragments, and proteins with extra TMHs in addition to the functional core. In this regard, we based our decision on Pfam models conserved in the (sub)family and the number of THMs the model spans according to predictions at the Pfam web site for the most well-characterized members of the (sub)family. For certain Pfam families, structures and annotations from the Protein Data Bank (PDB) linked with the family on the Pfam web site were used to decide whether or not the Pfam model encodes soluble proteins, provided that the family model has sufficient coverage by the structure available.
- In some TCDB families with multicomponent transport systems, the membrane-spanning core components were identified by the presence of conserved Pfam models in all members of the family. Whether Pfam domains were membrane-spanning was deduced by their descriptions at the Pfam web site and Pfam prediction of TMHs. The TCDB description of the family also helped in deciding which components were essential and membrane-spanning. If multiple such Pfam models were found, the proteins were assumed to function as an obligate heteromultimer, and such

proteins were considered as a single entity to assess the overall number of TMHs for the structural criterion.

Strikingly, of the 1534 subfamilies examined, only 602 in 167 families were found to be SLC-like according to our criteria. In particular, of the 556 subfamilies within superfamily TC# 2.A ("porters"), only 501 appeared to meet our selection criteria. This underlines once more that the term "porter" in relation to solute transport ambiguous in this field and a clearer definition of the perception of a solute transporter protein is needed.

Our curation efforts also yielded 211 Pfam families bearing SLC-like properties, which likely contain the membrane-spanning regions of SLC-like proteins selected from the 336 Pfam families present in total in the TCDB sequences analyzed. Special care has been taken to consistently include or exclude TCDB (sub)families and their corresponding Pfam family, if applicable, from our selection. Notably, many of those Pfam families that have been excluded represented soluble structural or regulatory domains. Following our initial round of selection, we took advantage of clan groupings in Pfam and extended our selection efforts to analyze Pfam families belonging to the same clan as the selected 211 SLC-like Pfam families. This was based on the observation that many SLC-like transporter families in Pfam appear grouped into clans, so that members of these clans might represent SLC-like transporters themselves. As noted before, Pfam clans represent remotely related protein families, and while protein function is not always conserved across different families within a Pfam clan, the inspection of these families whether they meet our SLC-likeness criteria is validated by their relatedness or similarity to SLC-like families. Such a "clan expansion" procedure resulted in 12 additional Pfam families that are evolutionarily related to SLC-like protein families, of which 8 Pfam families potentially meet our criteria as SLC-like, as evaluated according to the guidelines and criteria above. Interestingly, these 8 SLC-like Pfam families currently have no representatives with a modest score (bit score > 25) in the TCDB, while 4 of the 8 families are annotated in Pfam as domains of unknown function (DUF). Thus, a total of 219 SLC-like Pfam families were identified in our search (see S1 Table).

We would like to note that our selection of SLC-like TCDB families and subfamilies as well as Pfam families is limited by the availability of information on less well characterized protein families and may therefore need

to be revised as more information becomes available. Our criteria are objective, however, and can easily be used to revise the decision as to whether a particular family is SLC-like in the light of new information.

As a next step, we wanted to know whether the selected families either from the TCDB or from Pfam have representatives in the proteomes of human and other clinically relevant organisms. For this analysis, we selected 7 organisms due to their clinical relevance or scientific utility (*Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Gallus gallus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*), for which we downloaded all sequences from the UniProt database [27], including Swiss-Prot (curated) and TrEMBL (predicted) [32] entries. Sequences of proteins in the TCDB have been aligned within each SLC-like family and subfamily and the alignments converted to HMMs for sensitive sequence similarity searches (see Methods). In addition, HMMs of SLC-like Pfam families were downloaded from the Pfam database. HMM-based similarity searches were then performed on the sequences downloaded for the 7 organisms to find proteins similar to any of the SLC-like TCDB families or subfamilies, or SLC-like Pfam domains, followed by the clustering of sequence fragments to arrive at one representative protein sequence per gene (see Methods).

The results of our search for SLC-like proteins are summarized in numbers in Table 2. Briefly, 60-68 of the 167 TCDB families have representatives in the 7 organisms (67 in human), and the organisms seem to contain 435-676 SLC-like proteins (552 in human). In total, 3750 proteins have been found in the 7 organisms studied. Notably, the number of SLC-like transporters found in human in our search is ~130 higher than previously reported [2], indicating that the human SLC-ome may be significantly larger than previously thought.

Table 2. Results of the initial search for "SLC-like" proteins. The search was performed using HMM-based sequence similarity analysis based on selected families and subfamilies from the TCDB, as well as selected Pfam models that likely encode transmembrane domains of "SLC-like" transporters (see text for details). The table shows the number of families, subfamilies from the TCDB and the number of Pfam models that had representative sequences in each organism. In addition, the initial number of "SLC-like" proteins found is shown.

Species	Level 3 families (TC# x.y.z)	Level 4 subfamilies (TC# x.y.z.w)	Pfam models	Proteins
Human (*H. sapiens*)	67	310	82	552
Mouse (*M. musculus*)	68	316	82	535
Rat (*R. norvegicus*)	68	318	81	540
Zebrafish (*D. rerio*)	66	314	82	676
Chicken (*G. gallus*)	64	307	79	475
Fruitfly (*D. melanogaster*)	61	276	68	435
Roundworm (*C. elegans*)	60	285	71	537

After arriving at this initial set of SLC-like proteins, we proceeded by attempting to classify the proteins into protein families, followed by manual database and literature searches focused on outlier proteins (families with only a single protein) and human proteins in order to exclude false positives, as detailed below. This process was repeated iteratively to finally arrive at a consistent set of proteins and corresponding families. The numbers mentioned below correspond to this final set of proteins, and all subsequent analysis was carried out on this set.

Classification into families

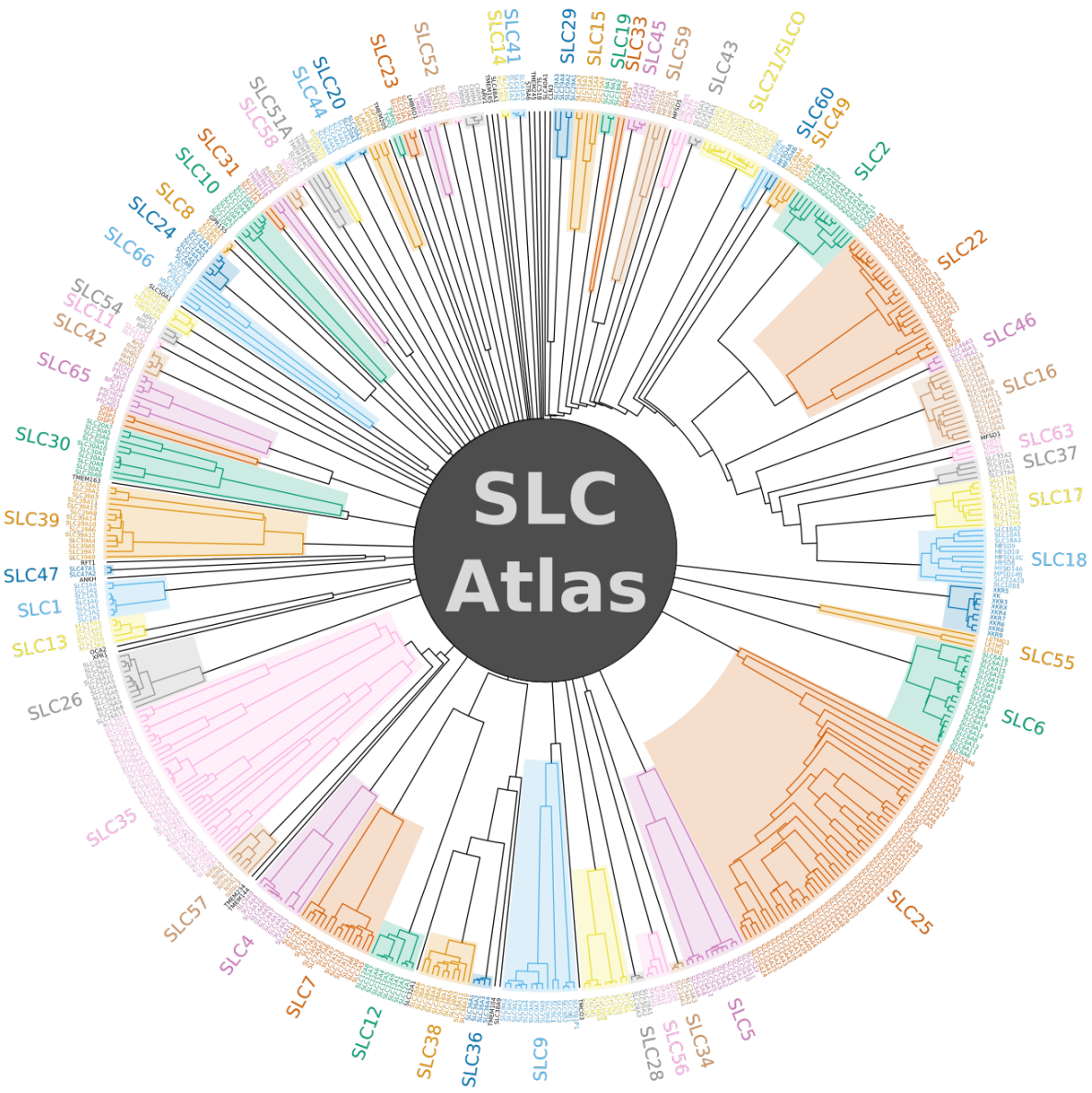
As mentioned in the introduction, SLC carriers are likely of polyphyletic origin, and individual families can be so diverse that even sensitive sequence similarity-based methods may have difficulties grouping related SLCs [12]. In our experience, multiple sequence alignment-based methods were not able to cluster the identified SLC-like sequences and to reproduce known SLC families, so we devised a custom method for clustering distantly related sequences into proteins families based on the introduction of “HMM fingerprints”. An HMM fingerprint is a mathematical vector of numbers assigned to a protein sequence, where the numbers represent the similarity scores of that protein sequence against each of the TCDB families, subfamilies and Pfam families that we have selected to be SLC-like. Thus, two protein sequences that show a similar pattern in their HMM fingerprints indicate their similarity. The usefulness of an HMM fingerprint depends on a meaningful definition of HMMs used in the fingerprint, whereby we capitalize on the evolutionary principle in the construction of TCDB families, subfamilies as well as Pfam families. Nevertheless, our goal was not to

reconstruct the evolutionary history of a set of proteins, but to group proteins that share similar sequence features. However, due to the transitivity of homology and since similarity to a group of proteins suggests homology, clusters derived using HMM fingerprinting are likely to contain homologous proteins.

The HMM fingerprint-based classification of the SLC-like proteins found in our search yielded 103 protein families in total, 94 of which had representatives in human (Fig 1). For existing SLC transporters, the generated families corresponded well to classical SLC families. Interestingly, outlier proteins were found in several families, which did not cluster with their families at the threshold we used. Examples include SLC5A7, SLC10A7, SLC25A46, SLC30A9, SLC39A9, MPDU1/SLC66A5 as well as the SLC9B family and SLC35 subfamilies. This shows that the HMM fingerprints, and thus likely the sequences of these outlier proteins diverge from those of other members of their families, and the sequences of subfamilies seem to diverge in certain cases. Several classical SLC families also clustered together, such as SLC32-SLC36-SLC38, SLC2-SLC22, and SLC17-SLC18-SLC37 proteins, likely due to their high sequence similarity. To ensure optimal correlation with the preexisting classification of classical SLC proteins, we have introduced split and join constraints to keep 1) closely related families separated, 2) outlier proteins merged with their families and 3) heterogenous families merged (see Methods). These constraints did not affect the classification of novel SLC-like proteins. The number of families proposed by our method is somewhat dependent on the threshold of clustering used. Raising it to 0.8 lowers the number of families to 98 (90 in human), joining families SLC58-pSLC.OSTC, SLC8-SLC24, SLC45-SLC59, and SLC65-pSLC.Dispatched, as well as pSLC.TSPO to a 2-membered subfamily of unannotated proteins from *D. rerio*. Since our clustering method is based on shared HMMs, these families are likely related. Indeed, SLC8 and SLC24 proteins constitute a superfamily of Na^{+}/Ca^{2+} antiporters [33], while NPC1/SLC65A1 and Dispatched proteins share structural similarity [34]. Lowering the clustering threshold to 0.6 generates 107 families (96 in human), splitting the pSLC.TMEM41-64 family to the TMEM41 and TMEM64 subfamilies, and the MFSD3 proteins off the SLC33 family, as well as splitting an unannotated protein family from *D. melanogaster* into 3 subfamilies.

Figure 1. Dendrogram of human SLC-like protein families based on HMM fingerprint-based clustering. Classical and newly incorporated SLC families are shown with colors. Since SLC proteins are polyphyletic,

which manifests itself in mathematically orthogonal HMM fingerprints that cannot be further clustered, the dendrogram does not join into a single branch. Branch lengths have been transformed for better visibility (see Methods).



Our analysis revealed 43 new proteins families in total in human, containing proteins that have not yet been annotated as SLCs. Interestingly, our search has also found new proteins that clustered into existing SLC families (Table 3, S4 Table).

Table 3. Novel SLC-like proteins. The table shows SLC-like human proteins from our search that are not in the classical group of SLCs (SLC1-52 families). The most similar (highest-scoring) TCDB family/subfamily and Pfam family are shown, as well as the most similar (highest-scoring) protein with structural information based on the pdb70 dataset. Substrate information, where available, was retrieved from available literature. Based on our search, SLC families SLC53-66 have recently been incorporated into the nomenclature in collaboration with the HUGO/HGNC. The same table can be found with additional details in S4 Table.

Gene symbol	TCDB family	Pfam family	Protein name	Structural fold	Substrates
SLC18 family					
MFSD8	2.A.1.2	MFS_1	Major facilitator superfamily domain-containing protein 8	MFS	orphan
MFSD9	2.A.1.2	MFS_1	Major facilitator superfamily domain-containing protein 9	MFS	orphan
MFSD10	2.A.1.2	MFS_1	Major facilitator superfamily domain-containing protein 10	MFS	indomethacin
MFSD14A	2.A.1.2	MFS_1	Hippocampus abundant transcript 1 protein	MFS	orphan
MFSD14B	2.A.1.2	MFS_1	Hippocampus abundant transcript-like protein 1	MFS	orphan
MFSD14C	2.A.1.2	MFS_1	Hippocampus abundant transcript-like protein 2	MFS	orphan
SLC22A18	2.A.1.2	MFS_1	Solute carrier family 22 member 18	MFS	orphan
SLC22 family					
SV2A (SLC22B1)	2.A.1.22	Sugar_tr	Synaptic vesicle glycoprotein 2A	MFS	galactose; also binds levetiracetam, seletracetam, brivaracetam.
SV2B (SLC22B2)	2.A.1.22	MFS_1	Synaptic vesicle glycoprotein 2B	MFS	orphan
SV2C (SLC22B3)	2.A.1.22	MFS_1	Synaptic vesicle glycoprotein 2C	MFS	orphan
SVOP (SLC22B4)	2.A.1.82	Sugar_tr	Synaptic vesicle 2-related protein	MFS	nicotinate; binds nucleotides in the TM region: 8-azido-ATP; ATP; GTP; TTP; CTP; NAD
SVOPL (SLC22B5)	2.A.1.82	MFS_1	Putative transporter SVOPL	MFS	orphan
SLC33 family					
MFSD3	2.A.1.57	Acatn	Major facilitator superfamily domain-containing protein 3	MFS	orphan
SLC35 family					
TMEM241	2.A.7.13	TPT	Transmembrane protein 241	NST	orphan
SLC51A family					
TMEM184A	2.A.82	Solute_trans_a	Transmembrane protein 184A		orphan
TMEM184B	2.A.82	Solute_trans_a	Transmembrane protein 184B		orphan
TMEM184C	2.A.82	Solute_trans_a	Transmembrane protein 184C		orphan
SLC53 family					
XPR1 (SLC53A1)	2.A.94	EXS	Xenotropic and polytropic retrovirus receptor 1		phosphate
SLC54 family					
MPC1 (SLC54A1)	2.A.105	MPC	Mitochondrial pyruvate carrier 1		pyruvate
MPC2 (SLC54A2)	2.A.105	MPC	Mitochondrial pyruvate carrier 2		pyruvate
MPC1L (SLC54A3)	2.A.105	MPC	Mitochondrial pyruvate carrier 1-like protein		pyruvate
SLC55 family					
LETM1 (SLC55A1)	2.A.97	LETM1	Mitochondrial proton/calcium exchanger protein		K^+/H^+
LETM2 (SLC55A2)	2.A.97	LETM1	LETM1 domain-containing protein LETM2, mitochondrial		orphan
LETMD1 (SLC55A3)	2.A.97	LETM1	LETM1 domain-containing protein 1		orphan
SLC56 family					

SFXN1 (SLC56A1)	2.A.54	Mtc	Sideroflexin-1		L-serine
SFXN2 (SLC56A2)	2.A.54	Mtc	Sideroflexin-2		orphan
SFXN3 (SLC56A3)	2.A.54	Mtc	Sideroflexin-3		orphan
SFXN4 (SLC56A4)	2.A.54	Mtc	Sideroflexin-4		orphan
SFXN5 (SLC56A5)	2.A.54	Mtc	Sideroflexin-5		orphan
SLC57 family					
NIPA1 (SLC57A1)	2.A.7.25	Mg_trans_NIPA	Magnesium transporter NIPA1	NST	Mg^{2+}; to a lesser extent Sr^{2+}, Fe^{2+}, Co^{2+}
NIPA2 (SLC57A2)	2.A.7.25	Mg_trans_NIPA	Magnesium transporter NIPA2	NST	very specific to Mg^{2+}
NIPAL1 (SLC57A3)	2.A.7.25	Mg_trans_NIPA	Magnesium transporter NIPA3	NST	Mg^{2+}, Sr^{2+}, Ba^{2+}, Fe^{2+}, Cu^{2+}
NIPAL2 (SLC57A4)	2.A.7.25	Mg_trans_NIPA	NIPA-like protein 2	NST	Mg^{2+}, Sr^{2+}, Ba^{2+}
NIPAL3 (SLC57A5)	2.A.7.25	Mg_trans_NIPA	NIPA-like protein 3	NST	orphan
NIPAL4 (SLC57A6)	2.A.7.25	Mg_trans_NIPA	Magnesium transporter NIPA4	NST	orphan
SLC58 family					
MAGT1 (SLC58A1)	1.A.76.1	OST3_OST6	Magnesium transporter protein 1	MagT	Mg^{2+}
TUSC3 (SLC58A2)	1.A.76.1	OST3_OST6	Tumor suppressor candidate 3	MagT	Mg^{2+}, Fe^{2+}, Cu^{2+}, Mn^{2+}
SLC59 family					
MFSD2A (SLC59A1)	2.A.2.3	MFS_2	Sodium-dependent lysophosphatidylcholine symporter 1	MFS	LPC DHA (docosahexaenoic acid); LPC palmitate, TopFluor LPE
MFSD2B (SLC59A2)	2.A.2.3	MFS_2	Major facilitator superfamily domain-containing protein 2B	MFS	sphingosine-1-phosphate (S1P)
MFSD12	2.A.2.7	MFS_2	Major facilitator superfamily domain-containing protein 12	MFS	cysteine
MFSD13A	2.A.2.3	MFS_2	Transmembrane protein 180	MFS	orphan
SLC60 family					
MFSD4A (SLC60A1)	2.A.1.7	MFS_1	Major facilitator superfamily domain-containing protein 4A	MFS	orphan
MFSD4B (SLC60A2)	2.A.1.7	MFS_1	Sodium-dependent glucose transporter 1	MFS	alpha-methyl-d-glucopyranoside, D-glucose, urea
SLC61 family					
MFSD5 (SLC61A1)	2.A.1.40	MFS_5	Molybdate-anion transporter	MFS	molybdate
SLC62 family					
ANKH (SLC62A1)	2.A.66.9	ANKH	Progressive ankylosis protein homolog	MATE	pyrophosphate
SLC63 family					
SPNS1 (SLC63A1)	2.A.1.49	MFS_1	Protein spinster homolog 1	MFS	orphan
SPNS2 (SLC63A2)	2.A.1.49	MFS_1	Protein spinster homolog 2	MFS	sphingosine-1-phosphate (S1P), phosphorylated Fingolimod (FTY720-P); dihydrosphingosine-1-phosphate (DH-S1P); phyto-S1P, C17-S1P
SPNS3 (SLC63A3)	2.A.1.49	MFS_1	Protein spinster homolog 3	MFS	orphan
SLC64 family					
TMEM165 (SLC64A1)	2.A.106.2	UPF0016	Transmembrane protein 165		Ca^{2+}/H^{+}, Mn^{2+}
SLC65 family					
NPC1 (SLC65A1)	2.A.6.6	Patched	NPC intracellular cholesterol transporter 1	RND	cholesterol
NPC1L1 (SLC65A2)	2.A.6.6	Patched	NPC1-like intracellular cholesterol transporter 1	RND	cholesterol
PTCH1	2.A.6.6	Patched	Protein patched homolog 1	RND	cholesterol
PTCH2	2.A.6.6	Patched	Protein patched homolog 2	RND	likely cholesterol based on functional similarity to PTCH1
PTCHD1	2.A.6.6	Patched	Patched domain-containing protein 1	RND	orphan
PTCHD3	2.A.6.6	Patched	Patched domain-containing protein 3	RND	orphan

PTCHD4	2.A.6.6	Patched	Patched domain-containing protein 4	RND	orphan
SCAP	2.A.6.6	Sterol-sensing	Sterol regulatory element-binding protein cleavage-activating protein	SCAP_full	cholesterol, 25-hydroxycholesterol
SLC66 family					
PQLC2 (SLC66A1)	2.A.43.2	PQ-loop	Lysosomal amino acid transporter 1 homolog	SWEET	L-Arg, L-Lys, L-His, L-canavanine, L-ornithine, cysteamine-cysteine mixed disulfide
PQLC1 (SLC66A2)	2.A.43	PQ-loop	PQ-loop repeat-containing protein 1	SWEET	orphan
PQLC3 (SLC66A3)	2.A.43.3	PQ-loop	PQ-loop repeat-containing protein 3	SWEET	orphan
CTNS (SLC66A4)	2.A.43.1	PQ-loop	Cystinosin	SWEET	L-cystine, L-selenocystine, L-cystathionine, H^+
MPDU1 (SLC66A5)	2.A.43.3	PQ-loop	Mannose-P-dolichol utilization defect 1 protein	SWEET	orphan
PQLC2L	2.A.43.2	PQ-loop	Putative uncharacterized protein PQLC2L	SWEET	orphan
pSLC.ARV1 family					
ARV1	9.A.19	Arv1	Protein ARV1		orphan
pSLC.Battenin					
CLN3	2.A.57.5	CLN3	Battenin	MFS	orphan
pSLC.CLCN					
CLCN1	2.A.49.2	Voltage_CLC	Chloride channel protein 1	CLC	channel for chloride, thiocyanate, perchlorate, bromide, nitrate, chlorate, iodide
CLCN2	2.A.49.2	Voltage_CLC	Chloride channel protein 2	CLC	channel for chloride, bromide, iodide
CLCN3	2.A.49.2	Voltage_CLC	H(+)/Cl(-) exchange transporter 3	CLC	antiport of chloride and H^+
CLCN4	2.A.49.2	Voltage_CLC	H(+)/Cl(-) exchange transporter 4	CLC	antiport of chloride and H^+, also nitrate, bromide, iodide
CLCN5	2.A.49.2	Voltage_CLC	H(+)/Cl(-) exchange transporter 5	CLC	antiport of chloride and H^+, also nitrate, bromide, iodide
CLCN6	2.A.49.3	Voltage_CLC	Chloride transport protein 6	CLC	antiport of chloride and H^+, also nitrate, iodide
CLCN7	2.A.49.3	Voltage_CLC	H(+)/Cl(-) exchange transporter 7	CLC	antiport of chloride and H^+
CLCNKA	2.A.49.2	Voltage_CLC	Chloride channel protein ClC-Ka	CLC	channel for chloride, bromide, nitrate, iodide
CLCNKB	2.A.49.2	Voltage_CLC	Chloride channel protein ClC-Kb	CLC	channel for chloride, bromide, nitrate, iodide
pSLC.CNNM					
CNNM1	1.A.112	DUF21	Metal transporter CNNM1	CorC_large	orphan
CNNM2	1.A.112	DUF21	Metal transporter CNNM2	CorC_large	Mg^{2+}, Co^{2+}, Mn^{2+}, Sr^{2+}, Ba^{2+}, Cu^{2+}, Fe^{2+}, Na^+
CNNM3	1.A.112	DUF21	Metal transporter CNNM3	CorC_large	orphan
CNNM4	1.A.112	DUF21	Metal transporter CNNM4	CorC_large	putatively Mg^{2+}, maybe antiport of Na^+
pSLC.CitMHS					
OCA2	2.A.45.2	CitMHS	P protein	AbgT	putatively chloride
pSLC.Dispatched					
DISP1	2.A.6.9	Patched	Protein dispatched homolog 1	RND	hedgehog protein
DISP2	2.A.6.9	Sterol-sensing	Protein dispatched homolog 2	RND	orphan
DISP3	2.A.6	Patched	Protein dispatched homolog 3	RND	putatively cholesterol
pSLC.GPR155					
GPR155	2.A.69.3	Mem_trans	Integral membrane protein GPR155	NhaA	orphan
pSLC.LAPTM					
	2.A.74.1	Mtp	cDNA FLJ61683, moderately similar to Lysosomal-associated multitransmembrane protein		orphan
LAPTM4A	2.A.74.1	Mtp	Lysosomal-associated transmembrane protein 4A	Tspn	thymidine, uridine, multi-drug resistance
LAPTM4B	2.A.74.1	Mtp	Lysosomal-associated transmembrane protein 4B		putatively ceramide, multi-drug resistance
LAPTM5	2.A.74.1	Mtp	Lysosomal-associated transmembrane protein 5	Tspn	orphan
pSLC.LMBR-A					
LMBRD1	9.A.54.1	LMBR1	Probable lysosomal cobalamin transporter		orphan
pSLC.LMBR-B					
LMBR1		LMBR1	Limb region 1 protein homolog		orphan
LMBR1L	9.A.54.1	LMBR1	Protein LMBR1L		orphan
LMBRD2	9.A.54.3	LMBR1	LMBR1 domain-containing protein 2		orphan
pSLC.MFSD1					
MFSD1	2.A.1.53	MFS_1	Major facilitator superfamily domain-containing protein 1	MFS	orphan
pSLC.MFSD6					
MFSD6	2.A.1.65	MFS_1_like	Major facilitator superfamily domain-containing protein 6	MFS	putatively sugars
MFSD6L	2.A.1.65	MFS_1_like	Major facilitator superfamily domain-containing protein 6-like	MFS	orphan
pSLC.OSTC					
	1.A.76.2	OST3_OST6	cDNA FLJ52625	MagT	orphan

OSTC	1.A.76.2	OST3_OST6	Oligosaccharyltransferase complex subunit OSTC	MagT	orphan
OSTCL	1.A.76.2	OST3_OST6	Oligosaccharyltransferase complex subunit-like	OSTC_part	orphan
pSLC.RFT1					
RFT1	2.A.66.3	Rft-1	Protein RFT1 homolog	MATE	orphan
pSLC.SIDT					
SIDT1	1.A.79.1	SID-1_RNA_chan	SID1 transmembrane family member 1		cholesterol
SIDT2	1.A.79.1	SID-1_RNA_chan	SID1 transmembrane family member 2		cholesterol
pSLC.STAR					
STARD3	9.B.64	MENTAL	StAR-related lipid transfer protein 3		probably cholesterol
STARD3NL	9.B.64	MENTAL	STARD3 N-terminal-like protein		binds cholesterol
pSLC.STRA					
STRA6	2.A.90.1	RBP_receptor	Receptor for retinol uptake STRA6	STRA6	retinol
pSLC.TMCO3					
TMCO3	2.A.37.1	Na_H_Exchanger	Transmembrane and coiled-coil domain-containing protein 3	NhaA	orphan
pSLC.TMEM14					
TMEM14A	2.A.126.1	Tmemb_14	Transmembrane protein 14A	TMEM14_alt	orphan
TMEM14B	2.A.126.1	Tmemb_14	Transmembrane protein 14B	TSPO	orphan
TMEM14C	2.A.126.1	Tmemb_14	Transmembrane protein 14C	TSPO	protoporphyrinogen IX
TMEM14DP	2.A.126.1	Tmemb_14	Transmembrane protein 14DP	TSPO	orphan
pSLC.TMEM41-64					
TMEM41A	9.B.27.1	SNARE_assoc	Transmembrane protein 41A		orphan
TMEM41B	9.B.27.1	SNARE_assoc	Transmembrane protein 41B		scramblase of cholesterol, phosphatidylserine
TMEM64	9.B.27.5	SNARE_assoc	Transmembrane protein 64		orphan
pSLC.TMEM104					
TMEM104	2.A.18.10	Aa_trans	Transmembrane protein 104	APC	orphan
pSLC.TMEM144					
TMEM144	2.A.7.8	TMEM144	Transmembrane protein 144	NST	orphan
pSLC.TMEM163					
TMEM163	2.A.4.8	Cation_efflux	Transmembrane protein 163	CDF	Zn^{2+}
pSLC.TMEM205					
TMEM205	9.A.55.1	DUF4149	Transmembrane protein 205		orphan
pSLC.TMEM234					
TMEM234	2.A.7.32	TMEM234	Transmembrane protein 234	NST	orphan
pSLC.TMEM245					
TMEM245	2.A.86	AI-2E_transport	Transmembrane protein 245		orphan
pSLC.TSPO					
TSPO	9.A.24.1	TspO_MBR	Translocator protein	TSPO	protoporphyrin IX, cholesterol
TSPO2	9.A.24.1	TspO_MBR	Translocator protein 2	TSPO	5-aminolevulinic acid
pSLC.UNC93					
MFSD11	2.A.1.58	UNC-93	UNC93-like protein MFSD11	MFS	orphan
UNC93A	2.A.1.58	UNC-93	Protein unc-93 homolog A	MFS	orphan
UNC93B1	2.A.1.58	UNC-93	Protein unc-93 homolog B1	MFS	orphan
pSLC.XK					
XK	2.A.112.1	XK-related	Membrane transport protein XK	XK	orphan
XKR3	2.A.112.1	XK-related	XK-related protein 3	XK	orphan
XKR4	2.A.112.1	XK-related	XK-related protein 4	XK	phosphatidylserine
XKR5	2.A.112.1	XK-related	XK-related protein 5	XK	orphan
XKR6	2.A.112.1	XK-related	XK-related protein 6	XK	orphan
XKR7	2.A.112.1	XK-related	XK-related protein 7	XK	orphan
XKR8	2.A.112.1	XK-related	XK-related protein 8	XK	phosphatidylserine
XKR9	2.A.112.1	XK-related	XK-related protein 9	XK	phosphatidylserine
XKRX	2.A.112.1	XK-related	XK-related protein 2	XK	orphan

Structural homologues

For polytopic transmembrane proteins, structural similarity can support evolutionary relatedness, while at the same time evolutionary relatedness can provide a basis for homology-based model building efforts [35,36]. On the other hand, the lack of predicted similarity to any protein of known structure could pin down interesting targets for structural biology efforts by highlighting proteins likely belonging to new fold families.

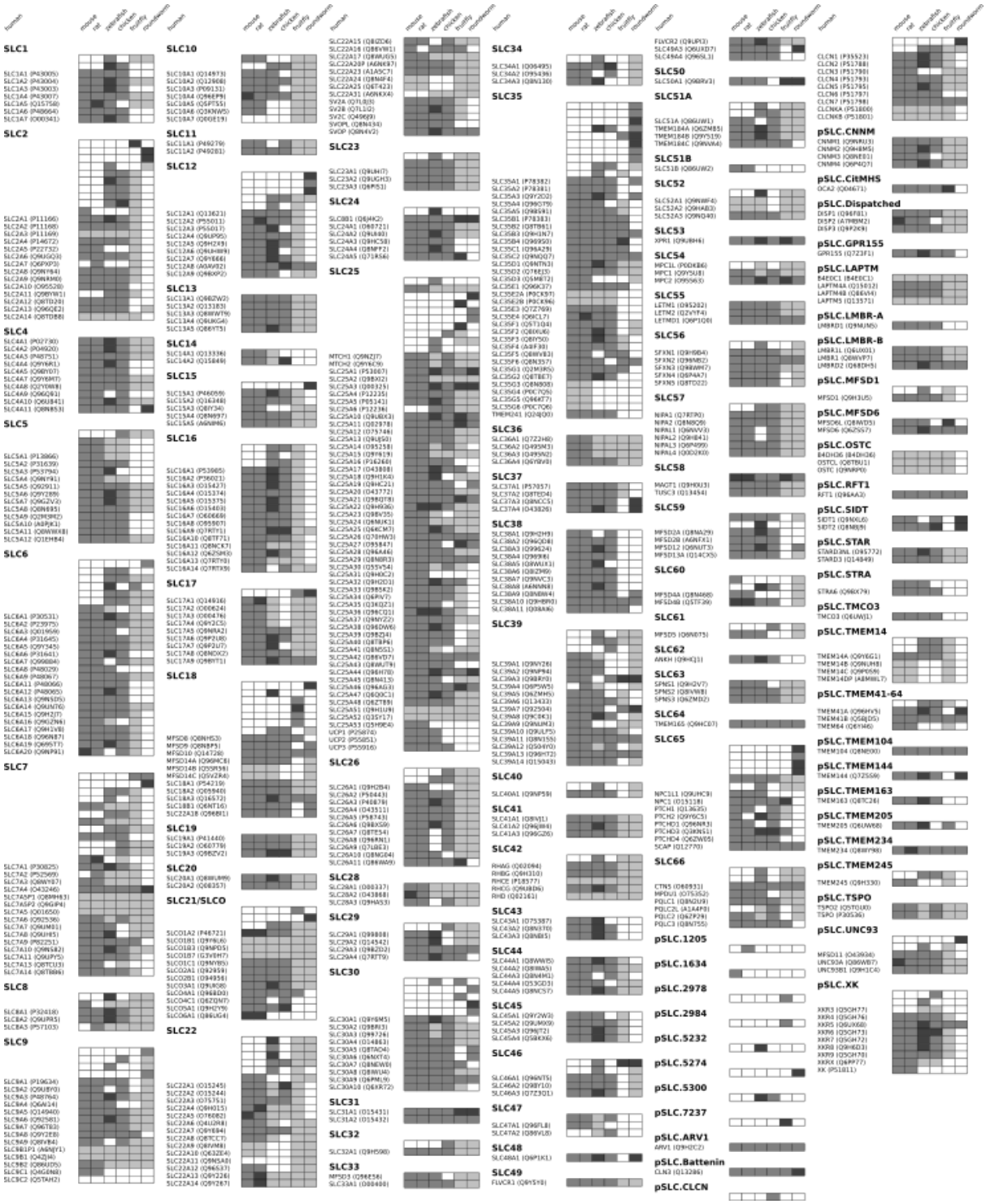
We performed HMM-based searches on the pdb70 database (sequences of the proteins in the Protein Data Bank clustered to 70% sequence identity, see Methods) to assess whether structural homologues are available for the proteins found. In total, for 80 of the 103 families, at least one similar protein was found with a corresponding structure in the Protein Data Bank. Importantly, 476 human SLC-like proteins likely have a homologue whose structure has been solved. On the other hand, 43 human SLC-like proteins belonging to 19 different families do not seem to have homologues with a known structure and thus are likely to constitute novel fold families. For the classical SLC proteins, their best-scoring similar proteins from the pdb70 dataset and the corresponding structural fold families are summarized in S3 Table. Based on this, it appears that classical SLCs from families SLC34, SLC44, SLC48 and SLC51 are still "structural orphans".

Phylogenetic analysis

Model organisms can be useful to study the biological function of various proteins, including solute transporters. In order to relate the results to human, however, knowledge of orthologous gene pairs is necessary. To this end, we performed phylogenetic analysis on each family of SLC-like proteins corresponding to our clustering. In brief, unrooted phylogenetic trees for each SLC-like protein family from all organisms have been generated and reconciled with the species tree of the 7 organisms in our study to identify gene duplication and speciation events in their evolutionary history (see Methods). The resulting evolutionary trees are deposited in S1 File. Based on these trees, we carried out orthology analysis focused on human proteins and the human lineage. The resulting data is presented in Fig 2, showing relationships between human genes and their orthologs in the other 6 organisms in our study. In addition, gene clusters are presented that have arisen through gene duplication events in the evolutionary history of the human lineage, but the corresponding human genes have likely been lost.

Figure 2. Orthology analysis of SLC-like proteins based on the human lineage. Each line shows a gene cluster that has evolved through gene duplication events along the human lineage. The human gene in the cluster, if present, is noted in the text label. Normal grey boxes in each column indicate a 1:1 orthology relationship between the human gene and a corresponding gene in the organism specified. Dark grey boxes indicate that the human gene has several orthologs, and light grey boxes indicate that several human genes

share a common ortholog in that organism. White boxes denote no ortholog in the organism specified, while lines without a human gene name correspond to genes that have been lost in human.



Literature search on newly found SLC-like proteins

As alluded to above, our initial search was followed by thorough investigation of the available literature to check whether a description of transport activity for the newly found proteins is available. Surprisingly, our search has revealed human proteins for which transported substrates are known. These data have been included in Table 3.

We also found several sequences that we believe could be false positives for various reasons. These can be fragments of other proteins, non-human proteins, translated sequences of pseudogenes or proteins that have shown partial similarity to existing transporters or proteins in Table 3, but have well-known non-transporter functions. These proteins have been omitted from Table 3, and are instead summarized in S2 Table. After such filtering, 3676 SLC-like proteins have remained, out of which 519 were from human, and 130 of them were from non-classical SLC families, indicating putative novel SLC-like proteins.

Discussion

Our curation and search efforts have revealed a surprising 130 human proteins that are SLC-like but have not been officially part of the SLC nomenclature before the start of our study. Interestingly, around 30 of them were addressed in an earlier study and referred to as “atypical SLC transporters” [10,11]. All of these atypical transporters have also been identified in our study, together with many others, including several that are not part of the major facilitator superfamily (MFS) or the amino acid-polyamine-cation (APC) transporter superfamily.

Our HMM fingerprint-based classification method yielded 103 protein families. Interestingly, in order to reproduce some of the classic SLC families, we had to introduce clustering constraints to artificially split or merge predicted families. In turn, increasing or decreasing the clustering threshold by 0.1 changes the number of families to 98 and 107, respectively. Researchers have reported similar problems with the classification of the highly diverse protein set of the Pfam database [29]. In particular, Finn et al. note that closely related Pfam families may have artificially high thresholds to prevent them from overlapping, while divergent families cannot always be covered by a single model [29]. Thus, we believe that however protein

families are shaped, the discovery of evolutionary relationships between them will always be necessary to define higher-order superfamilies, similarly to Pfam clans [29]. Similarly, the analysis of evolutionary relationship of proteins within each family is equally validated, and several classic SLC families indeed contain established subfamilies based on sequence and/or functional comparisons [37–40]. In this regard, we believe that our work provides correlation between transporter proteins in the studied organisms with the hierarchy of families and subfamilies in the TCDB database on the one hand, and a scale-free framework for protein grouping and classification through HMM fingerprints on the other hand. Still, more work is needed to examine the evolutionary relatedness of different SLC-like families using more robust phylogenetic approaches.

The phylogenetic trees of SLC-like protein families presented in S1 File can be instrumental for functional annotation based on orthology analysis in lower organisms, as well for studying species-specific differences in transport pathways. Of particular note, different orthologs of hepatic drug transporters are present in humans and experimental mammals used in pre-clinical studies, contributing to imperfect prediction of drug half-life and toxicity in animal models [41]. On the other hand, genetic ablation of transporters and phenotype studies in lower organisms could shed light on the function of their human orthologs. However, as can be seen in Fig 2 and the phylogenetic trees in S1 File, a significant number of SLC-like protein sequences were found in *D. melanogaster* and *C. elegans* that do not appear to have orthologs in the higher organisms studied here. Most of these protein sequences are poorly annotated and their expression and function have yet to be confirmed.

The interesting cases of apparent outlier sequences in several families (SLC5A7, SLC10A7, SLC25A46, SLC30A9, SLC39A9, MPDU1/SLC66A5 as well as the SLC9B and SLC35 families) have previously been partially documented. These members show striking sequence and/or functional divergences from the other members of their SLC families. SLC5A7 diverges from other SLC5 proteins in phylogenetic trees [42,43] and shares only 20-25% sequence identity with other members of the SLC5 family [44]. The sequence of SLC10A7 is more similar to its bacterial relatives than to other SLC10 members, and its genetic structure, particularly its number of exons, is also different from other proteins in the SLC10 family [45,46]. SLC10A7 also appears

to have diverged at the functional level, as it has been suggested to be a regulator of Ca^{2+} influx [47], while having no documented transport activity. SLC25A46 turns out to be an outer mitochondrial membrane protein [48], in contrast to most other members of the SLC25 family, which are generally located on the inner mitochondrial membrane. Consistent with this, its yeast ortholog Ugo1, referred to as "a degenerate member of the mitochondrial metabolite carrier family", was reported to be part of the mitochondrial outer membrane fusion machinery [49]. Similarly, MTCH1 and MTCH2, which also seem to be distantly related to other SLC25 members according to our dendrogram (Fig 1), are also outer mitochondrial membrane proteins [50,51], and together with SLC25A46, are collectively referred to as "peculiar" members of the family [52]. SLC30A9 is found in the cytosol and nuclear fractions and functions as a coactivator of nuclear receptors after hormonal stimulation [53], in contrast to other members of the family, which are Zn^{2+} transporters. SLC39A9, the only member of "subfamily I" of Zrt/Irt-like proteins (ZIPs) [37] might function as an androgen hormone receptor [54], while most other family members are primarily transporters of divalent metal ions. SLC9B subfamily proteins NHA1 and NHA2 of the SLC9 Na^+/H^+ exchanger family are more similar to their bacterial homologs than to SLC9A members [55].

In the following sections, we discuss the non-classical set of human SLC-like proteins resulting from our search, with a particular focus on the identities of the transported substrates and, if known, the structural and mechanistic aspects of transport. In certain cases, the proteins have already been officially included in the SLC nomenclature as per our initiative and as a result of this work, following approval by the Human Gene Nomenclature Committee (HGNC). Where existing information about substrate and function is not available, we have speculated on these aspects using available information in the literature and considerations of sequence similarity. For several proteins, however their classification requires further consideration. We would also like to articulate open questions about what further work is needed in order to identify additional transporters. We believe our approach has been useful to pinpoint proteins that have a high probability of being novel transporters. Thus, specific biological, biochemical and structural efforts could focus on these specific targets highlighted in our work, all of which would contribute to a complete assessment of the SLC-ome in human cells.

Transporters with existing evidence for transport function

Importantly, our search highlighted several proteins for which our literature search uncovered previously reported evidence of transporter activity. Several of the proteins in these families have been assigned SLC family numbers and, in collaboration with the HGNC, included in the SLC nomenclature. Among these hits are several mitochondrial transporters (MPC/SLC54, LETM/SLC55, Sideroflexins/SLC56), which have been reviewed before [56]. In terms of transported substrate, many of the proteins with documented transport activity appear to be ion transporters or exchangers (Table 3). In the next paragraph, we will highlight certain proteins and families that have particularly caught our attention.

Interestingly, the **SLC60** family contains two MFS-like proteins (MFSD4A, MFSD4B), of which MFSD4B has been shown to transport D-glucose and urea [57,58].

The **SLC61A1** protein (MFSD5) is the only protein in human that shows similarity to the #2.A.1.40 family of molybdate transporters and contains the "MFS_5" Pfam domain. It has been claimed to be the homolog of similar transporters from algae and plants, and complementation assays suggested its ability to transport molybdate [59]. While molybdenum is a biologically active trace element, not much is known about its transport and homeostasis in human [60].

Interestingly, the **TMEM163** protein clustered together with SLC30 zinc transporter (ZnT) proteins, since the "Cation_efflux" Pfam domain, representative of the SLC30 family, was present in its sequence, although at a low score and non-significant e-value (2.9e-3). In the TCDB, TMEM163 is also classified under subfamily #2.A.4.8, sharing a common family with SLC30 transporters (#2.A.4.2). Multiple sequence alignment as well as pairwise alignments with existing SLC30 members reveal very low sequence identity with SLC30 proteins (4.2-14.4%), albeit these numbers are similar to those of SLC30A9 (6.5-13.4%). Given the marginal similarity to the "Cation_efflux" domain, it is tempting to assume that SLC30 proteins and TMEM163 are distantly related. Indeed, TMEM163 has been shown to bind [61] and transport Zn^{2+} [62–64], and substitution of its proposed substrate-binding residues with alanine abolished Zn^{2+} efflux activity [64]. Transport has been demonstrated to be H^{+}-coupled, and the protein functioning as a dimer [62], while extruding Zn^{2+} from the

cell [64]. Intracellularly, TMEM163 was originally shown to be expressed in synaptic vesicles [65]. In overexpression systems, it is localized to both the plasma membrane and intracellular membrane compartments [64]. TMEM163 has been linked to Parkinson's disease (PD) [66], even though the opposite conclusion has also been drawn [67]. TMEM163 has also been reported to be upregulated by olanzapine, a psychotropic drug prescribed for PD patients [68]. In addition, TMEM163 was also shown to be highly expressed in insulin secretory vesicles in human pancreas [69], and has been identified as a risk factor in type 2 diabetes [70,71]. Disruption of TMEM163 expression might impair insulin secretion at high glucose stimuli [69].

The **TMEM165** protein clustered into its own family and is the only protein in human containing the "UPF0016" Pfam domain and showing similarity to TCDB family #2.A.106.2. TMEM165 is a member of a highly conserved family of transmembrane proteins that is present in many species of eukaryotes and bacteria [72]. Initially, TMEM165 and its yeast homolog, Gdt1p, have been hypothesized to be Ca^{2+}/H^{+} exchangers [73,74]. However, recently, evidence has been mounting about its involvement in manganese (Mn^{2+}) homeostasis [74], and both Ca^{2+} and Mn^{2+} transport activity has directly been shown [75]. TMEM165 is localized to the trans-Golgi in human cells [72], and is proposed to play a crucial role in regulating Mn^{2+} uptake into the Golgi apparatus [72,74]. In line with this, its homologs in other organisms, also containing the UPF0016 domain, are also annotated as Mn^{2+} transporters [74]. Manganese plays an important role as a co-factor for enzymes involved in glycosylation, and impairment of TMEM165 function results in glycosylation defects. Indeed, mutations of TMEM165 found in patients with congenital disorder of glycosylation (CDG) type II hamper the transport function or localization of TMEM165 [75]. Due to the importance of TMEM165 in lactate biosynthesis [76], it has also been suggested that TMEM165 could be a transporter importing both Ca^{2+} [77] and Mn^{2+} into the Golgi in exchange for protons [74]. TMEM165 proteins contain two copies of the UPF0016 domain, and each domain contains a signature motif, E-φ-G-D-(K/R)-(T/S), where φ denotes a hydrophobic amino acid. The glutamic acid of the second motif, E248, has been shown to be crucial for affecting the glycosylation function of the Golgi but not the expression of the protein [78], and so can be speculated to form part of a binding site for transporter function. However, in

the absence of an experimentally determined structure, further investigation will be required to understand the transport mechanism of TMEM165.

Proteins with sequence similarity to existing transporters

Our search uncovered a large number of proteins that show sequence similarity and thus possible relationships to existing transporters in the SLC nomenclature. Since transport activity has not been demonstrated, these proteins are either orphan transporters or they could have transceptor functions. What follows is a comprehensive discussion of these proteins, as their similarity to transporters makes them ideal targets for further studies to elucidate their putative transporter activity.

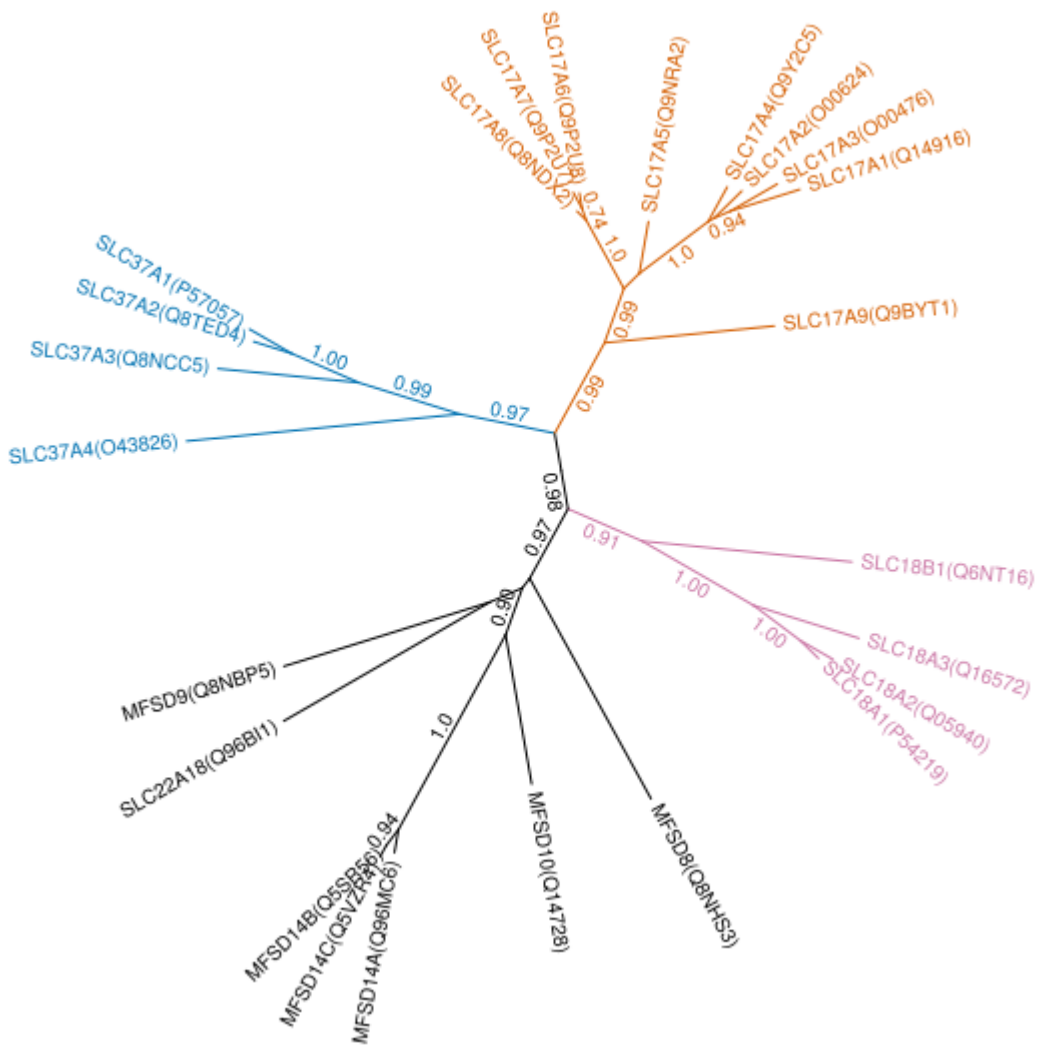
Atypical transporters

A previous effort by Perland and coworkers has uncovered novel transporter-like proteins mostly from the MFS and APC superfamilies [11], which have also been recognized by our search. In general, the function of these atypical transporters is not well known, but some have been reported to be expressed in the brain, and their expression levels seem to be affected by nutrient availability [79–82]. For MFSD1, MFSD6 and UNC93A, the study of the *D. melanogaster* and *C. elegans* orthologs have provided some information on the loss-of-function phenotype [83–86]. MFSD8 and MFSD10 have been linked to the Wolf-Hirschhorn syndrome and to LINCL (late-infantile-onset neuronal ceroid lipofuscinoses), respectively [87,88], and MFSD8 seems to be localized in the lysosomes [89,90]. More studies about the biological function and transport activity of these proteins is required to fully understand their physiological roles.

Some of the “atypical” SLC-like transporters (e.g. MFSD8, MFSD9, MFSD10 and MFSD14 proteins) clustered together with members of the classical SLC18 family, which prompted us to examine the relationship of these and neighboring proteins in more detail. We constructed multiple alignments and a phylogenetic tree of the proteins one level above these proteins in our clustering dendrogram (i.e. members of the SLC17, SLC18, SLC37 families as well as MFSD8, MFSD9, MFSD10, MFSD14A-C and SLC22A18 proteins, Fig 3). As expected, the phylogenetic tree gives a better separation of these very similar proteins than the HMM fingerprint-based dendrogram, and the branch support values suggest a clear separation of the SLC17, SLC18

and SLC37 families. In addition, the phylogenetic tree highlights that the atypical SLC proteins MFSD8, MFSD9, MFSD10, MFSD14A-C as well as SLC22A18, while being more divergent, are likely to have evolved from a single common ancestor. The relationship between the MFSD9, MFSD10 and MFSD14A-B proteins also agrees with earlier studies [Perland2017a, Perland2017b]. Similarly, the evolutionary dendrogram created using all 7 organisms in our study for the SLC18 family (S1 File) suggests that MFSD9, MFSD10, MFSD14A-B and SLC22A18 likely share a common evolutionary origin and are thus more closely related to each other than to SLC18 proteins, while MFSD8 is more distantly related. Further studies may be required to elucidate the particular evolutionary relationship between these proteins.

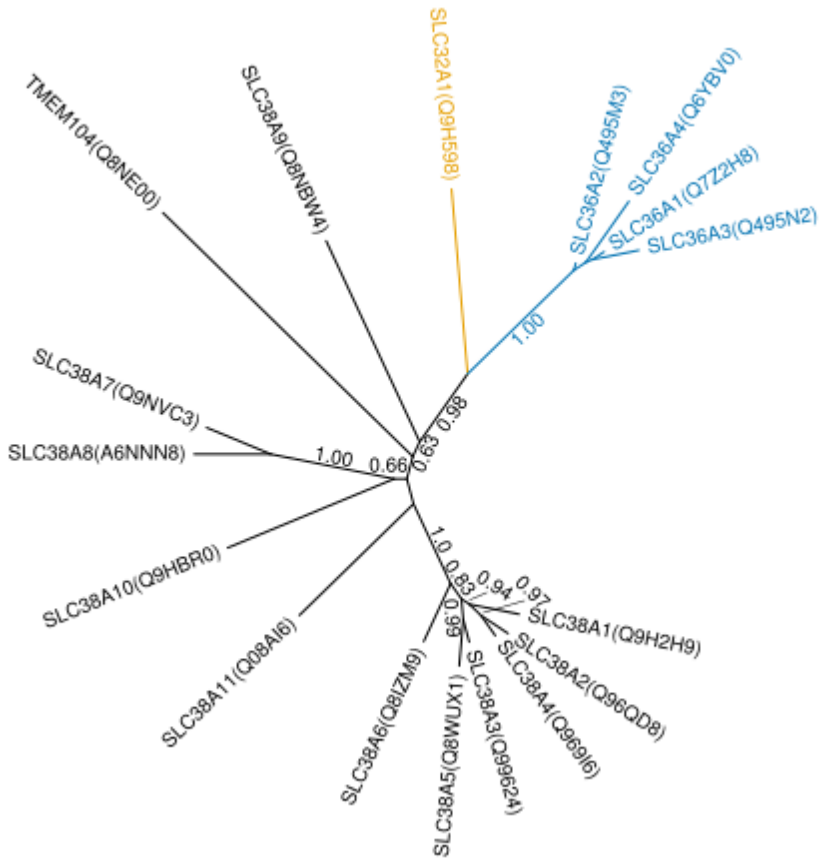
Figure 3. Phylogenetic tree of the SLC17, SLC18 and SLC37 families and proteins clustering in their neighborhood. Branch support values ≥ 0.7 are shown.



Interestingly, **MFSD3** has clustered together with the SLC33A1 protein in our HMM fingerprint-based clustering analysis. Indeed, the "Acatn" (Acetyl-coenzyme A transporter 1) Pfam domain is present in MFSD3, albeit with a relatively low score, but with significant e-value (5.6e-12). Sequence alignment between SLC33A1 and MFSD3 gives 18.2% sequence identity. Even though the sequence identity between MFSD3 and SLC33A1 is relatively low, the "Acatn" domain was found only in these proteins. The relatedness of MFSD3 and SLC33A1 is also corroborated by previous results of other groups [81]. The biological function of MFSD3 is still unclear [81,91,92].

The **TMEM104** protein in our analysis clustered together with amino acid transporter families SLC32, SLC36 and SLC38. Based on multiple and pairwise sequence alignments and sequence identity, TMEM104 was most similar to SLC38A7 (13.3-15.1%), SLC38A8 (13.1-15.1%), and SLC36A1 (10.9-16.1%). Interestingly, TMEM104 also bears moderate similarity to the "Aa_trans" Pfam domain, which describes the transmembrane region of SLC38 proteins. In our SLC classification dendrogram (Fig 1), TMEM104 clustered with SLC38 proteins, even though it seems to be an outlier from the family, similarly to SLC38A9. In addition, TMEM104, SLC38A7 and SLC38A8 all show low similarity to the "Trp_Tyr_perm" Pfam domain, which describes bacterial tyrosine and tryptophan permeases. Despite the low sequence similarity to SLC38 members, these data suggest that TMEM104 might be an amino acid transporter distantly related to the SLC38 family. To get a more detailed picture of the evolutionary relationship of TMEM104 and the SLC38, SLC36 and SLC32 families, we constructed a multiple alignment and a phylogenetic tree of these proteins (Fig 4). While the tree undoubtedly separates the SLC32 and SLC36 clades due to high branch support values, TMEM104 could not be clearly separated from the SLC38 family, and it likely has a similar relationship to the rest of the family as SLC38A9, playing a transceptor role in cells. However, there is currently no experimental evidence of this and the biological function of TMEM104 remains elusive.

Figure 4. Phylogenetic tree of proteins clustering in the SLC38-36-32 families and TMEM104. Branch support values ≥ 0.5 are shown. SLC32 and SLC36 family members are colored gold and blue, respectively.



Proteins similar to SLC35 transporters

Interestingly, our search revealed several proteins that show sequence similarity to transporters of the SLC35 family. SLC35 proteins are currently classified into subfamilies A-G, which have relatively low sequence identity among them (4.0-22.0%). SLC35 transporters belong to the family of “DMT” (drug-metabolite transporters), which is classified in TCDB family #2.A.7, and corresponds to a clan of Pfam families, called DMT. Currently known substrates of human SLC35 members include nucleotide-sugar conjugates [40]. However, the substrate range of this superfamily is substantially larger [93].

The **TMEM144** proteins harbors its dedicated Pfam domain called “TMEM144”, which itself is a member of the DMT clan of transporters. Its relatedness to the DMT family is further corroborated by high-scoring similarity of the TCDB subfamily #2.A.7.8 to the sequence of TMEM144. Otherwise, functionally, the protein is uncharacterized, although it might be related to sterol metabolism/transport, because its function has

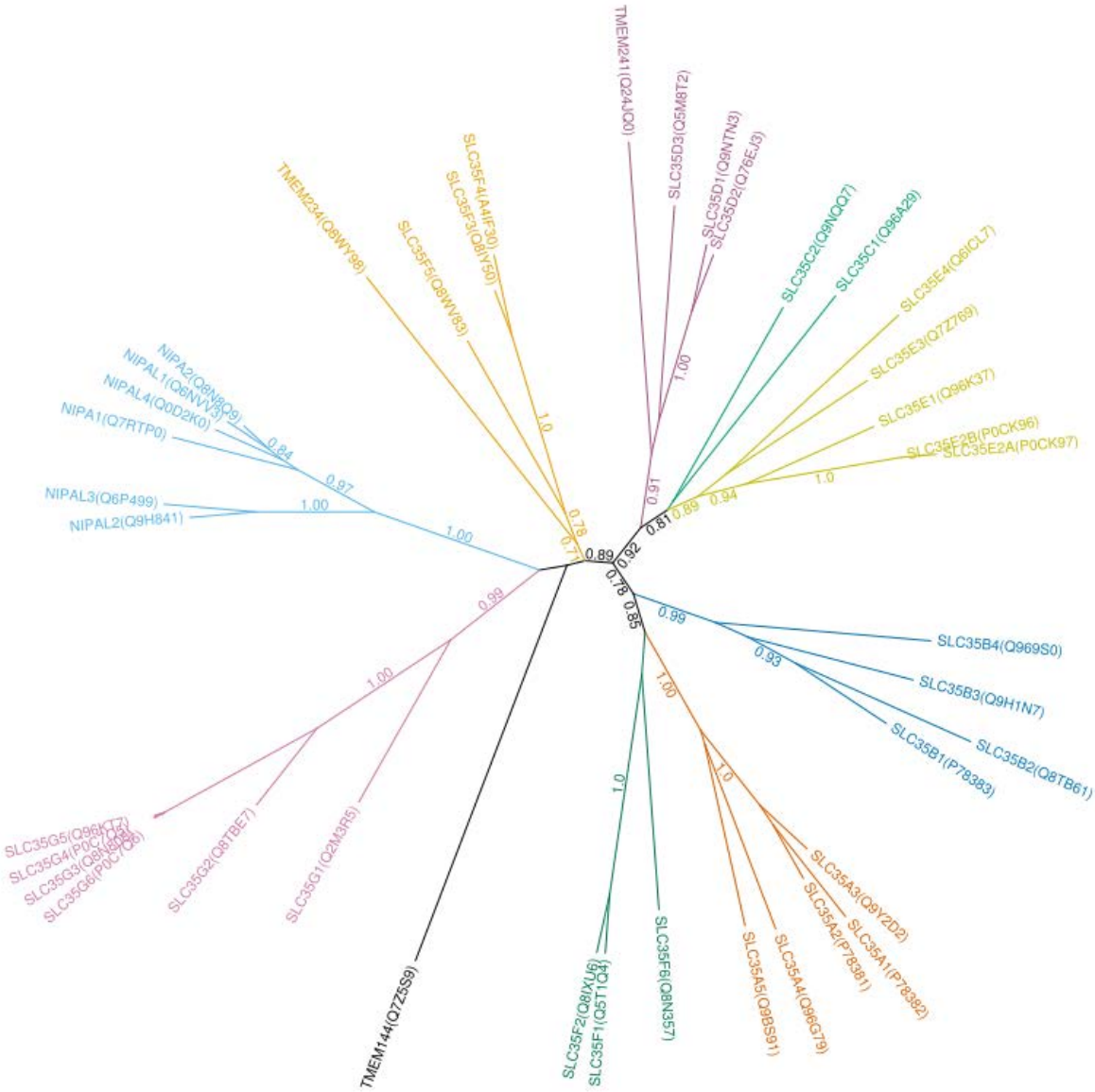
been linked to bovine milk cholesterol levels [94], the hypothalamic-gonadal axis and testosterone response [95]. It is also highly expressed in the hypothalamus [95].

TMEM234 is classified in the TCDB #2.A.7.32 family and also contains a corresponding "TMEM234" Pfam domain, which is a member of the "DMT" clan of Pfam domains as well. The physiological role of TMEM234 is not known. However, in zebrafish, its homolog might play a role in the formation of the kidney filtration barrier, as its knockdown causes proteinurea [96].

In our clustering analysis, **TMEM241** clustered with the SLC35 family very closely. Its HMM fingerprint shows similarity to the TC# 2.A.7.13 subfamily, and weak similarity to the "TPT" Pfam domain (which also belongs to the "DMT" Pfam clan) over the whole length of the protein. The proteins in the #2.A.7.13 family are Golgi GDP-mannose:GMP antiporters from plants, yeast and other organisms [97,98], but not from vertebrates. Nevertheless, the protein seems to be present in many higher organisms according to the Swiss-Prot database. However, these protein are not listed in the TCDB. The biological function of TMEM241 is still unknown, but it has been suggested to affect serum triglyceride levels [99].

Due to the sequence diversity of the SLC35 family, we were interested in the relationships between individual proteins. To this end, we have built a phylogenetic tree of human proteins that showed similarity to existing SLC35 transporters (Fig 5). In the tree, most SLC35 subfamilies could be resolved as a single clade, while TMEM241 and TMEM234 form clades with SLC35D and SLC35F3-5 proteins with a support value of 0.91 and 0.71, respectively. The relationship of TMEM241 with the SLC35D subfamily is also supported by our HMM fingerprint-based clustering results. TMEM241 shows 12.0-21.4% sequence identity with SLC35D proteins. In contrast, our phylogenetic tree with SLC35 proteins from all 7 organisms (S1 File) indicated that TMEM241 is most closely related to SLC35E4. On the other hand, TMEM234 only weakly associated with SLC35F proteins, with sequence identities 3.6-9.2%. TMEM144 appears to be only distantly related to SLC35 proteins. The elucidation of evolutionary relationships between proteins in the SLC35 thus likely requires further investigation.

Figure 5. Phylogenetic tree of proteins with HMM fingerprints overlapping with the SLC35 family. Branch support values ≥ 0.7 are shown. The different subfamilies are colored in various colors.



Others

GPR155 is an enigmatic protein that seems to be a concatenation of a membrane transporter domain (Pfam: "Mem_trans") and a G-protein coupled receptor (GPCR) domain, which might be the reason why it is annotated as a GPCR. The membrane transporter part seems to be most similar to TCDB #2.A.69.3 subfamily proteins, which are annotated as malate/malonate transporters in the Auxin Efflux Carrier (AEC) family

(#2.A.69). Gene structure analysis suggests that the concatenation is real [100], and both the human, mouse and fruitfly proteins seem to contain 17 TMHs according to UniProt annotations.

The “Mem_trans” domain is only present in GPR155 from all human proteins analyzed, and matches the first 10 TMHs of the protein in a 5+5 arrangement. In our structural search, the second half of this transporter domain (TMHs 6-10) of GPR155 exhibits similarity to the N-terminal half of sodium/bile transporters of the AsbT fold (SLC10 family). On the other hand, the last 7 TMHs of GPR155 (TMHs 11-17) indeed show similarity to GPCR-fold (7-TM) proteins with known structure, with highest similarity to structures of the human Smoothened receptor homolog (PDB ID: 6OT0). Interestingly, in our search, the N-terminal half of the transporter domain of GPR155 did not show any structural homologues.

The precise function of GPR155 still remains elusive. However, because highest expression levels were found in the brain, especially in GABAergic neurons, it might play a role in GABAergic neurotransmission [100]. It also has been suggested that GPR155 might play a role in neurons involved in motor brain function as well as sensory information processing [100]. In *D. melanogaster*, knockdown of the homologous gene, “anchor”, resulted in increased wing size and thickened veins [101]. This phenotype was similar to what appeared in bone morphogenetic protein (BMP) signaling gain-of-function experiments [101]. GPR155 has also been linked to a number of different cancers [102,103].

Somewhat surprisingly, **KDEL** receptor proteins also turned up in our search (S2 Table). The structure of KDEL receptors is known and interestingly reveals similarity to the structure of SWEET transporters, having 7 TMHs in a 3+1+3 arrangement [104]. Indeed KDEL receptors have a common evolutionary origin with SWEET as well as PQ-loop transporters [105], which are represented in human by the SLC50 and SLC66 families, respectively. Oddly, while KDEL receptors have been well characterized, they are not known to harbor transport activity.

The **RFT1** protein was originally thought to be a scramblase of lipid-linked origosaccharides [106]. However, these molecules have at least 12-14 sugar moieties, so given their size, it is unlikely that a single transporter

could catalyze their flipping. Later studies refuted the scramblase concept and suggested instead that RFT1 could serve as an accessory protein to a flippase, but would not act as a flippase itself [107–109].

Nevertheless, the corresponding "Rft1" Pfam domain shows similarity to multidrug and toxic compound extrusion (MATE) transporters (SLC47 family) and belongs to a clan of Pfam domains ("MviN_MATE") that contain transporters as well. In line with this, human RFT1 showed significant similarity to MATE transporters in our search for structural homologs, indicating likely structural similarity. Some members of the corresponding TC# 2.A.66.3 subfamily also contain weak hits of the "MatE" Pfam domain. Thus, while RFT1 shows similarity to existing transporters, its biological function is still unclear.

The C-terminal half of the **TMEM245** protein shows weak similarity to TC# 2.A.86 proteins (Autoinducer-2 Exporter/AI-2E family), which contain both small-molecule exporters [110,111] as well as Na^+/H^+ antiporter proteins [112–114]. Accordingly, TMEM245 also has weak similarity to the corresponding Pfam domain ("AI-2E_transport").

The HMMs of TC# 2.A.86 and #2.A.86.1 match from residues 444-866 of human TMEM245, which are the last 6 TMHs according to UniProt predictions. The last 5 TMHs are separated from the previous ones by a slightly larger loop. This architecture is similar to the 3+5 arrangement of the previously described bacterial $Na^+(Li^+)/H^+$ antiporter TC# 2.A.86.1.14 according to UniProt predictions (accession code: NLHAP_HALAA). This bacterial protein also matched the full-length "AI-2E_transport" domain from Pfam, while only the last 5 TMHs of TMEM245 match with C-terminal region of "AI-2E_transport". The human TMEM245 protein contains 14 TMHs in total according to UniProt predictions. Thus, TMEM245 might have a transporter-like domain at the C-terminus. In terms of the structure, we have not found any similarities to proteins with known structure. Therefore both TMEM245 and bacterial exporters and antiporters in the TC# 2.A.86 family are likely to have a yet uncharacterized tertiary structure. Functionally, the TMEM245 protein also remains elusive.

The **TMEM41A**, **TMEM41B**, and **TMEM64** proteins clustered to the same family in our results. These are the only proteins in human that show any similarity to the "SNARE_assoc" Pfam domain, as well as to the TCDB

family #9.B.27. While no protein with this domain or from this family has direct evidence for transport activity, Pfam reports SCOOP-based similarity [31] of the "SNARE_assoc" domain with "Sm_multidrug_ex", which is a domain encoding transporter proteins. Some members of the family in the TCDB have been proposed to be "cation:proton importers" (#9.B.27.2.2) or "selenite transport proteins" (#9.B.27.2.3).

The most well-characterized member of the human protein family is TMEM41B. Interestingly, a recent study reported a putative structure generated *ab initio* using evolutionary covariance-derived information [115]. Strikingly, this structural model shows features reminiscent of secondary transporters, such as a tandem internal repeat with two-fold rotational symmetry, and the authors suggest a H^+ antiporter activity as a mechanism of transport [115].

While the exact function of TMEM41B is still unclear, it forms a complex with vacuole membrane protein 1 (VMP1), also harboring the "SNARE_assoc" domain, and both are required for autophagosome formation [116,117]. Tmem41b localized to mitochondria-associated ER membranes [118–120]. Interestingly, TMEM41B seems to be an absolutely required factor for SARS-CoV-2 [121], and probably also flaviviral [122] infection, possibly by facilitating a membrane curvature that is beneficial for viral replication [122].

The proteins **TMEM184A**, **TMEM184B**, **TMEM184C** clustered together with **SLC51A** (family of transporters of steroid-derived molecules) in our analysis. While human TMEM184B and SLC51A are included in the TCDB as members of family #2.A.82, TMEM184A and C are not. Independently, the "Solute_trans_a" Pfam domain was found in all four proteins with high scores and significance, but not in other human proteins. Therefore, it is likely that the four proteins, TMEM184A-C and SLC51A, are homologous. In spite of this, sequence identity between TMEM184 proteins and SLC51A is low (12.3%-13.6%), but moderate among TMEM184 proteins (26.5%-62.0%). All four proteins are predicted to harbor 7 TM helices according to UniProt, yet our search has found no similar proteins with known structures.

TMEM184A was identified as a heparin receptor in vascular cells [123], but no transport activity has been reported. Interestingly, while SLC51A is known to function as a bile acid transporter [124–126], TMEM184B has been proposed to be responsible for ibuprofen uptake [127]. This is interesting in view of the partial

chemical similarity between steroid acids and ibuprofen, both harboring a hydrophobic hydrocarbon part and a carboxyl moiety. TMEM184C resides in a locus that has been suggested to be responsible for the pathogenicity of X-linked congenital hypertrichosis syndrome [128], but no transport activity has been suggested.

Putative transporters

Our search also identified proteins whose transport activity is either controversial or not characterized, and which do not show sequence similarity to transporters of known function. Thus, the proteins in these families require further investigation to uncover their function.

The **CNNM1-4** proteins (also called ACDP1-4) are distant homologs of the cyclins, but have no documented enzymatic activity. Instead, CNNM proteins belong to a highly conserved family of Mg^{2+} transport-related proteins [129], and CNNM2 and CNNM4 have been proposed to be the long sought after basolateral Na^{+}/Mg^{2+} exchangers in the kidney and intestine, respectively [130,131]. The function of these proteins is, however, controversial [132–135], and there are hypotheses that CNNM proteins *per se* are not Mg^{2+} transporters [136]. Most recently, however, the structure of a bacterial homolog, CorC, has been resolved, revealing its membrane topology, as well as a conserved Mg^{2+}-binding site [137]. Strikingly, the Mg^{2+} ions in the structure are fully dehydrated, in contrast to those in other known Mg^{2+} channel structures [137], which makes it unlikely that the proteins function via a channel-like transport mechanism. In line with this, the authors suggest an alternating-access exchange mechanism [137], however, further studies are required to understand how and whether CNNM proteins might be able to mediate the translocation of Mg^{2+} ions across the membrane.

A family of 4 lysosomal-associated transmembrane proteins (**LAPTM4A**, **LAPTM4B**, **LAPTM5** and sequence **B4E0C1**) turned up in our search, corresponding to the TCDB subfamily #2.A.74.1 and Pfam domain “Mtp” (mouse transporter protein). The family also includes an uncharacterized transcript with the UniProt accession “B4E0C1”. Originally, the mouse transporter protein (Mtp, ortholog of LAPTM4A) was characterized as a transporter mediating the transport of nucleosides and nucleobases between the

cytoplasm and intracellular compartments [138], and was later also associated with multidrug-resistance (MDR) in yeast, where its expression changed the subcellular compartmentalization of a heterogenous group of compounds [139,140]. LAPTM4A was shown to be involved in glycosylation and glycolipid regulation [141,142]. All three proteins seem to be lysosomal [138,143,144]. Nevertheless, these proteins appear to interact with other characterized transporters, such as SLC22A2 (hOCT2), SLC7A5/SLC3A2 (LAT1/4F2hc) and MDR-related ABC transporters [145–147]. But it has been claimed that they are not *per se* transporters, but rather regulatory factors, either assisting the localization and targeting or the function of other transporters [147]. Interestingly, the transcript "B4E0C1" appears to have 4 TMHs at its N-terminus, which is identical to human LAPTM5 apart from a ~40-amino acid insertion between TMH3 and TMH4. This region also shows significant similarity to both the "Mtp" Pfam domain and the TC# 2.A.74.1 subfamily. However, the C-terminal region of the transcript is identical to the C-terminal segment of "actin filament-associated protein 1-like 1" protein (UniProt accession Q8TED9). We did not find any similar fusion sequences in the other organisms we analyzed. The "Mtp" Pfam domain, which is the hallmark of the family, belongs to the "Tetraspannin" Pfam clan, which has no other domains with annotated transporter function and no similarity to existing transporters. Structural information is also not available. Therefore, the transport function of these proteins requires further investigation.

Our search identified four proteins in human (**LMBR1**, **LMBR1L**, **LMBD1/LMBRD1**, **LMBRD2**) bearing the "LMBR1" Pfam domain, which clustered into two families in our analysis. These proteins correspond to TCDB family #9.A.54. The LMBD1 protein, encoded by the LMBRD1 gene, was suggested to function as a vitamin B12 (cobalamin) transporter, exporting vitamin B12 from the lysosomes into the cytoplasm [148]. However, it was later shown that LMBD1 actually interacts with ABCD4 and assists in its lysosomal trafficking [149], and that ABCD4 transports vitamin B12 even in the absence of LMBD1 [150]. Therefore, it is likely that LMBD1 itself is not a vitamin B12 transporter. LMBD1 was originally coined as having "significant homology" to lipocalin membrane receptors [148], and indeed the LIMR (lipocalin-1-interacting membrane receptor) protein, encoded by the LMBR1L gene, is responsible for binding lipocain 1 (LCN-1) with high affinity [151,152]. LMBRD2 was proposed to be a regulator of β2-adrenoceptor signaling [153], while the first

protein identified in the family, LMBR1, was associated with polydactyly and limb malformations [154,155]. However, its physiological role is still elusive. The proteins seem to contain 9 TMHs in a 5+4 arrangement according to UniProt predictions, but the tertiary structure of the proteins is still unknown, and no homologs with a known structure were found in our search.

Two MagT1-like proteins (**MAGT1** and **TUSC3**), as well as **OSTC** and **OSTCL** (oligosaccharyltransferase complex subunit) turned up in our search, showing similarity to TCDB family #1.A.76 members. MAGT1 and TUSC3 also have high-scoring hits for the Pfam domain "OST3_OST6", which is characteristic of members of the oligosaccharyltransferase (OST) complex. TUSC3 (also called N33) was first identified as a tumor suppressor gene [156], and its presence, together with that of MAGT1 in the OST complex has been attested later on [157–161]. Therefore, it was suggested that these proteins act as oxidoreductases [160]. Meanwhile, MAGT1 and TUSC3 were also proposed to act as Mg^{2+} transporters [162,163]. On the other hand, recent structural findings of human MAGT1 [161] indicated that this protein may not function as a transporter or channel due to the lack of substrate-binding site or pore. OSTC (also called DC2) has similarly been shown to be part of the OST complex and to have a structure similar to MAGT1 [161]. Whether MagT1-like proteins still have a transport function remains to be clarified.

The **TMEM14A**, **TMEM14B** and **TMEM14C** proteins are the only ones in human containing the "Tmemb_14" Pfam domain. While Pfam lists this domain as functionally uncharacterized, a plant protein (FAX1) containing this domain was suggested to be involved in fatty acid export from chloroplasts [164]. However, the physiological roles of TMEM14A and TMEM14B in human remain elusive. The third member of the family, TMEM14C, was identified as a putative mitochondrial protein whose transcript is consistently coexpressed with proteins from the core machinery of heme biosynthesis [165]. It was later shown that TMEM14C mediates the import of protoporphyrinogen IX (PPgenIX) into the mitochondrial matrix [166,167]. While the structure of TMEM14C was solved using nuclear magnetic resonance (NMR) [168], showing a bundle of three TM helices and an amphipathic helix, the transport mechanism remains elusive. Interestingly, despite their proposed function in the mitochondria, TargetP-2.0 [169] did not predict a mitochondrial targeting sequence in the amino acid sequence of any of the human TMEM14 proteins in our hands.

TMEM205 is a 4-TMH protein according to UniProt annotations, which was linked to cisplatin resistance [170]. The protein is expressed mostly in liver, pancreas and adrenal glands, and is present on the plasma membrane [170]. TMEM205-mediated resistance was shown to be selective towards platinum-based drugs, such as cisplatin and oxaliplatin, but not carboplatin [171]. While structural information about the protein is not available, mutagenesis studies of TMEM205 showed that mutating sulfur-containing residues, especially in TMH2 and TMH4, diminishes the effect of cisplatin resistance [171]. Nevertheless, neither the biological function nor the physiological substrates of TMEM205 are known.

Transporters with hydrophobic substrates

In addition to proteins that transport solutes or are similar to transporters that typically translocate water-soluble small molecule compounds, our search has uncovered numerous proteins that have been reported to take part in modulating the intracellular distribution of hydrophobic or amphipathic compounds, such as cholesterol, fat-soluble vitamins, lipids and fatty acids (Table 3). Although these proteins do not, strictly speaking, transport so-called “solutes”, they translocate small hydrophobic molecules that have fundamental biological functions. Thus they belong to the SLC superfamily as well. Accordingly, they have been integrated in our search and some of them have already been included in the SLC nomenclature (SLC59, SLC63, SLC65). In view of the biological and pharmaceutical importance of the transport mechanisms of hydrophobic substrates, our results on this topic will be discussed in a separate paper.

Conclusions

Our study represents the first systematic correlation of the SLC and TCDB nomenclature schemes. Many of the transport proteins discovered in our search are underexplored and there is limited information about them, although they often have important physiological roles and/or potentially represent new therapeutic targets. Even with proteins that have been studied for their physiological involvement, it was often not taken into account that they could have a transport function. Numerous proteins uncovered in our search have similarity to proteins with transport function in other organisms, but their physiological substrates remain unknown. These proteins will be interesting targets for deorphanization studies, to reveal their natural

substrates. In our work, we also highlight proteins for which transport activity has been controversial, and more specific analyses are required to clarify their biological function. In addition, our search reveals new SLC-like proteins that have no structural information. This hinders a deeper understanding of their transport mechanism. Future structure determinations would be of crucial importance to accelerate validation of the identified proteins. The combination of all these efforts would greatly facilitate the completion of the SLC-ome in human cells. Thus, our study points out important directions in which future studies could help resolve the lack of information about SLC transporters, which will help unlock their therapeutic potential.

Acknowledgements

We acknowledge support by the Swiss National Science Foundation for the grants # CRSII5_180326 ("The role of mitochondrial carriers in metabolic tuning and reprogramming by calcium flow across membrane contact sites") and # 310030_182272 ("Intestinal absorption of transition metals in human health and disease") as well as by the Swiss National Research Programme NRP 78, # 4078P0_198281 ("New insights into the COVID-19 pandemic"). Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

Methods

HMM building for TCDB families and subfamilies

Sequences of selected TCDB families and subfamilies were aligned using PSI-Coffee 11.00 [172] and NCBI BLAST+ 2.6.0 [173] using the "nr" BLAST database of 2018-02-12. PSI-Coffee was run with BLAST mode "LOCAL" and otherwise with default options. Altogether, 4221 different sequences from the TCDB were present in the subfamily and family alignments, and the profiles used for alignments contained 1-1070 sequences. In total, 130 sequences from the TCDB returned zero hits from the "nr" database in the iterative BLAST searches, meaning their profiles just contained the query sequence itself. PSI-Coffee uses the profiles to guide the generation of multiple alignments of the query sequences within each subfamily and family. The alignments were subsequently turned into HMMs using the "hmmbuild" command of HMMER 3.1b2 [30] using default settings.

Sequence similarity search

All protein belonging to each of the 7 organisms studied were downloaded from UniProt on 2019-07-31 into a FASTA file. Sequence similarity search was performed using the HMMs downloaded for selected Pfam families on 2019-02-14 and those generated for selected TCDB families and subfamilies using "hmmsearch" from HMMER 3.1b2. Hits with bit scores larger than 50 were used for further analysis. These hits presented a maximal hit e-value of 7.9e-12, while hits with bit score larger than 25, used for HMM fingerprint-based clustering (see below), had hit e-values below 6e-4.

Sequence clustering for fragment elimination

Since the downloaded protein set from UniProt contained fragments as well as predicted open reading frames and sequences from genomic screening methods, we strived to retain one sequence per gene for further analysis. In order to achieve this, hits yielded by the sequence similarity search were clustered using the following method. First, all-against-all similarity searches were performed using NCBI BLAST+ 2.4.0 [173]. Sequences were assigned the same cluster if they either share common gene annotations according to their UniProt records, or a high-scoring segment pair (HSP) with more than 95% sequence identity based on the all-against-all BLAST search. For gene annotation, the fields Gene Symbol (GN), HGNC symbol, GeneID, UniGene, FlyBase, KEGG identifiers were used from UniProt records. Ambiguous or conflicting annotations, as well as annotations conflicting with BLAST-reported sequence similarity were detected and resolved manually. Afterwards, clusters were reduced to representative sequences. For clusters containing a single Swiss-Prot sequence, that sequence was taken as representative. For cluster with no Swiss-Prot sequence, the longest sequence of the cluster was taken as representative. Clusters with more than one Swiss-Prot sequence were manually analyzed and split if necessary.

HMM fingerprint-based sequence clustering

We introduce the concept of a "HMM fingerprint", which is a mathematical vector of numbers assigned to a protein sequence, corresponding to the bit scores of similarity to each of the set of HMMs used in our analysis, consisting of the HMMs of TCDB families and subfamilies, as well as Pfam HMMs. We have restricted the number of HMMs to those that gave a hit with bit score > 25, in total 513 HMMs. Two proteins

sequences that are related are expected to show similarity to a similar subset of TCDB families, subfamilies, or Pfam families, and therefore a similar pattern in their HMM fingerprints. In turn, if two protein sequences show similarity to the same subset of TCDB families, subfamilies or Pfam families, as indicated by a similar HMM fingerprint, then they can be expected to be related. Once the HMM fingerprint has been assigned to each protein sequence found in our search, the unweighted pair group method with arithmetic mean (UPGMA) method [174] was used using the cosine metric to arrive at a hierarchical clustering of the sequences. The tree representing the clustering was cut at 0.7 cosine distance to arrive at branches that formed the basis of protein families. Join and split constraints have been introduced to keep certain proteins artificially grouped or separated. If a join constraint was acting between a pair of proteins and they were not grouped into the same cluster by the default tree cut threshold, then the clustering tree was cut just above the branch that leads to the smallest cluster containing both proteins. Similarly, split constraints between two proteins caused the clustering tree to be cut just below the branch containing both proteins, unless the two proteins were already in different clusters. Join constraints were introduced between the following protein pairs: SLC5A1-SLC5A7, SLC9A1-SLC9B1, SLC10A1-SLC10A7, SLC25A1-SLC25A46, SLC30A1-SLC30A9, SLC35C1-SLC35G1, SLC39A1-SLC39A9, and SLC66A1-SLC66A5. Split constraints were introduced between the following protein pairs: SLC2A1-SLC22A1, SLC17A1-SLC18A1, SLC17A1-SLC37A1, SLC32A1-SLC36A1, and SLC32A1-SLC38A1. The clustering tree with the resulting families was shown in Fig 1 in a polar coordinate system, with the radial component ($d \in [0; 1]$) transformed according to $d' = 0.5 + \frac{\arcsin(2d-1)}{\pi}$ to magnify tree details around $d = 0$ and $d = 1$ for better visual representation.

Structural homolog search

Sequences of SLC-like proteins were turned into HMMs using “hhblits” from the HH-suite3 package [175], using the UniRef30 database of 2020-06 [35,176], 3 iterations, an E-value threshold of 1e-3 for inclusion, and a probability threshold of 0.35 for MAC re-alignment. The HMMs were searched against the “pdb70” database of 2021-08-04 [35] with no MAC realignment, “predicted vs predicted” secondary structure scoring, and amino acid score of 1. The resulting hits were checked against PDB annotations of

transmembrane helices, and were accepted if at least 3 TM segments were contained within the aligned region and the E-value of the hit was less than 1e-4.

Phylogenetic trees

Selected groups of SLC-like proteins were aligned using Clustal Omega 1.2.1 [177,178] with 5 iterations and default settings. Smart Model Selection 1.8.4 [179] and PhyML 3.3.20190909 [180] were used to generate the phylogenetic trees, with 10 random starting trees and using the approximate likelihood ratio test aLRT method [181]. Trees in main figures were visualized using TreeViewer 1.2.2 (https://treeviewer.org/). Tree rooting, rearrangement (with threshold 0.9) and reconciliation with the species tree was done using NOTUNG 2.9.1.5 [182]. Reconciled phylogenetic trees were visualized using custom Python scripts in the style used by NOTUNG. Orthologs were identified using the reconciled phylogenetic trees and custom Python scripts.

Supplementary information

S1 Table. "SLC-like" Pfam domains and their clan memberships. The dataset is a result of our manual curation based on our "SLC-like" criteria (see text) and Pfam families present in protein sequences listed in the TCDB. Family names and data are from the Pfam database.

S2 Table. False positive hits or incorrectly annotated human sequences found in our search. Gene symbol and protein name shown based on UniProt data. Most similar (highest scoring) TCDB families/subfamilies and Pfam families are shown.

S3 Table. Table of existing ("original"/classical) SLC transporters, showing correlation with TCDB families/subfamilies, Pfam families and PDB structures and structural folds. Gene symbol and protein name based on UniProt information are shown. Most similar (highest-scoring) TCDB families and subfamilies, as well as Pfam families are shown for each protein. The PDB structure from the pdb70 dataset that is most similar (highest-scoring) to each protein is shown, along with its fold family assignment by us, partly based on TCDB family names.

AbgT: p-Aminobenzoyl-glutamate Transporter family; Amt: Ammonium Transporter family; APC: Amino acid-Polyamine-Cation family; CDF: Cation Diffusion Facilitator family; CNT: Concentrative Nucleoside Transporter family; Ctr: Copper Transporter family; DAACS: Dicarboxylate/Amino Acid:Cation (Na^+ or H^+) Symporter family; MATE: Multidrug And Toxic compound Extrusion family; MCF: Mitochondrial Carrier Family; MFS: Major Facilitator Superfamily; MgtE: Mg^{2+} transporter-E family; NAT: Nucleobase/Ascorbate Transporter or Nucleobase:Cation Symporter-2 (NCS2) family; NCX: Sodium/Calcium exchanger family; NhaA: Sodium/proton antiporter family; NST: Nucleoside-Sugar Transporter family; PiT: Type III Sodium/phosphate cotransporter family; SWEET: Sugar Will Eventually be Transported family.; ZIP: Zrt/Irt-like Transporter family.

S4 Table. "Novel" SLC-like transporters of Table 3, extended with additional information. Protein sequence identifiers, most similar (highest-scoring) PDB structure from the pdb70 dataset, chemical identifiers, PubMed links to references, and comments have been included.

S1 File. Phylogenetic trees of SLC-like protein families with more than two members. Trees were generated using multiple alignment by ClustalO, maximum likelihood tree generation by PhyML, followed by tree reconciliation with the species tree using NOTUNG (see Methods). The species tree with internal names of putative ancestor taxa is shown on each page on the upper-right hand corner. The trees are shown as dendrograms and branch lengths are not indicative of evolutionary distance. Each tree leaf corresponds to an SLC-like protein sequence denoting a gene, labels show the gene symbol, UniProt accession and taxon name. Leaves with labels ending with "*LOST" denote putative genes lost in the indicated ancestral species. Red "D" denote gene duplication nodes, normal nodes correspond to speciation nodes. Light green numbers denote branch support values as calculated by NOTUNG.

References

1. Hediger MA, Romero MF, Peng J-B, Rolfs A, Takanaga H, Bruford EA. The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteinsIntroduction. Pflugers Arch. 2004;447: 465–468. doi:10.1007/s00424-003-1192-y

2. Hediger MA, Clémençon B, Burrier RE, Bruford EA. The ABCs of membrane transporters in health and disease (SLC series): introduction. Mol Aspects Med. 2013;34: 95–107. doi:10.1016/j.mam.2012.12.009
3. César-Razquin A, Snijder B, Frappier-Brinton T, Isserlin R, Gyimesi G, Bai X, et al. A Call for Systematic Research on Solute Carriers. Cell. 2015;162: 478–487. doi:10.1016/j.cell.2015.07.022
4. Bai X, Moraes TF, Reithmeier RAF. Structural biology of solute carrier (SLC) membrane transport proteins. Mol Membr Biol. 2018; 1–32. doi:10.1080/09687688.2018.1448123
5. Choi S, Jeon J, Yang J-S, Kim S. Common occurrence of internal repeat symmetry in membrane proteins. Proteins. 2008;71: 68–80. doi:10.1002/prot.21656
6. Forrest LR. Structural Symmetry in Membrane Proteins. Annu Rev Biophys. 2015;44: 311–337. doi:10.1146/annurev-biophys-051013-023008
7. Jardetzky O. Simple allosteric model for membrane pumps. Nature. 1966;211: 969–970. doi:10.1038/211969a0
8. Forrest LR, Krämer R, Ziegler C. The structural basis of secondary active transport mechanisms. Biochim Biophys Acta. 2011;1807: 167–188. doi:10.1016/j.bbabio.2010.10.014
9. Fredriksson R, Nordström KJV, Stephansson O, Hägglund MGA, Schiöth HB. The solute carrier (SLC) complement of the human genome: phylogenetic classification reveals four major families. FEBS Lett. 2008;582: 3811–3816. doi:10.1016/j.febslet.2008.10.016
10. Perland E, Fredriksson R. Classification Systems of Secondary Active Transporters. Trends Pharmacol Sci. 2017;38: 305–315. doi:10.1016/j.tips.2016.11.008
11. Perland E, Bagchi S, Klaesson A, Fredriksson R. Characteristics of 29 novel atypical solute carriers of major facilitator superfamily type: evolutionary conservation, predicted structure and neuronal co-expression. Open Biol. 2017;7. doi:10.1098/rsob.170142
12. Schlessinger A, Matsson P, Shima JE, Pieper U, Yee SW, Kelly L, et al. Comparison of human solute carriers. Protein Science. 2010;19: 412–428. doi:10.1002/pro.320
13. Schlessinger A, Yee SW, Sali A, Giacomini KM. SLC classification: an update. Clin Pharmacol Ther. 2013;94: 19–23. doi:10.1038/clpt.2013.73
14. Saier MH. Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya. Adv Microb Physiol. 1998;40: 81–136.
15. Saier MH, Reddy VS, Moreno-Hagelsieb G, Hendargo KJ, Zhang Y, Iddamsetty V, et al. The Transporter Classification Database (TCDB): 2021 update. Nucleic Acids Res. 2021;49: D461–D467. doi:10.1093/nar/gkaa1004
16. Cao Y, Jin X, Huang H, Derebe MG, Levin EJ, Kabaleeswaran V, et al. Crystal structure of a potassium ion transporter, TrkH. Nature. 2011;471: 336–340. doi:10.1038/nature09731
17. Bosson R, Jaquenoud M, Conzelmann A. GUP1 of Saccharomyces cerevisiae encodes an O-acyltransferase involved in remodeling of the GPI anchor. Mol Biol Cell. 2006;17: 2636–2645. doi:10.1091/mbc.e06-02-0104
18. Berks BC, Sargent F, Palmer T. The Tat protein export pathway. Mol Microbiol. 2000;35: 260–274. doi:10.1046/j.1365-2958.2000.01719.x
19. Berks BC, Sargent F, De Leeuw E, Hinsley AP, Stanley NR, Jack RL, et al. A novel protein transport system involved in the biogenesis of bacterial electron transfer chains. Biochim Biophys Acta. 2000;1459: 325–330. doi:10.1016/s0005-2728(00)00168-7
20. Ren Q, Kang KH, Paulsen IT. TransportDB: a relational database of cellular membrane transport systems. Nucleic Acids Res. 2004;32: D284-288. doi:10.1093/nar/gkh016
21. Ren Q, Chen K, Paulsen IT. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res. 2007;35: D274-279. doi:10.1093/nar/gkl925

22. Elbourne LDH, Tetu SG, Hassan KA, Paulsen IT. TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. Nucleic Acids Res. 2017;45: D320–D324. doi:10.1093/nar/gkw1068
23. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001;29: 22–28. doi:10.1093/nar/29.1.22
24. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res. 2013;41: D387-395. doi:10.1093/nar/gks1234
25. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44: D279-285. doi:10.1093/nar/gkv1344
26. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305: 567–580. doi:10.1006/jmbi.2000.4315
27. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47: D506–D515. doi:10.1093/nar/gky1049
28. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432. doi:10.1093/nar/gky995
29. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, et al. Pfam: clans, web tools and services. Nucleic Acids Res. 2006;34: D247-251. doi:10.1093/nar/gkj149
30. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195. doi:10.1371/journal.pcbi.1002195
31. Bateman A, Finn RD. SCOOP: a simple method for identification of novel protein superfamily relationships. Bioinformatics. 2007;23: 809–814. doi:10.1093/bioinformatics/btm034
32. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol. 2007;406: 89–112. doi:10.1007/978-1-59745-535-0_4
33. Quednau BD, Nicoll DA, Philipson KD. The sodium/calcium exchanger family-SLC8. Pflugers Arch. 2004;447: 543–548. doi:10.1007/s00424-003-1065-4
34. Chen H, Liu Y, Li X. Structure of human Dispatched-1 provides insights into Hedgehog ligand biogenesis. Life Sci Alliance. 2020;3: e202000776. doi:10.26508/lsa.202000776
35. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9: 173–175. doi:10.1038/nmeth.1818
36. Söding J, Remmert M. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr Opin Struct Biol. 2011;21: 404–411. doi:10.1016/j.sbi.2011.03.005
37. Jeong J, Eide DJ. The SLC39 family of zinc transporters. Mol Aspects Med. 2013;34: 612–619. doi:10.1016/j.mam.2012.05.011
38. Donowitz M, Ming Tse C, Fuster D. SLC9/NHE gene family, a plasma membrane and organellar family of Na^+/H^+ exchangers. Mol Aspects Med. 2013;34: 236–251. doi:10.1016/j.mam.2012.05.001
39. Pramod AB, Foster J, Carvelli L, Henry LK. SLC6 transporters: structure, function, regulation, disease association and therapeutics. Mol Aspects Med. 2013;34: 197–219. doi:10.1016/j.mam.2012.07.002
40. Song Z. Roles of the nucleotide sugar transporters (SLC35 family) in health and disease. Mol Aspects Med. 2013;34: 590–600. doi:10.1016/j.mam.2012.12.004

41. Chu X, Bleasby K, Evers R. Species differences in drug transporters and implications for translating preclinical findings to humans. Expert Opin Drug Metab Toxicol. 2013;9: 237–252. doi:10.1517/17425255.2013.741589
42. Wright EM. Glucose transport families SLC5 and SLC50. Mol Aspects Med. 2013;34: 183–196. doi:10.1016/j.mam.2012.11.002
43. Gyimesi G, Pujol-Giménez J, Kanai Y, Hediger MA. Sodium-coupled glucose transport, the SLC5 family, and therapeutically relevant inhibitors: from molecular discovery to clinical application. Pflugers Arch. 2020;472: 1177–1206. doi:10.1007/s00424-020-02433-x
44. Haga T. Molecular properties of the high-affinity choline transporter CHT1. J Biochem. 2014;156: 181–194. doi:10.1093/jb/mvu047
45. Claro da Silva T, Polli JE, Swaan PW. The solute carrier family 10 (SLC10): beyond bile acid transport. Mol Aspects Med. 2013;34: 252–269. doi:10.1016/j.mam.2012.07.004
46. Godoy JR, Fernandes C, Döring B, Beuerlein K, Petzinger E, Geyer J. Molecular and phylogenetic characterization of a novel putative membrane transporter (SLC10A7), conserved in vertebrates and bacteria. Eur J Cell Biol. 2007;86: 445–460. doi:10.1016/j.ejcb.2007.06.001
47. Karakus E, Wannowius M, Müller SF, Leiting S, Leidolf R, Noppes S, et al. The orphan solute carrier SLC10A7 is a novel negative regulator of intracellular calcium signaling. Sci Rep. 2020;10: 7248. doi:10.1038/s41598-020-64006-3
48. Janer A, Prudent J, Paupe V, Fahiminiya S, Majewski J, Sgarioto N, et al. SLC25A46 is required for mitochondrial lipid homeostasis and cristae maintenance and is responsible for Leigh syndrome. EMBO Mol Med. 2016;8: 1019–1038. doi:10.15252/emmm.201506159
49. Sesaki H, Jensen RE. UGO1 encodes an outer membrane protein required for mitochondrial fusion. J Cell Biol. 2001;152: 1123–1134. doi:10.1083/jcb.152.6.1123
50. Lamarca V, Sanz-Clemente A, Pérez-Pé R, Martínez-Lorenzo MJ, Halaihel N, Muniesa P, et al. Two isoforms of PSAP/MTCH1 share two proapoptotic domains and multiple internal signals for import into the mitochondrial outer membrane. Am J Physiol, Cell Physiol. 2007;293: C1347-1361. doi:10.1152/ajpcell.00431.2006
51. Labbé K, Mookerjee S, Le Vasseur M, Gibbs E, Lerner C, Nunnari J. The modified mitochondrial outer membrane carrier MTCH2 links mitochondrial fusion to lipogenesis. J Cell Biol. 2021;220: e202103122. doi:10.1083/jcb.202103122
52. Ruprecht JJ, Kunji ERS. The SLC25 Mitochondrial Carrier Family: Structure and Mechanism. Trends Biochem Sci. 2020;45: 244–258. doi:10.1016/j.tibs.2019.11.001
53. Chen Y-H, Kim JH, Stallcup MR. GAC63, a GRIP1-dependent nuclear receptor coactivator. Mol Cell Biol. 2005;25: 5965–5972. doi:10.1128/MCB.25.14.5965-5972.2005
54. Thomas P, Pang Y, Dong J, Berg AH. Identification and characterization of membrane androgen receptors in the ZIP9 zinc transporter subfamily: II. Role of human ZIP9 in testosterone-induced prostate and breast cancer cell apoptosis. Endocrinology. 2014;155: 4250–4265. doi:10.1210/en.2014-1201
55. Brett CL, Donowitz M, Rao R. Evolutionary origins of eukaryotic sodium/proton exchangers. Am J Physiol Cell Physiol. 2005;288: C223-239. doi:10.1152/ajpcell.00360.2004
56. Gyimesi G, Hediger MA. Sequence Features of Mitochondrial Transporter Protein Families. Biomolecules. 2020;10. doi:10.3390/biom10121611
57. Horiba N, Masuda S, Takeuchi A, Takeuchi D, Okuda M, Inui K. Cloning and characterization of a novel Na+-dependent glucose transporter (NaGLT1) in rat kidney. J Biol Chem. 2003;278: 14669–14676. doi:10.1074/jbc.M212240200

58. Nawata CM, Dantzler WH, Pannabecker TL. Alternative channels for urea in the inner medulla of the rat kidney. Am J Physiol Renal Physiol. 2015;309: F916-924. doi:10.1152/ajprenal.00392.2015
59. Tejada-Jiménez M, Galván A, Fernández E. Algae and humans share a molybdate transporter. Proc Natl Acad Sci USA. 2011;108: 6420–6425. doi:10.1073/pnas.1100700108
60. Zhu W, Spiga L, Winter S. Transition metals and host-microbe interactions in the inflamed intestine. Biometals. 2019;32: 369–384. doi:10.1007/s10534-019-00182-8
61. Barth J, Zimmermann H, Volknandt W. SV31 is a Zn2+-binding synaptic vesicle protein. J Neurochem. 2011;118: 558–570. doi:10.1111/j.1471-4159.2011.07344.x
62. Waberer L, Henrich E, Peetz O, Morgner N, Dötsch V, Bernhard F, et al. The synaptic vesicle protein SV31 assembles into a dimer and transports Zn2. J Neurochem. 2017;140: 280–293. doi:10.1111/jnc.13886
63. Cuajungco MP, Kiselyov K. The mucolipin-1 (TRPML1) ion channel, transmembrane-163 (TMEM163) protein, and lysosomal zinc handling. Front Biosci (Landmark Ed). 2017;22: 1330–1343.
64. Sanchez VB, Ali S, Escobar A, Cuajungco MP. Transmembrane 163 (TMEM163) protein effluxes zinc. Arch Biochem Biophys. 2019;677: 108166. doi:10.1016/j.abb.2019.108166
65. Burré J, Zimmermann H, Volknandt W. Identification and characterization of SV31, a novel synaptic vesicle membrane protein and potential transporter. J Neurochem. 2007;103: 276–287. doi:10.1111/j.1471-4159.2007.04758.x
66. Wang L, Li N-N, Lu Z-J, Li J-Y, Peng J-X, Duan L-R, et al. Association of three candidate genetic variants in ACMSD/TMEM163, GPNMB and BCKDK /STX1B with sporadic Parkinson's disease in Han Chinese. Neurosci Lett. 2019;703: 45–48. doi:10.1016/j.neulet.2019.03.019
67. Chang K-H, Chen C-M, Chen Y-C, Fung H-C, Wu Y-R. Polymorphisms of ACMSD-TMEM163, MCCC1, and BCKDK-STX1B Are Not Associated with Parkinson's Disease in Taiwan. Parkinsons Dis. 2019;2019: 3489638. doi:10.1155/2019/3489638
68. Lauterbach EC. Psychotropic drug effects on gene transcriptomics relevant to Parkinson's disease. Prog Neuropsychopharmacol Biol Psychiatry. 2012;38: 107–115. doi:10.1016/j.pnpbp.2012.03.011
69. Chakraborty S, Vellarikkal SK, Sivasubbu S, Roy SS, Tandon N, Bharadwaj D. Role of Tmem163 in zinc-regulated insulin storage of MIN6 cells: Functional exploration of an Indian type 2 diabetes GWAS associated gene. Biochem Biophys Res Commun. 2019. doi:10.1016/j.bbrc.2019.11.117
70. Tabassum R, Chauhan G, Dwivedi OP, Mahajan A, Jaiswal A, Kaur I, et al. Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. Diabetes. 2013;62: 977–986. doi:10.2337/db12-0406
71. Bai H, Liu H, Suyalatu S, Guo X, Chu S, Chen Y, et al. Association Analysis of Genetic Variants with Type 2 Diabetes in a Mongolian Population in China. J Diabetes Res. 2015;2015: 613236. doi:10.1155/2015/613236
72. Foulquier F, Amyere M, Jaeken J, Zeevaert R, Schollen E, Race V, et al. TMEM165 deficiency causes a congenital disorder of glycosylation. Am J Hum Genet. 2012;91: 15–26. doi:10.1016/j.ajhg.2012.05.002
73. Demaegd D, Foulquier F, Colinet A-S, Gremillon L, Legrand D, Mariot P, et al. Newly characterized Golgi-localized family of proteins is involved in calcium and pH homeostasis in yeast and human cells. Proc Natl Acad Sci USA. 2013;110: 6859–6864. doi:10.1073/pnas.1219871110
74. Foulquier F, Legrand D. Biometals and glycosylation in humans: Congenital disorders of glycosylation shed lights into the crucial role of Golgi manganese homeostasis. Biochim Biophys Acta Gen Subj. 2020;1864: 129674. doi:10.1016/j.bbagen.2020.129674

75. Stribny J, Thines L, Deschamps A, Goffin P, Morsomme P. The human Golgi protein TMEM165 transports calcium and manganese in yeast and bacterial cells. J Biol Chem. 2020;295: 3865–3874. doi:10.1074/jbc.RA119.012249
76. Snyder NA, Palmer MV, Reinhardt TA, Cunningham KW. Milk biosynthesis requires the Golgi cation exchanger TMEM165. J Biol Chem. 2019;294: 3181–3191. doi:10.1074/jbc.RA118.006270
77. Reinhardt TA, Lippolis JD, Sacco RE. The Ca(2+)/H(+) antiporter TMEM165 expression, localization in the developing, lactating and involuting mammary gland parallels the secretory pathway Ca(2+) ATPase (SPCA1). Biochem Biophys Res Commun. 2014;445: 417–421. doi:10.1016/j.bbrc.2014.02.020
78. Lebredonchel E, Houdou M, Potelle S, de Bettignies G, Schulz C, Krzewinski Recchi M-A, et al. Dissection of TMEM165 function in Golgi glycosylation and its Mn2+ sensitivity. Biochimie. 2019;165: 123–130. doi:10.1016/j.biochi.2019.07.016
79. Perland E, Lekholm E, Eriksson MM, Bagchi S, Arapi V, Fredriksson R. The Putative SLC Transporters Mfsd5 and Mfsd11 Are Abundantly Expressed in the Mouse Brain and Have a Potential Role in Energy Homeostasis. PLoS ONE. 2016;11: e0156912. doi:10.1371/journal.pone.0156912
80. Lekholm E, Perland E, Eriksson MM, Hellsten SV, Lindberg FA, Rostami J, et al. Putative Membrane-Bound Transporters MFSD14A and MFSD14B Are Neuronal and Affected by Nutrient Availability. Front Mol Neurosci. 2017;10: 11. doi:10.3389/fnmol.2017.00011
81. Perland E, Hellsten SV, Lekholm E, Eriksson MM, Arapi V, Fredriksson R. The Novel Membrane-Bound Proteins MFSD1 and MFSD3 are Putative SLC Transporters Affected by Altered Nutrient Intake. J Mol Neurosci. 2017;61: 199–214. doi:10.1007/s12031-016-0867-8
82. Bagchi S, Perland E, Hosseini K, Lundgren J, Al-Walai N, Kheder S, et al. Probable role for major facilitator superfamily domain containing 6 (MFSD6) in the brain during variable energy consumption. Int J Neurosci. 2020;130: 476–489. doi:10.1080/00207454.2019.1694020
83. Valoskova K, Biebl J, Roblek M, Emtenani S, Gyoergy A, Misova M, et al. A conserved major facilitator superfamily member orchestrates a subset of O-glycosylation to aid macrophage tissue invasion. Elife. 2019;8. doi:10.7554/eLife.41801
84. Landis GN, Bhole D, Tower J. A search for doxycycline-dependent mutations that increase Drosophila melanogaster life span identifies the VhaSFD, Sugar baby, filamin, fwd and Cctl genes. Genome Biol. 2003;4: R8. doi:10.1186/gb-2003-4-2-r8
85. Kim KW, Tang NH, Piggott CA, Andrusiak MG, Park S, Zhu M, et al. Expanded genetic screening in Caenorhabditis elegans identifies new regulators and an inhibitory role for NAD+ in axon regeneration. Elife. 2018;7: e39756. doi:10.7554/eLife.39756
86. Ceder MM, Aggarwal T, Hosseini K, Maturi V, Patil S, Perland E, et al. CG4928 Is Vital for Renal Function in Fruit Flies and Membrane Potential in Cells: A First In-Depth Characterization of the Putative Solute Carrier UNC93A. Front Cell Dev Biol. 2020;8: 580291. doi:10.3389/fcell.2020.580291
87. Hannes F, Hammond P, Quarrell O, Fryns J-P, Devriendt K, Vermeesch JR. A microdeletion proximal of the critical deletion region is associated with mild Wolf-Hirschhorn syndrome. Am J Med Genet A. 2012;158A: 996–1004. doi:10.1002/ajmg.a.35299
88. Damme M, Brandenstein L, Fehr S, Jankowiak W, Bartsch U, Schweizer M, et al. Gene disruption of Mfsd8 in mice provides the first animal model for CLN7 disease. Neurobiol Dis. 2014;65: 12–24. doi:10.1016/j.nbd.2014.01.003
89. Siintola E, Topcu M, Aula N, Lohi H, Minassian BA, Paterson AD, et al. The novel neuronal ceroid lipofuscinosis gene MFSD8 encodes a putative lysosomal transporter. Am J Hum Genet. 2007;81: 136–146. doi:10.1086/518902

90. von Kleist L, Ariunbat K, Braren I, Stauber T, Storch S, Danyukova T. A newly generated neuronal cell model of CLN7 disease reveals aberrant lysosome motility and impaired cell survival. Mol Genet Metab. 2019;126: 196–205. doi:10.1016/j.ymgme.2018.09.009
91. Li Y, Yang X, Yang J, Wang H, Wei W. An 11-gene-based prognostic signature for uveal melanoma metastasis based on gene expression and DNA methylation profile. J Cell Biochem. 2018. doi:10.1002/jcb.28151
92. Nicoletti CF, Pinhel MS, Noronha NY, Jácome A, Crujeiras AB, Nonino CB. Association of MFSD3 promoter methylation level and weight regain after gastric bypass: Assessment for 3 y after surgery. Nutrition. 2020;70: 110499. doi:10.1016/j.nut.2019.04.010
93. Västermark Å, Almén MS, Simmen MW, Fredriksson R, Schiöth HB. Functional specialization in nucleotide sugar transporters occurred through differentiation of the gene cluster EamA (DUF6) before the radiation of Viridiplantae. BMC Evol Biol. 2011;11: 123. doi:10.1186/1471-2148-11-123
94. Do DN, Schenkel FS, Miglior F, Zhao X, Ibeagha-Awemu EM. Genome wide association study identifies novel potential candidate genes for bovine milk cholesterol content. Sci Rep. 2018;8: 13239. doi:10.1038/s41598-018-31427-0
95. Prentice LM, d'Anglemont de Tassigny X, McKinney S, Ruiz de Algara T, Yap D, Turashvili G, et al. The testosterone-dependent and independent transcriptional networks in the hypothalamus of Gpr54 and Kiss1 knockout male mice are not fully equivalent. BMC Genomics. 2011;12: 209. doi:10.1186/1471-2164-12-209
96. Rodriguez PQ, Oddsson A, Ebarasi L, He B, Hultenby K, Wernerson A, et al. Knockdown of Tmem234 in zebrafish results in proteinuria. Am J Physiol Renal Physiol. 2015;309: F955-966. doi:10.1152/ajprenal.00525.2014
97. Dean N, Zhang YB, Poster JB. The VRG4 gene is required for GDP-mannose transport into the lumen of the Golgi in the yeast, Saccharomyces cerevisiae. J Biol Chem. 1997;272: 31908–31914. doi:10.1074/jbc.272.50.31908
98. Baldwin TC, Handford MG, Yuseff MI, Orellana A, Dupree P. Identification and characterization of GONST1, a golgi-localized GDP-mannose transporter in Arabidopsis. Plant Cell. 2001;13: 2283–2295. doi:10.1105/tpc.010247
99. Rodríguez A, Gonzalez L, Ko A, Alvarez M, Miao Z, Bhagat Y, et al. Molecular Characterization of the Lipid Genome-Wide Association Study Signal on Chromosome 18q11.2 Implicates HNF4A-Mediated Regulation of the TMEM241 Gene. Arterioscler Thromb Vasc Biol. 2016;36: 1350–1355. doi:10.1161/ATVBAHA.116.307182
100. Trifonov S, Houtani T, Shimizu J-I, Hamada S, Kase M, Maruyama M, et al. GPR155: Gene organization, multiple mRNA splice variants and expression in mouse central nervous system. Biochem Biophys Res Commun. 2010;398: 19–25. doi:10.1016/j.bbrc.2010.05.162
101. Wang XC, Liu Z, Jin LH. Anchor negatively regulates BMP signalling to control Drosophila wing development. Eur J Cell Biol. 2018;97: 308–317. doi:10.1016/j.ejcb.2018.04.007
102. Shimizu D, Kanda M, Tanaka H, Kobayashi D, Tanaka C, Hayashi M, et al. GPR155 Serves as a Predictive Biomarker for Hematogenous Metastasis in Patients with Gastric Cancer. Sci Rep. 2017;7: 42089. doi:10.1038/srep42089
103. Umeda S, Kanda M, Sugimoto H, Tanaka H, Hayashi M, Yamada S, et al. Downregulation of GPR155 as a prognostic factor after curative resection of hepatocellular carcinoma. BMC Cancer. 2017;17: 610. doi:10.1186/s12885-017-3629-2
104. Bräuer P, Parker JL, Gerondopoulos A, Zimmermann I, Seeger MA, Barr FA, et al. Structural basis for pH-dependent retrieval of ER proteins from the Golgi by the KDEL receptor. Science. 2019;363: 1103–1107. doi:10.1126/science.aaw2859

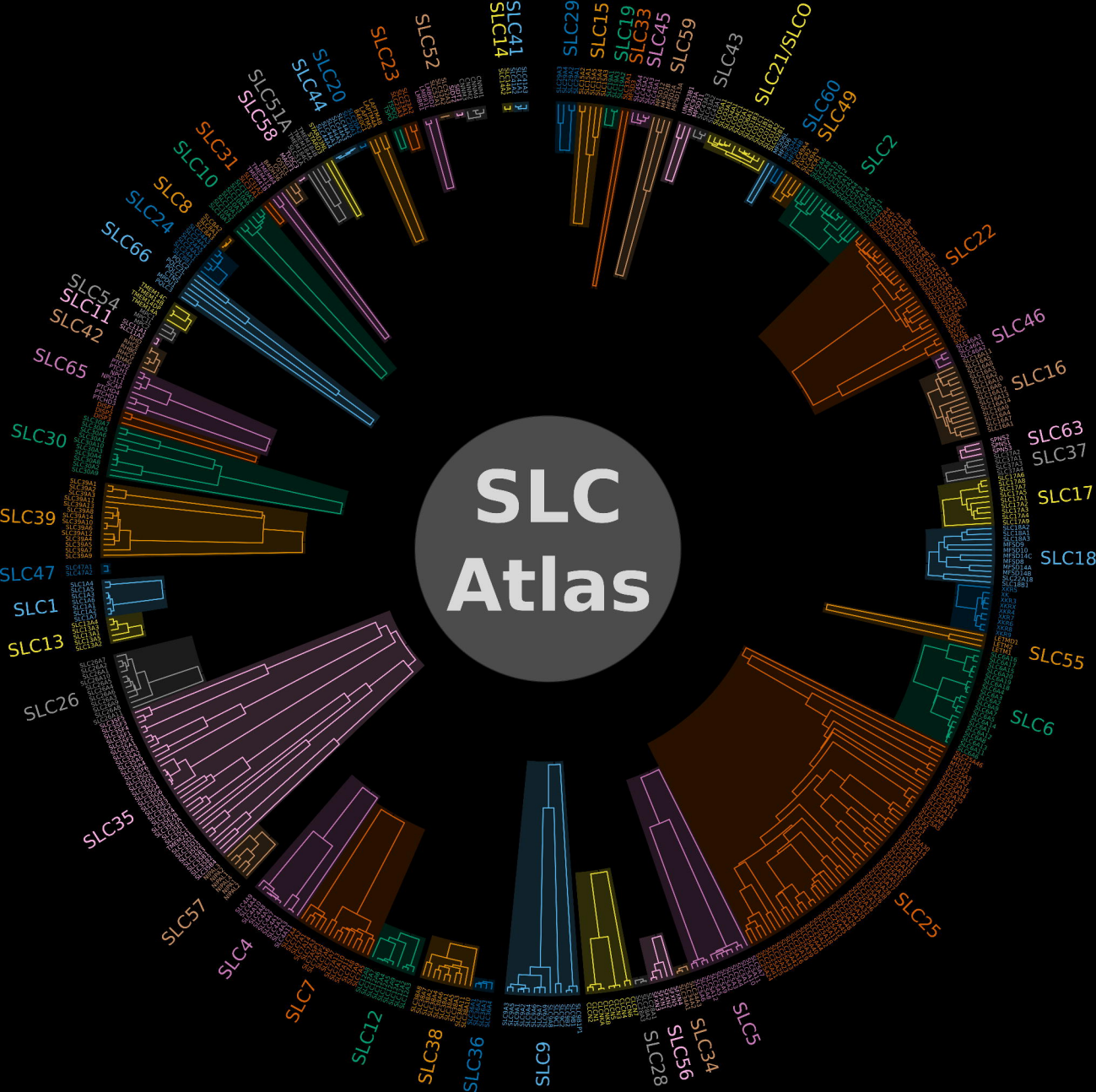
105. Saudek V. Cystinosin, MPDU1, SWEETs and KDELR belong to a well-defined protein family with putative function of cargo receptors involved in vesicle trafficking. PLoS ONE. 2012;7: e30876. doi:10.1371/journal.pone.0030876
106. Helenius J, Ng DTW, Marolda CL, Walter P, Valvano MA, Aebi M. Translocation of lipid-linked oligosaccharides across the ER membrane requires Rft1 protein. Nature. 2002;415: 447–450. doi:10.1038/415447a
107. Frank CG, Sanyal S, Rush JS, Waechter CJ, Menon AK. Does Rft1 flip an N-glycan lipid precursor? Nature. 2008;454: E3-4; discussion E4-5. doi:10.1038/nature07165
108. Gottier P, Gonzalez-Salgado A, Menon AK, Liu Y-C, Acosta-Serrano A, Bütikofer P. RFT1 Protein Affects Glycosylphosphatidylinositol (GPI) Anchor Glycosylation. J Biol Chem. 2017;292: 1103–1111. doi:10.1074/jbc.M116.758367
109. Verchère A, Cowton A, Jenni A, Rauch M, Häner R, Graumann J, et al. Complexity of the eukaryotic dolichol-linked oligosaccharide scramblase suggested by activity correlation profiling mass spectrometry. Sci Rep. 2021;11: 1411. doi:10.1038/s41598-020-80956-0
110. Nobre LS, Al-Shahrour F, Dopazo J, Saraiva LM. Exploring the antimicrobial action of a carbon monoxide-releasing compound through whole-genome transcription profiling of Escherichia coli. Microbiology (Reading). 2009;155: 813–824. doi:10.1099/mic.0.023911-0
111. Herzberg M, Kaye IK, Peti W, Wood TK. YdgG (TqsA) controls biofilm formation in Escherichia coli K-12 through autoinducer 2 transport. J Bacteriol. 2006;188: 587–598. doi:10.1128/JB.188.2.587-598.2006
112. Dong P, Wang L, Song N, Yang L, Chen J, Yan M, et al. A UPF0118 family protein with uncharacterized function from the moderate halophile Halobacillus andaensis represents a novel class of Na+(Li+)/H+ antiporter. Sci Rep. 2017;7: 45936. doi:10.1038/srep45936
113. Wang L, Zou Q, Yan M, Wang Y, Guo S, Zhang R, et al. Polar or Charged Residues Located in Four Highly Conserved Motifs Play a Vital Role in the Function or pH Response of a UPF0118 Family Na+(Li+)/H+ Antiporter. Front Microbiol. 2020;11: 841. doi:10.3389/fmicb.2020.00841
114. Shao L, Xu T, Zheng X, Shao D, Zhang H, Chen H, et al. A novel three-TMH Na+/H+ antiporter and the functional role of its oligomerization. J Mol Biol. 2021;433: 166730. doi:10.1016/j.jmb.2020.166730
115. Mesdaghi S, Murphy DL, Sánchez Rodríguez F, Burgos-Mármol JJ, Rigden DJ. In silico prediction of structure and function for a large family of transmembrane proteins that includes human Tmem41b. F1000Res. 2020;9: 1395. doi:10.12688/f1000research.27676.2
116. Morita K, Hama Y, Izume T, Tamura N, Ueno T, Yamashita Y, et al. Genome-wide CRISPR screen identifies TMEM41B as a gene required for autophagosome formation. J Cell Biol. 2018;217: 3817–3828. doi:10.1083/jcb.201804132
117. Moretti F, Bergman P, Dodgson S, Marcellin D, Claerr I, Goodwin JM, et al. TMEM41B is a novel regulator of autophagy and lipid mobilization. EMBO Rep. 2018;19. doi:10.15252/embr.201845889
118. Van Alstyne M, Lotti F, Dal Mas A, Area-Gomez E, Pellizzoni L. Stasimon/Tmem41b localizes to mitochondria-associated ER membranes and is essential for mouse embryonic development. Biochem Biophys Res Commun. 2018;506: 463–470. doi:10.1016/j.bbrc.2018.10.073
119. Morita K, Hama Y, Mizushima N. TMEM41B functions with VMP1 in autophagosome formation. Autophagy. 2019;15: 922–923. doi:10.1080/15548627.2019.1582952
120. Shoemaker CJ, Huang TQ, Weir NR, Polyakov NJ, Schultz SW, Denic V. CRISPR screening using an expanded toolkit of autophagy reporters identifies TMEM41B as a novel autophagy factor. PLoS Biol. 2019;17: e2007044. doi:10.1371/journal.pbio.2007044

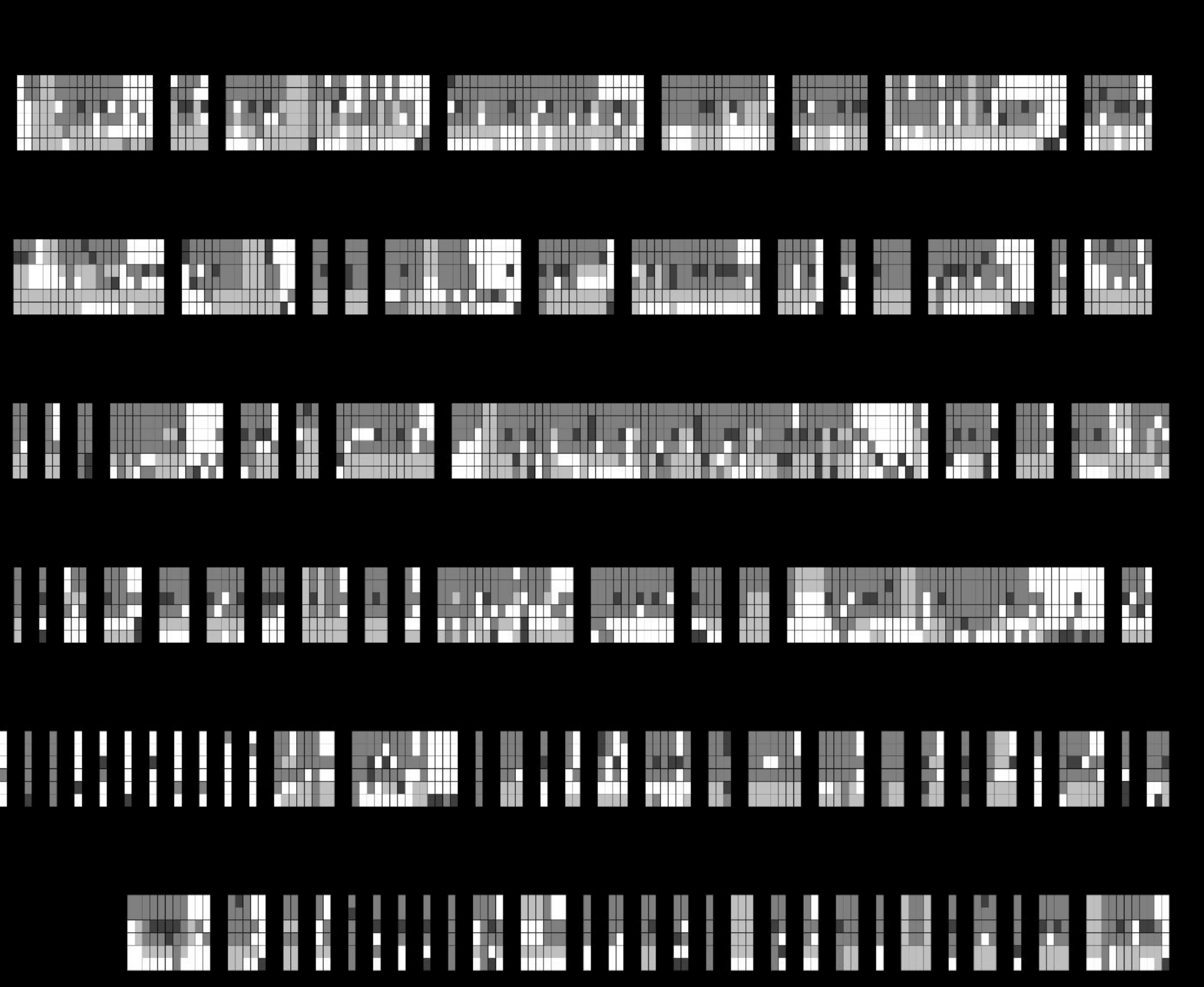
121. Schneider WM, Luna JM, Hoffmann H-H, Sánchez-Rivera FJ, Leal AA, Ashbrook AW, et al. Genome-scale identification of SARS-CoV-2 and pan-coronavirus host factor networks. bioRxiv. 2020; 2020.10.07.326462. doi:10.1101/2020.10.07.326462
122. Hoffmann H-H, Schneider WM, Rozen-Gagnon K, Miles LA, Schuster F, Razooky B, et al. TMEM41B Is a Pan-flavivirus Host Factor. Cell. 2021;184: 133-148.e20. doi:10.1016/j.cell.2020.12.005
123. Farwell SLN, Kanyi D, Hamel M, Slee JB, Miller EA, Cipolle MD, et al. Heparin Decreases in Tumor Necrosis Factor α (TNFα)-induced Endothelial Stress Responses Require Transmembrane Protein 184A and Induction of Dual Specificity Phosphatase 1. J Biol Chem. 2016;291: 5342–5354. doi:10.1074/jbc.M115.681288
124. Dawson PA, Hubbert M, Haywood J, Craddock AL, Zerangue N, Christian WV, et al. The heteromeric organic solute transporter alpha-beta, Ostalpha-Ostbeta, is an ileal basolateral bile acid transporter. J Biol Chem. 2005;280: 6960–6968. doi:10.1074/jbc.M412752200
125. Seward DJ, Koh AS, Boyer JL, Ballatori N. Functional complementation between a novel mammalian polygenic transport complex and an evolutionarily ancient organic solute transporter, OSTalpha-OSTbeta. J Biol Chem. 2003;278: 27473–27482. doi:10.1074/jbc.M301106200
126. Wang W, Seward DJ, Li L, Boyer JL, Ballatori N. Expression cloning of two genes that together mediate organic solute and steroid transport in the liver of a marine vertebrate. Proc Natl Acad Sci USA. 2001;98: 9431–9436. doi:10.1073/pnas.161099898
127. Rasmussen RN, Christensen KV, Holm R, Nielsen CU. Nfat5 is involved in the hyperosmotic regulation of Tmem184b: a putative modulator of ibuprofen transport in renal MDCK I cells. FEBS Open Bio. 2019;9: 1071–1081. doi:10.1002/2211-5463.12630
128. Zhu H, Shang D, Sun M, Choi S, Liu Q, Hao J, et al. X-linked congenital hypertrichosis syndrome is associated with interchromosomal insertions mediated by a human-specific palindrome near SOX3. Am J Hum Genet. 2011;88: 819–826. doi:10.1016/j.ajhg.2011.05.004
129. Quamme GA. Molecular identification of ancient and modern mammalian magnesium transporters. Am J Physiol, Cell Physiol. 2010;298: C407-429. doi:10.1152/ajpcell.00124.2009
130. Stuiver M, Lainez S, Will C, Terryn S, Günzel D, Debaix H, et al. CNNM2, encoding a basolateral protein required for renal Mg2+ handling, is mutated in dominant hypomagnesemia. Am J Hum Genet. 2011;88: 333–343. doi:10.1016/j.ajhg.2011.02.005
131. Yamazaki D, Funato Y, Miura J, Sato S, Toyosawa S, Furutani K, et al. Basolateral Mg2+ extrusion via CNNM4 mediates transcellular Mg2+ transport across epithelia: a mouse model. PLoS Genet. 2013;9: e1003983. doi:10.1371/journal.pgen.1003983
132. Funato Y, Furutani K, Kurachi Y, Miki H. CrossTalk proposal: CNNM proteins are Na+ /Mg2+ exchangers playing a central role in transepithelial Mg2+ (re)absorption. J Physiol. 2018;596: 743–746. doi:10.1113/JP275248
133. Arjona FJ, de Baaij JHF. CrossTalk opposing view: CNNM proteins are not Na+ /Mg2+ exchangers but Mg2+ transport regulators playing a central role in transepithelial Mg2+ (re)absorption. J Physiol (Lond). 2018;596: 747–750. doi:10.1113/JP275249
134. Funato Y, Furutani K, Kurachi Y, Miki H. Rebuttal from Yosuke Funato, Kazuharu Furutani, Yoshihisa Kurachi and Hiroaki Miki. J Physiol. 2018;596: 751. doi:10.1113/JP275706
135. Arjona FJ, de Baaij JHF. Rebuttal from Francisco J. Arjona and Jeroen H. F. de Baaij. J Physiol. 2018;596: 753–754. doi:10.1113/JP275705
136. Sponder G, Mastrototaro L, Kurth K, Merolle L, Zhang Z, Abdulhanan N, et al. Human CNNM2 is not a Mg(2+) transporter per se. Pflugers Arch. 2016;468: 1223–1240. doi:10.1007/s00424-016-1816-7

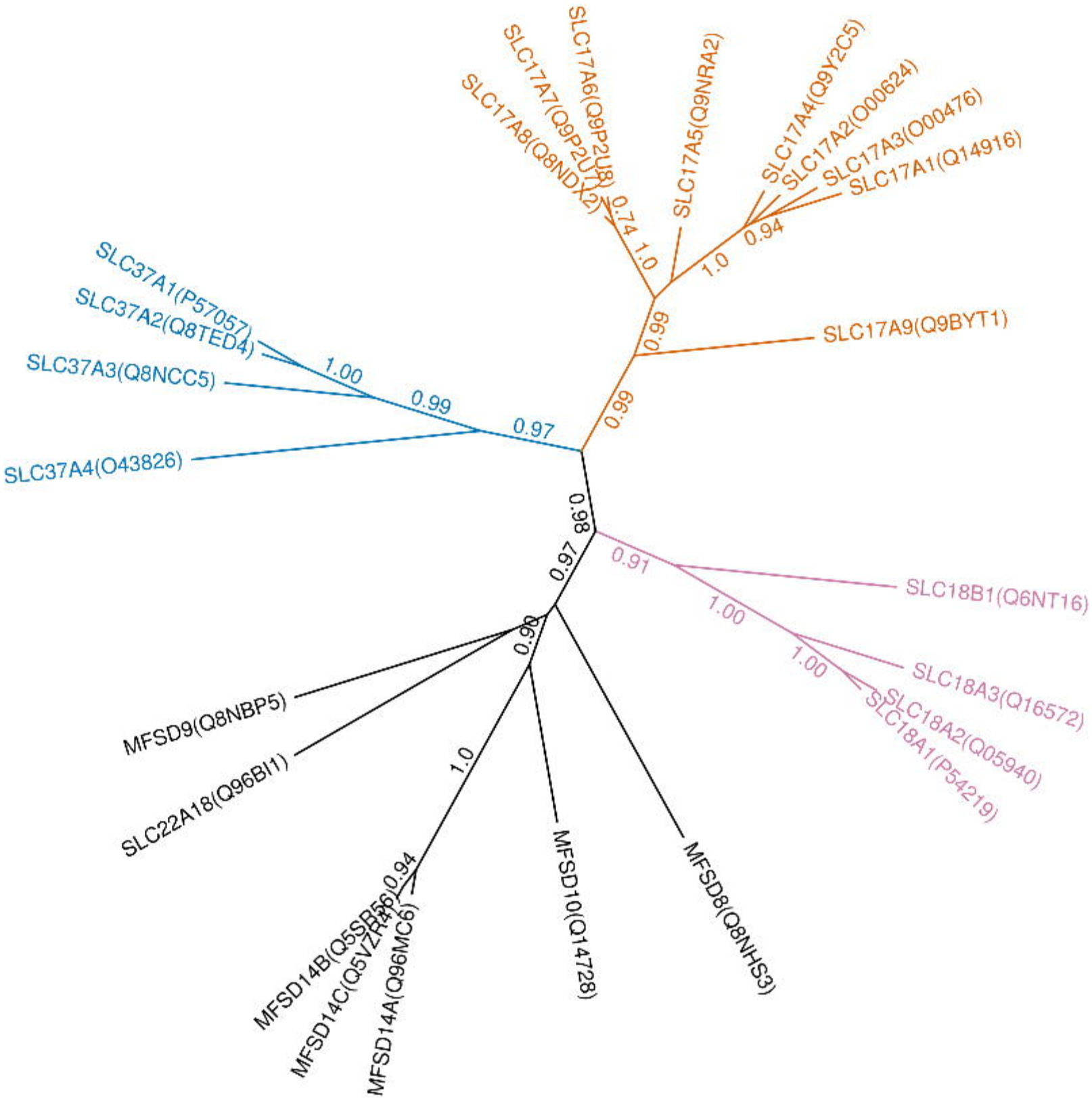
137. Huang Y, Jin F, Funato Y, Xu Z, Zhu W, Wang J, et al. Structural basis for the Mg2+ recognition and regulation of the CorC Mg2+ transporter. Sci Adv. 2021;7: eabe6140. doi:10.1126/sciadv.abe6140
138. Hogue DL, Ellison MJ, Young JD, Cass CE. Identification of a novel membrane transporter associated with intracellular membranes by phenotypic complementation in the yeast Saccharomyces cerevisiae. J Biol Chem. 1996;271: 9801–9808. doi:10.1074/jbc.271.16.9801
139. Cabrita MA, Hobman TC, Hogue DL, King KM, Cass CE. Mouse transporter protein, a membrane protein that regulates cellular multidrug resistance, is localized to lysosomes. Cancer Res. 1999;59: 4890–4897.
140. Hogue DL, Kerby L, Ling V. A mammalian lysosomal membrane protein confers multidrug resistance upon expression in Saccharomyces cerevisiae. J Biol Chem. 1999;274: 12877–12882. doi:10.1074/jbc.274.18.12877
141. Tian S, Muneeruddin K, Choi MY, Tao L, Bhuiyan RH, Ohmi Y, et al. Genome-wide CRISPR screens for Shiga toxins and ricin reveal Golgi proteins critical for glycosylation. PLoS Biol. 2018;16: e2006951. doi:10.1371/journal.pbio.2006951
142. Yamaji T, Sekizuka T, Tachida Y, Sakuma C, Morimoto K, Kuroda M, et al. A CRISPR Screen Identifies LAPTM4A and TM9SF Proteins as Glycolipid-Regulating Factors. iScience. 2019;11: 409–424. doi:10.1016/j.isci.2018.12.039
143. Shao G-Z, Zhou R-L, Zhang Q-Y, Zhang Y, Liu J-J, Rui J-A, et al. Molecular cloning and characterization of LAPTM4B, a novel gene upregulated in hepatocellular carcinoma. Oncogene. 2003;22: 5060–5069. doi:10.1038/sj.onc.1206832
144. Adra CN, Zhu S, Ko JL, Guillemot JC, Cuervo AM, Kobayashi H, et al. LAPTM5: a novel lysosomal-associated multispanning membrane protein preferentially expressed in hematopoietic cells. Genomics. 1996;35: 328–337. doi:10.1006/geno.1996.0364
145. Grabner A, Brast S, Sucic S, Bierer S, Hirsch B, Pavenstädt H, et al. LAPTM4A interacts with hOCT2 and regulates its endocytotic recruitment. Cell Mol Life Sci. 2011;68: 4079–4090. doi:10.1007/s00018-011-0694-6
146. Milkereit R, Persaud A, Vanoaica L, Guetg A, Verrey F, Rotin D. LAPTM4b recruits the LAT1-4F2hc Leu transporter to lysosomes and promotes mTORC1 activation. Nat Commun. 2015;6: 7250. doi:10.1038/ncomms8250
147. Li L, Wei XH, Pan YP, Li HC, Yang H, He QH, et al. LAPTM4B: a novel cancer-associated gene motivates multidrug resistance through efflux and activating PI3K/AKT signaling. Oncogene. 2010;29: 5785–5795. doi:10.1038/onc.2010.303
148. Rutsch F, Gailus S, Miousse IR, Suormala T, Sagné C, Toliat MR, et al. Identification of a putative lysosomal cobalamin exporter altered in the cblF defect of vitamin B12 metabolism. Nat Genet. 2009;41: 234–239. doi:10.1038/ng.294
149. Kawaguchi K, Okamoto T, Morita M, Imanaka T. Translocation of the ABC transporter ABCD4 from the endoplasmic reticulum to lysosomes requires the escort protein LMBD1. Sci Rep. 2016;6: 30183. doi:10.1038/srep30183
150. Kitai K, Kawaguchi K, Tomohiro T, Morita M, So T, Imanaka T. The lysosomal protein ABCD4 can transport vitamin B12 across liposomal membranes in vitro. J Biol Chem. 2021;296: 100654. doi:10.1016/j.jbc.2021.100654
151. Wojnar P, Lechner M, Merschak P, Redl B. Molecular cloning of a novel lipocalin-1 interacting human cell membrane receptor using phage display. J Biol Chem. 2001;276: 20206–20212. doi:10.1074/jbc.M101762200

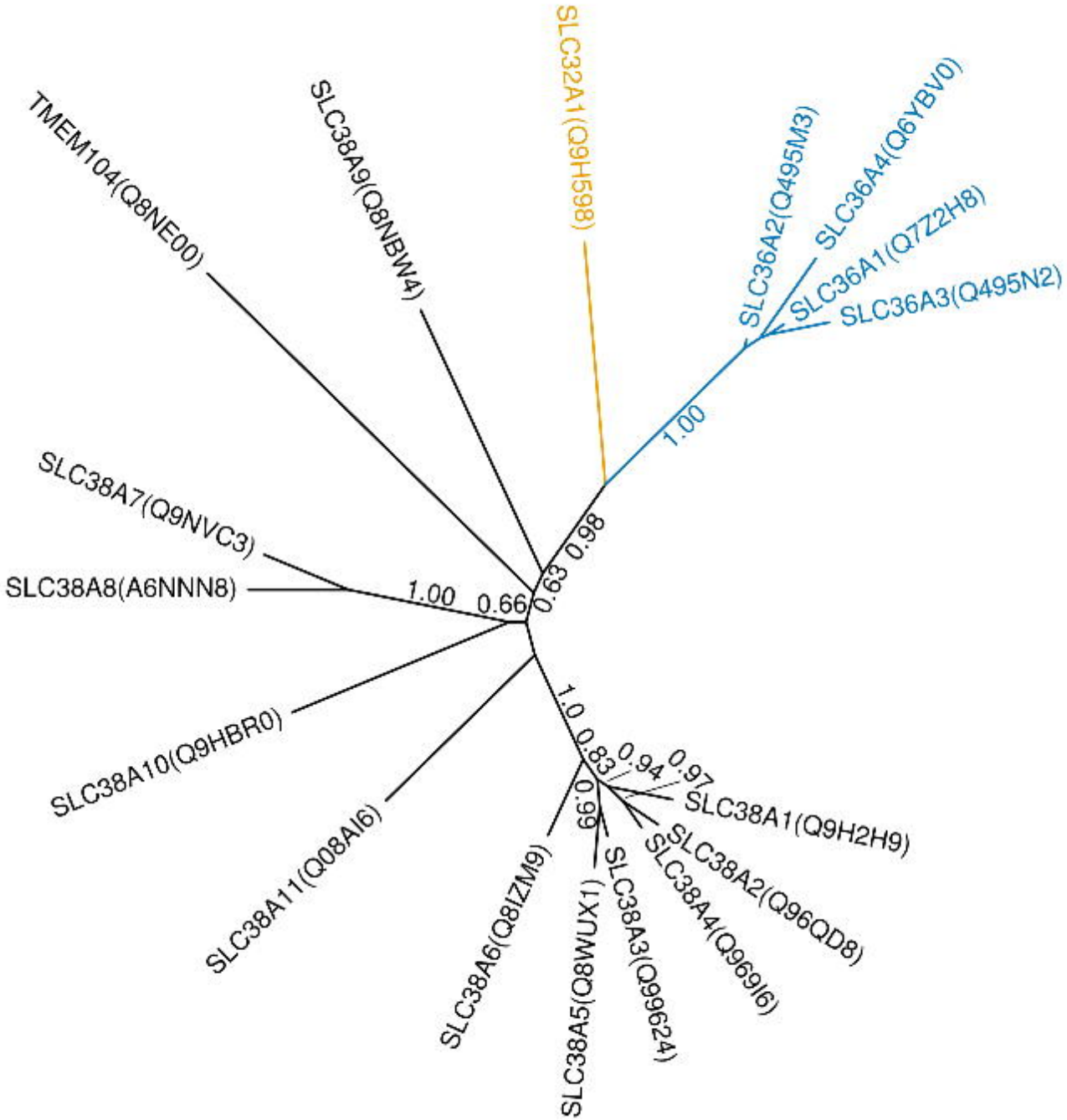
152. Hesselink RW, Findlay JBC. Expression, characterization and ligand specificity of lipocalin-1 interacting membrane receptor (LIMR). Mol Membr Biol. 2013;30: 327–337. doi:10.3109/09687688.2013.823018
153. Paek J, Kalocsay M, Staus DP, Wingler L, Pascolutti R, Paulo JA, et al. Multidimensional Tracking of GPCR Signaling via Peroxidase-Catalyzed Proximity Labeling. Cell. 2017;169: 338-349.e11. doi:10.1016/j.cell.2017.03.028
154. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 2003;12: 1725–1735. doi:10.1093/hmg/ddg180
155. Ianakiev P, van Baren MJ null, Daly MJ, Toledo SP, Cavalcanti MG, Neto JC, et al. Acheiropodia is caused by a genomic deletion in C7orf2, the human orthologue of the Lmbr1 gene. Am J Hum Genet. 2001;68: 38–45. doi:10.1086/316955
156. MacGrogan D, Levy A, Bova GS, Isaacs WB, Bookstein R. Structure and methylation-associated silencing of a gene within a homozygously deleted region of human chromosome band 8p22. Genomics. 1996;35: 55–65. doi:10.1006/geno.1996.0322
157. Knauer R, Lehle L. The oligosaccharyltransferase complex from Saccharomyces cerevisiae. Isolation of the OST6 gene, its synthetic interaction with OST3, and analysis of the native complex. J Biol Chem. 1999;274: 17249–17256. doi:10.1074/jbc.274.24.17249
158. Knauer R, Lehle L. The oligosaccharyltransferase complex from yeast. Biochim Biophys Acta. 1999;1426: 259–273. doi:10.1016/s0304-4165(98)00128-7
159. Kelleher DJ, Karaoglu D, Mandon EC, Gilmore R. Oligosaccharyltransferase isoforms that contain different catalytic STT3 subunits have distinct enzymatic properties. Mol Cell. 2003;12: 101–111. doi:10.1016/s1097-2765(03)00243-0
160. Cherepanova NA, Shrimal S, Gilmore R. Oxidoreductase activity is necessary for N-glycosylation of cysteine-proximal acceptor sites in glycoproteins. J Cell Biol. 2014;206: 525–539. doi:10.1083/jcb.201404083
161. Ramírez AS, Kowal J, Locher KP. Cryo-electron microscopy structures of human oligosaccharyltransferase complexes OST-A and OST-B. Science. 2019;366: 1372–1375. doi:10.1126/science.aaz3505
162. Goytain A, Quamme GA. Identification and characterization of a novel mammalian Mg2+ transporter with channel-like properties. BMC Genomics. 2005;6: 48. doi:10.1186/1471-2164-6-48
163. Zhou H, Clapham DE. Mammalian MagT1 and TUSC3 are required for cellular magnesium uptake and vertebrate embryonic development. Proc Natl Acad Sci USA. 2009;106: 15750–15755. doi:10.1073/pnas.0908332106
164. Li N, Gügel IL, Giavalisco P, Zeisler V, Schreiber L, Soll J, et al. FAX1, a novel membrane protein mediating plastid fatty acid export. PLoS Biol. 2015;13: e1002053. doi:10.1371/journal.pbio.1002053
165. Nilsson R, Schultz IJ, Pierce EL, Soltis KA, Naranuntarat A, Ward DM, et al. Discovery of genes essential for heme biosynthesis through large-scale gene expression analysis. Cell Metab. 2009;10: 119–130. doi:10.1016/j.cmet.2009.06.012
166. Yien YY, Robledo RF, Schultz IJ, Takahashi-Makise N, Gwynn B, Bauer DE, et al. TMEM14C is required for erythroid mitochondrial heme metabolism. J Clin Invest. 2014;124: 4294–4304. doi:10.1172/JCI76979
167. Yien YY, Ringel AR, Paw BH. Mitochondrial transport of protoporphyrinogen IX in erythroid cells. Oncotarget. 2015;6: 20742–20743. doi:10.18632/oncotarget.5124

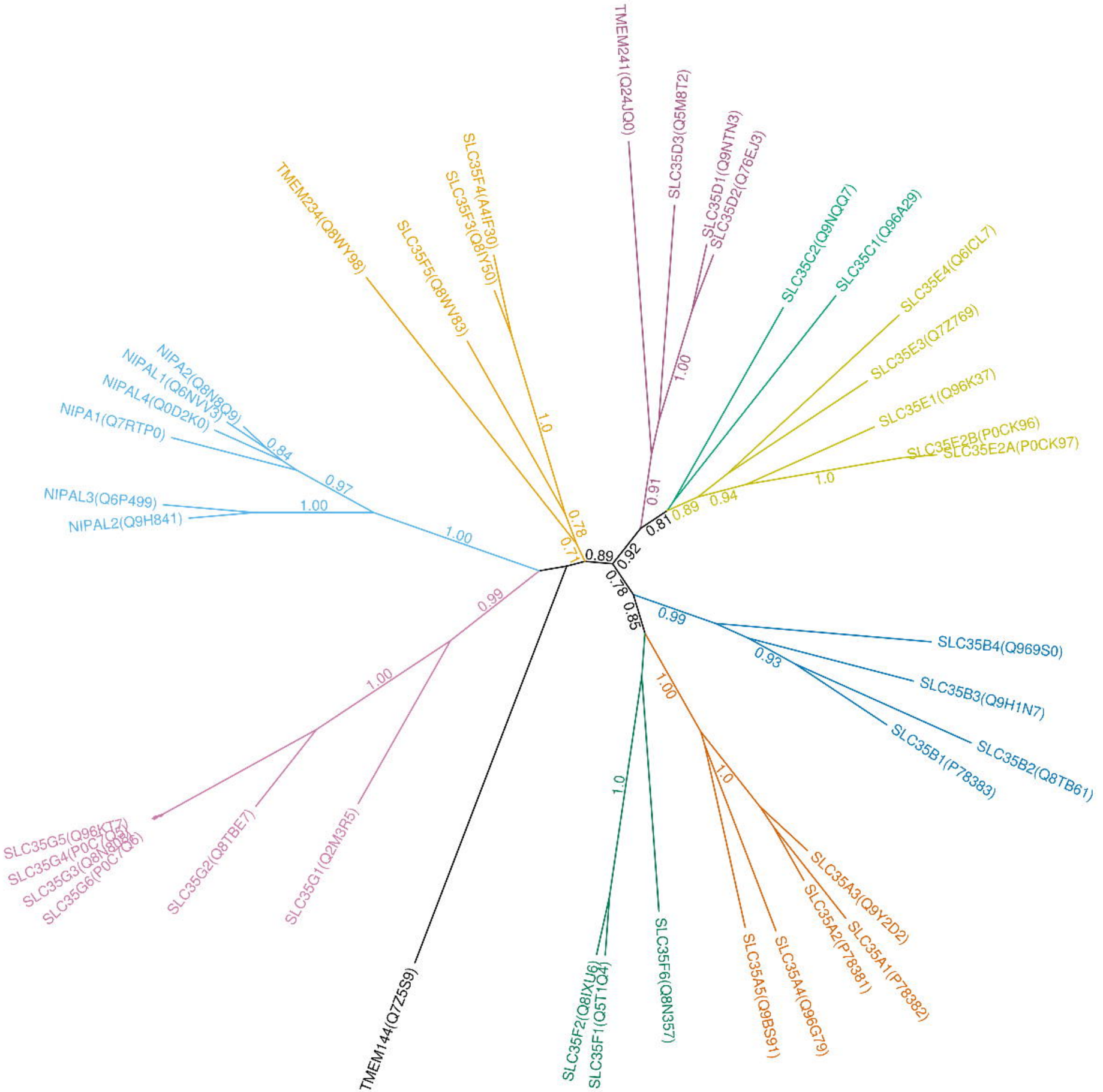
168. Klammt C, Maslennikov I, Bayrhuber M, Eichmann C, Vajpai N, Chiu EJC, et al. Facile backbone structure determination of human membrane proteins by NMR spectroscopy. Nat Methods. 2012;9: 834–839. doi:10.1038/nmeth.2033
169. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance. 2019;2. doi:10.26508/lsa.201900429
170. Shen D-W, Ma J, Okabe M, Zhang G, Xia D, Gottesman MM. Elevated expression of TMEM205, a hypothetical membrane protein, is associated with cisplatin resistance. J Cell Physiol. 2010;225: 822–828. doi:10.1002/jcp.22287
171. Gallenito MJ, Qasim TS, Tutol JN, Prakash V, Dodani SC, Meloni G. A recombinant platform to characterize the role of transmembrane protein hTMEM205 in Pt(II)-drug resistance and extrusion. Metallomics. 2020;12: 1542–1554. doi:10.1039/d0mt00114g
172. Chang J-M, Di Tommaso P, Taly J-F, Notredame C. Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics. 2012;13 Suppl 4: S1. doi:10.1186/1471-2105-13-S4-S1
173. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421
174. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin. 1958;38: 1409–1438.
175. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019;20: 473. doi:10.1186/s12859-019-3019-7
176. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45: D170–D176. doi:10.1093/nar/gkw1081
177. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7: 539. doi:10.1038/msb.2011.75
178. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018;27: 135–145. doi:10.1002/pro.3290
179. Lefort V, Longueville J-E, Gascuel O. SMS: Smart Model Selection in PhyML. Mol Biol Evol. 2017;34: 2422–2424. doi:10.1093/molbev/msx149
180. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59: 307–321. doi:10.1093/sysbio/syq010
181. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 2006;55: 539–552. doi:10.1080/10635150600755453
182. Chen K, Durand D, Farach-Colton M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol. 2000;7: 429–447. doi:10.1089/106652700750050871


SLC
Atlas
SLC41
SLC14
SLC29
SLC15
SLC19
SLC33
SLC45
SLC59
SLC43
SLC21/SLCO
SLC60
SLC49
SLC2
SLC22
SLC46
SLC16
SLC63
SLC37
SLC17
SLC18
SLC55
SLC6
SLC25
SLC5
SLC34
SLC56
SLC28
SLC9
SLC36
SLC38
SLC12
SLC7
SLC4
SLC57
SLC35
SLC26
SLC13
SLC1
SLC47
SLC39
SLC30
SLC65
SLC42
SLC11
SLC54
SLC66
SLC24
SLC8
SLC10
SLC31
SLC58
SLC51A
SLC44
SLC20
SLC23
SLC52




SLC17A6(Q9P2U8)
SLC17A7(Q9P2U7)
SLC17A8(Q8ND)2)
SLC17A5(Q9NRA2)
SLC17A4(Q9Y2C5)
SLC17A2(O00624)
SLC17A3(O00476)
SLC17A1(Q14916)
0.74
1.0
1.0
0.94
0.99
SLC17A9(Q9BYT1)
0.99
SLC37A1(P57057)
SLC37A2(Q8TED4)
SLC37A3(Q8NCC5)
SLC37A4(O43826)
1.00
0.99
0.97
0.98
0.97
0.91
SLC18B1(Q6NT16)
1.00
1.00
SLC18A3(Q16572)
SLC18A2(Q05940)
SLC18A1(P54219)
0.99
MFSD9(Q8NBP5)
SLC22A18(Q96BI1)
1.0
0.94
MFSD14B(Q5SR56)
MFSD14C(Q5VZR4)
MFSD14A(Q96MC6)
MFSD10(Q14728)
MFSD8(Q8NHS3)


SLC32A1(Q9H598)
TMEM104(Q8NE00)
SLC38A9(Q8NBW4)
SLC36A2(Q495M3)
SLC36A4(Q6YBV0)
SLC36A1(Q7Z2H8)
SLC36A3(Q495N2)
1.00
SLC38A7(Q9NVC3)
SLC38A8(A6NNN8)
1.00
0.66
0.63
0.98
SLC38A10(Q9HBR0)
SLC38A11(Q08AI6)
1.0
0.83
0.94
0.97
0.99
SLC38A1(Q9H2H9)
SLC38A2(Q96QD8)
SLC38A4(Q969I6)
SLC38A3(Q99624)
SLC38A5(Q8WUX1)
SLC38A6(Q8IZM9)


TMEM241(Q24JQ0)
SLC35D3(Q5M8T2)
SLC35D1(Q9NTN3)
SLC35D2(Q76EJ3)
SLC35F4(A4IF30)
SLC35F3(Q8IY50)
SLC35F5(Q8WV83)
TMEM234(Q8WY98)
SLC35C2(Q9NQQ7)
SLC35C1(Q96A29)
SLC35E4(Q6ICL7)
SLC35E3(Q7Z769)
SLC35E1(Q96K37)
SLC35E2B(P0CK96)
SLC35E2A(P0CK97)
NIPA2(Q8N8Q9)
NIPAL1(Q6NVV3)
NIPAL4(Q0D2K0)
NIPA1(Q7RTP0)
NIPAL3(Q6P499)
NIPAL2(Q9H841)
SLC35B4(Q969S0)
SLC35B3(Q9H1N7)
SLC35B1(P78383)
SLC35B2(Q8TB61)
SLC35G5(Q96KT7)
SLC35G4(P0C7Q5)
SLC35G3(Q8N8Q6)
SLC35G6(P0C7Q6)
SLC35G2(Q8TBE7)
SLC35G1(Q2M3R5)
TMEM144(Q7Z5S9)
SLC35F2(Q8IXU6)
SLC35F1(Q5T1Q4)
SLC35F6(Q8N357)
SLC35A5(Q9BS91)
SLC35A4(Q96G79)
SLC35A3(Q9Y2D2)
SLC35A2(P78381)
SLC35A1(P78382)
1.00
0.91
1.0
0.78
0.71
0.89
0.92
0.81
0.89
0.94
1.0
0.84
0.97
1.00
1.00
0.99
1.00
0.78
0.85
0.99
0.93
1.00
1.0
1.0