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 ~120 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

31 for over 50% of all transport-related proteins and about 10% of all membrane proteins encoded by the
32 human genome. With more than 400 annotated members, it is the largest superfamily of membrane
33 transporter proteins [1,2]. The roles of SLC transporters as cellular gatekeepers, determinants of nutrient
34 homeostasis and facilitators of drug metabolism and drug targeting has recently been revisited [3]. Around
35 ~50% of currently annotated SLCs are predicted to be associated with human disease phenotypes and many
36 SLCs are considered to represent promising drug targets or drug delivery systems or to affect drug ADMET
37 (absorption, distribution, metabolism, extrusion, toxicity). It has recently become evident that the SLC
38 superfamily offers an enormous unexplored therapeutic treasure. But while the list of approved drugs that
39 target transporter proteins is increasing, many promising SLCs still remain unexplored, uncharacterized and
40 underrepresented in the literature.

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

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

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

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

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

82 The Transporter Classification DataBase (TCDB) is an alternative classification system that was created in the
83 1990s in parallel to the SLC nomenclature series [14,15]. It collects transport-related membrane proteins,
84 including membrane receptors, transporters, ion channels, and membrane-anchored enzymes from all
85 kingdoms of life, with a particular focus on proteins from lower organisms. The proteins in the TCDB are
86 organized hierarchically into subfamilies, families and superfamilies based on phylogenetic and functional
87 considerations, and each member in the database is given a five-segmented TC# similar to the EC# that is
88 used for enzyme classification. In addition, a brief description is provided for each family that introduces
89 identified members and contains links to the most important relevant papers.

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

95 It also seems problematic to consider all proteins in the TCDB that are annotated as part of the secondary
96 transporter superfamily TC# 2.A as being SLC-like. Indeed, many of the TCDB families annotated as part of
97 TC# 2.A exhibit structural or sequence features that do not match the characteristics of existing SLCs.
98 Examples of this are the Trk K^+ transporters (#2.A.38), which display an ion-channel like structural fold [16],
99 the GUP glycerol uptake proteins (#2.A.50), which exhibit enzymatic activity based on follow-up studies [17],
100 and the Twin Arginine Targeting (Tat) family (#2.A.64), which are actually protein secretion complexes
101 [18,19]. Since none of these families correspond to structural or functional characteristics of currently known
102 SLC transporters, it is likely that not all proteins annotated under the TC# 2.A superfamily are "SLC-like".
103 Thus, while the TCDB could be a rich source of information for finding new transporters, it is clear that the
104 perception of what a "transporter" should be according to TCDB does not always correspond to the typical
105 properties of well-characterized SLC proteins. Additional filtering of TCDB data is therefore necessary.

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

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

125 Therefore, there is a clear need in the field to define the criteria of "SLC-likeness" and to identify and classify
126 all proteins in humans and other species that exhibit "SLC-likeness". Thus, in our work, we interpret the term
127 "SLC-likeness" by defining the essential criteria for it, and we carry out an exhaustive search for proteins that
128 potentially meet these criteria, both with manual curation of datasets and with automatic sequence
129 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). Proteins with a β-barrel architecture, 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 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.

(4) Nucleotide-driven transporters (e.g. ATP-binding cassette/ABC, Energy-coupling factor/ECF) were excluded.

(5) 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.

(6) 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.

156 (7) Proteins where transport activity was used as a synonym for trafficking (i.e. protein or vesicle
157 translocation within the cell) but otherwise seemingly having no small-molecule transmembrane
158 transport activity were excluded. On this basis, chaperones and other proteins helping the
159 insertion of nascent proteins into a cellular membrane were also excluded.
160 (8) For some (mostly putative) transporter families, TCDB does not give an explanation why the
161 proteins would be considered as transporters. Families with no resemblance to known
162 transporters and no indication or argument as to why they would be transporters were
163 excluded.

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

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

184 Search for novel "SLC-like" proteins

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

192 **Table 1. Families and subfamilies examined from the TCDB for SLC-like proteins.** Superfamilies marked in
193 boldface have been analyzed. The superfamily #1.A was examined since several existing SLC families are
194 classified here, while superfamily #2.A was expected to contain most known SLC proteins. Superfamilies #9.A
195 and #9.B were also examined. Total number of families and subfamilies in the first-level classes are indicated
196 in parentheses. Numbers show the total number of level 3 families and level 4 subfamilies in each
197 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**	**24**		**535**	**43**
	Total:	(1367)			(2795)		
			616	**166**		**1534**	**600**

198

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

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

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

223 • The criterion about the size of the substrate was judged based on the known or putative substrate
224 according to TCDB family descriptions. The assumption that certain proteins are transporters,
225 whether based on experimental observations or purely on prediction, always entails the presence of
226 either a tested or an assumed class of substrates. Families where no indication was available why
227 they would function as transporters were thus excluded for the moment until further information on
228 their function becomes available.

229 • The structural criterion was assessed based on the "number of transmembrane segments"
230 annotations in the TCDB for each protein. (Sub)families where their TCDB descriptions mentioned a
231 β-barrel structure or relationship with such proteins were excluded. Care was taken to identify
232 outliers, i.e. proteins with less TMHs as the functional unit, which could potentially be fragments,
233 and proteins with extra TMHs in addition to the functional core. In this regard, we based our
234 decision on Pfam models conserved in the (sub)family and the number of THMs the model spans
235 according to predictions at the Pfam web site for the most characterized members of the
236 (sub)family. For certain Pfam families, structures and annotations from the Protein Data Bank (PDB)
237 linked with the family on the Pfam web site were used to decide whether or not the Pfam model
238 encodes soluble proteins, provided that the family model has sufficient coverage by the structure
239 available.

240 • In some TCDB families with multicomponent transport systems, the membrane-spanning core
241 components were identified by the presence of conserved Pfam models in all members of the
242 family. Whether Pfam domains were membrane-spanning was deduced by their descriptions at the
243 Pfam web site and Pfam prediction of TMHs. The TCDB description of the family also helped in
244 deciding which components were essential and membrane-spanning. If multiple such Pfam models
245 were found, the proteins were assumed to function as an obligate heteromultimer, and such

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

248 Strikingly, of the 1534 subfamilies examined, only 600 in 166 families were found to be SLC-like according to
249 our criteria. In particular, of the 556 subfamilies within superfamily TC# 2.A ("porters"), only 501 appeared to
250 meet our selection criteria. This underlines once more that the term "porter" in relation to solute transport
251 is ambiguous in this field and a clearer definition of the perception of a solute transporter protein is needed.

252 Our curation efforts also yielded 209 Pfam families bearing SLC-like properties, which likely contain the
253 membrane-spanning regions of SLC-like proteins selected from the 336 Pfam families present in total in the
254 TCDB sequences analyzed. Special care has been taken to consistently include or exclude TCDB (sub)families
255 and their corresponding Pfam family, if applicable, from our selection. Notably, many of those Pfam families
256 that have been excluded represented soluble structural or regulatory domains. Following our initial round of
257 selection, we took advantage of clan groupings in Pfam and extended our selection efforts to analyze Pfam
258 families belonging to the same clan as the selected 209 SLC-like Pfam families. This was based on the
259 observation that many SLC-like transporter families in Pfam appear grouped into clans, so that members of
260 these clans might represent SLC-like transporters themselves. As noted before, Pfam clans represent
261 remotely related protein families, and while protein function is not always conserved across different
262 families within a Pfam clan, the inspection of these families whether they meet our SLC-likeness criteria is
263 validated by their relatedness or similarity to SLC-like families. Such a "clan expansion" procedure resulted in
264 12 additional Pfam families that are evolutionarily related to SLC-like protein families, of which 8 Pfam
265 families potentially meet our criteria as SLC-like, as evaluated according to the guidelines and criteria above.
266 Interestingly, these 8 SLC-like Pfam families currently have no representatives with a modest score (bit score
267 > 25) in the TCDB, while 4 of the 8 families are annotated in Pfam as domains of unknown function (DUF).
268 Thus, a total of 217 SLC-like Pfam families were identified in our search (see S1 Table).

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

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

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

284 The results of our search for SLC-like proteins are summarized in numbers in Table 2. Briefly, 59-67 of the
285 166 TCDB families have representatives in the 7 organisms (66 in human), and the organisms seem to
286 contain 434-673 SLC-like proteins (549 in human). In total, 3733 proteins have been found in the 7 organisms
287 studied. Notably, the number of SLC-like transporters found in human in our search is ~130 higher than
288 previously reported [2], indicating that the human SLC-ome may be significantly larger than previously
289 thought.

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

Species	Level 3 families (TC# x.y.z)	Level 4 subfamilies (TC# x.y.z.w)	Pfam models	Proteins
Human (*H. sapiens*)	66	308	79	549
Mouse (*M. musculus*)	67	314	79	532
Rat (*R. norvegicus*)	67	316	79	537
Zebrafish (*D. rerio*)	65	312	80	673
Chicken (*G. gallus*)	63	305	77	473
Fruitfly (*D. melanogaster*)	60	273	66	434
Roundworm (*C. elegans*)	59	283	69	535

After arriving at this initial set of SLC-like proteins, we proceeded by a performing a first round of sanitization based on human proteins. First, we manually investigated the UniProt records of all human hits that are non Swiss-Prot-validated sequences, which revealed that 19 out of 23 are likely non-human or fragment sequences (S2 Table). After excluding these, we investigated the number of helical transmembrane (TM) segments annotated in UniProt for 144 of all remaining 530 human hits that do not correspond to classic SLC transporters (families SLC1-SLC52). In total, 17 of these proteins, showing similarity to 8 different TCDB families, contained less than 3 TMHs based on UniProt annotations and thus did not satisfy our structural criterion (1). These proteins have been investigated individually based on their coverage of HMMs used in the search and the location of TMHs according to UniProt annotations in other SLC-like proteins covering the same HMMs, as well as available literature on oligomeric state and function. This way, we have retained 8 sequences where the functional unit contains 3 or more TMHs, or the protein spans at least 3 TMHs based on HMM coverage, or for other reasons (see S4 Table). These proteins have been included in Table 3 and S4 Table. In contrast, 10 of the 17 proteins seemed to cover regions of HMMs where other members did not seem to contain TMHs, while in addition containing less than 3 TMHs or known non-transporter TM domains. Notably, certain Pfam models seem to cover both TM and non-TM domains, and several TCDB families and their corresponding HMMs encode multi-domain proteins, leading to hits that show similarity to the model in the non-TM region, but not in the TM region that would be crucial for SLC-likeness. These 10 proteins have been excluded from further analyses and included in S2 Table. Furthermore, initial SLC-like hits

315 that only match with such non-TM regions of these HMMs (#2.A.19.3 positions 250-725, #9.B.64.1 positions
316 above 200, “OST3_OST6” positions below 200, “RBP_receptor” positions above 400) in all other organisms
317 have also been excluded from further analyses. Thus, in this initial sanitization step, 29 human proteins, as
318 well as 8, 10, 2, 9, and 6 proteins from *M. musculus*, *R. norvegicus*, *C. elegans*, *D. rerio*, and *G. gallus*,
319 respectively, have been excluded, leaving a remaining 3669 proteins for further analysis, including 520 from
320 human.

321 As a next step, we proceeded by attempting to classify the remaining proteins into protein families, followed
322 by manual database and literature searches focused on novel human proteins, as detailed below.

323 Classification into families

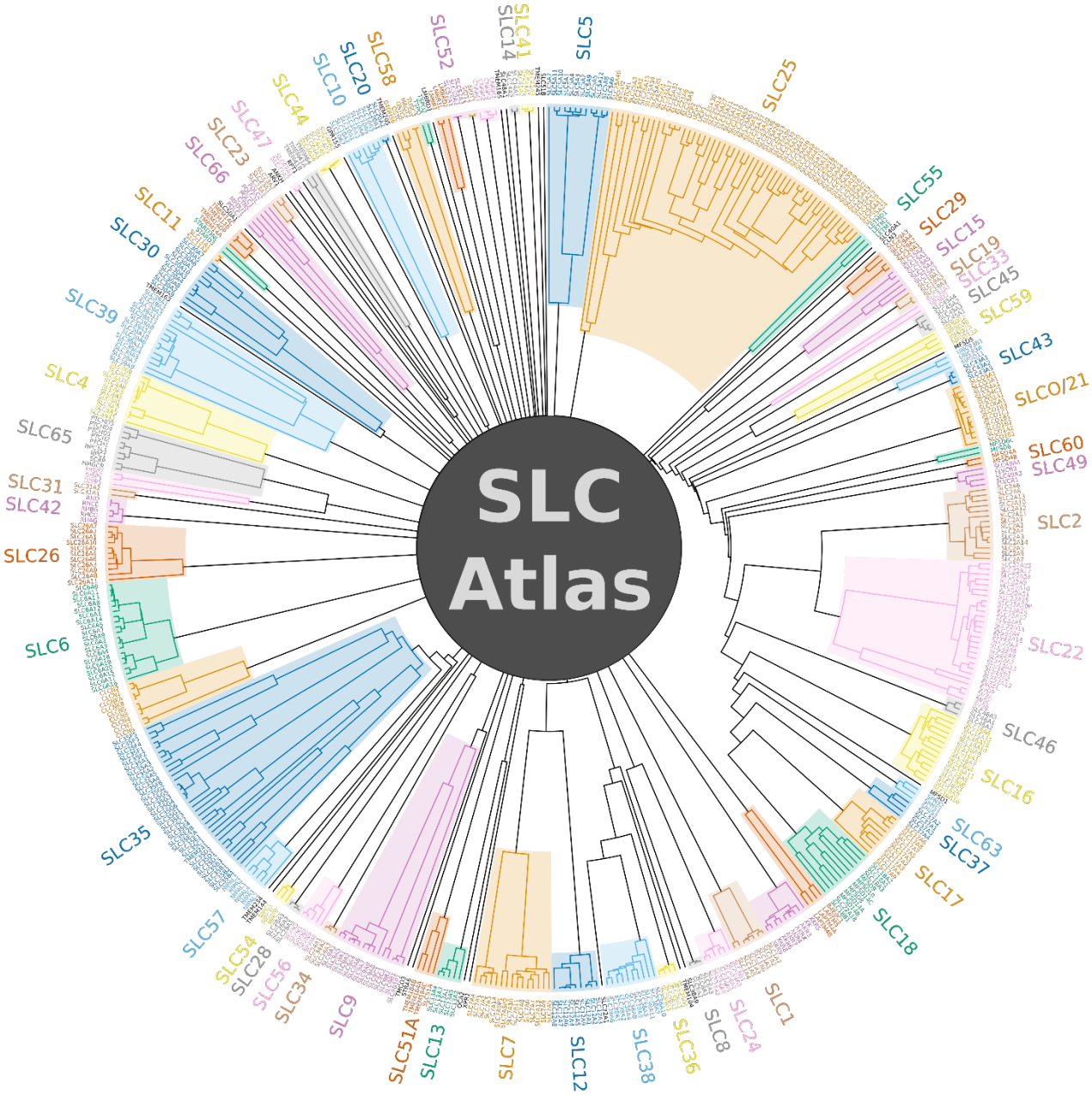
324 As mentioned in the introduction, SLC carriers are likely of polyphyletic origin, and individual families can be
325 so diverse that even sensitive sequence similarity-based methods may have difficulties grouping related SLCs
326 [12]. In our experience, multiple sequence alignment-based methods were not able to cluster the identified
327 SLC-like sequences and to reproduce known SLC families, so we devised a custom method for clustering
328 distantly related sequences into proteins families based on the introduction of “HMM fingerprints”. An
329 HMM fingerprint is a mathematical vector of numbers assigned to a protein sequence, where the numbers
330 represent the similarity scores of that protein sequence against each of the TCDB families, subfamilies and
331 Pfam families that we have selected to be SLC-like. Thus, two protein sequences that show a similar pattern
332 in their HMM fingerprints indicate their similarity. The usefulness of an HMM fingerprint depends on a
333 meaningful definition of HMMs used in the fingerprint, whereby we capitalize on the evolutionary principle
334 in the construction of TCDB families, subfamilies as well as Pfam families. Nevertheless, our goal was not to
335 reconstruct the evolutionary history of a set of proteins, but to group proteins that share similar sequence
336 features. However, due to the transitivity of homology and since similarity to a group of proteins suggests
337 homology, clusters derived using HMM fingerprinting are likely to contain homologous proteins.

338 The HMM fingerprint-based classification of the SLC-like proteins found in our search yielded 102 protein
339 families in total, 94 of which had representatives in human (Fig 1). For existing SLC transporters, the

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

360 **Figure 1. Dendrogram of human SLC-like protein families based on HMM fingerprint-based clustering.**
361 Classic and newly incorporated SLC families are shown with colors. Since SLC proteins are polyphyletic, which
362 manifests itself in mathematically orthogonal HMM fingerprints that cannot be further clustered, the
363 dendrogram does not join into a single branch. Branch lengths have been transformed for better visibility
364 (see Methods).

365



366

367 Our analysis revealed 42 new proteins families in total in human, containing proteins that have not yet been

368 annotated as SLCs. Interestingly, our search has also found new proteins that clustered into existing SLC

369 families (Table 3, S4 Table).

370 **Table 3. Novel SLC-like proteins.** The table shows SLC-like human proteins from our search that are not in

371 the classic group of SLCs (SLC1-52 families). The most similar (highest-scoring) TCDB family/subfamily and

372 Pfam family are shown, as well as the most similar (highest-scoring) protein with structural information

373 based on the pdb70 dataset (see Methods). Substrate information, where available, was retrieved from

374 available literature. Based on our search, SLC families SLC53-66 have recently been incorporated into the

375 nomenclature in collaboration with the HUGO/HGNC. Proteins marked with a single asterisk are included

376 due to sequence similarity but may have functions other than transporter. Proteins marked with double

377 asterisks might have a channel-like transport mechanism. For additional details, see 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
SLC42 family					
RHCE*	1.A.11.4	Ammonium_transp	Blood group Rh(CE) polypeptide	Amt	orphan
RHD*	1.A.11.4	Ammonium_transp	Blood group Rh(D) polypeptide	Amt	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; also binds 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	Patched	Sterol regulatory element-binding protein cleavage-activating protein	RND (half)	cholesterol, 25-hydroxycholesterol
HMGCR*	2.A.6.6	Patched	3-hydroxy-3-methylglutaryl-coenzyme A reductase	RND (half)	probably sterols
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	CorB	orphan
CNNM2	1.A.112	DUF21	Metal transporter CNNM2	CorB	Mg^{2+}, Co^{2+}, Mn^{2+}, Sr^{2+}, Ba^{2+}, Cu^{2+}, Fe^{2+}, Na^+
CNNM3	1.A.112	DUF21	Metal transporter CNNM3	CorB	orphan
CNNM4	1.A.112	DUF21	Metal transporter CNNM4	CorB	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	MagT	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	orphan
TMEM14B	2.A.126.1	Tmemb_14	Transmembrane protein 14B	TMEM14	orphan
TMEM14C	2.A.126.1	Tmemb_14	Transmembrane protein 14C	TMEM14	protoporphyrinogen IX
TMEM14DP	2.A.126.1	Tmemb_14	Transmembrane protein 14DP	TMEM14	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

382 [35,36]. On the other hand, the lack of predicted similarity to any protein of known structure could pin down
383 interesting targets for structural biology efforts by highlighting proteins likely belonging to new fold families.

384 We performed HMM-based searches on the pdb70 database (sequences of the proteins represented in the
385 Protein Data Bank clustered to 70% sequence identity, see Methods) to assess whether structural
386 homologues are available for the proteins found. In total, for 79 of the 102 families, at least one similar
387 protein was found with a corresponding structure in the Protein Data Bank. Importantly, 477 human SLC-like
388 proteins likely have a homologue whose structure has been solved. On the other hand, 43 human SLC-like
389 proteins belonging to 19 different families do not seem to have homologues with a known structure and
390 thus are likely to constitute novel fold families. For the classical SLC proteins, their best-scoring similar
391 proteins from the pdb70 dataset and the corresponding structural fold families are summarized in S3 Table.
392 Based on this, it appears that classical SLCs from families SLC34, SLC44, SLC48 and SLC51 are still "structural
393 orphans".

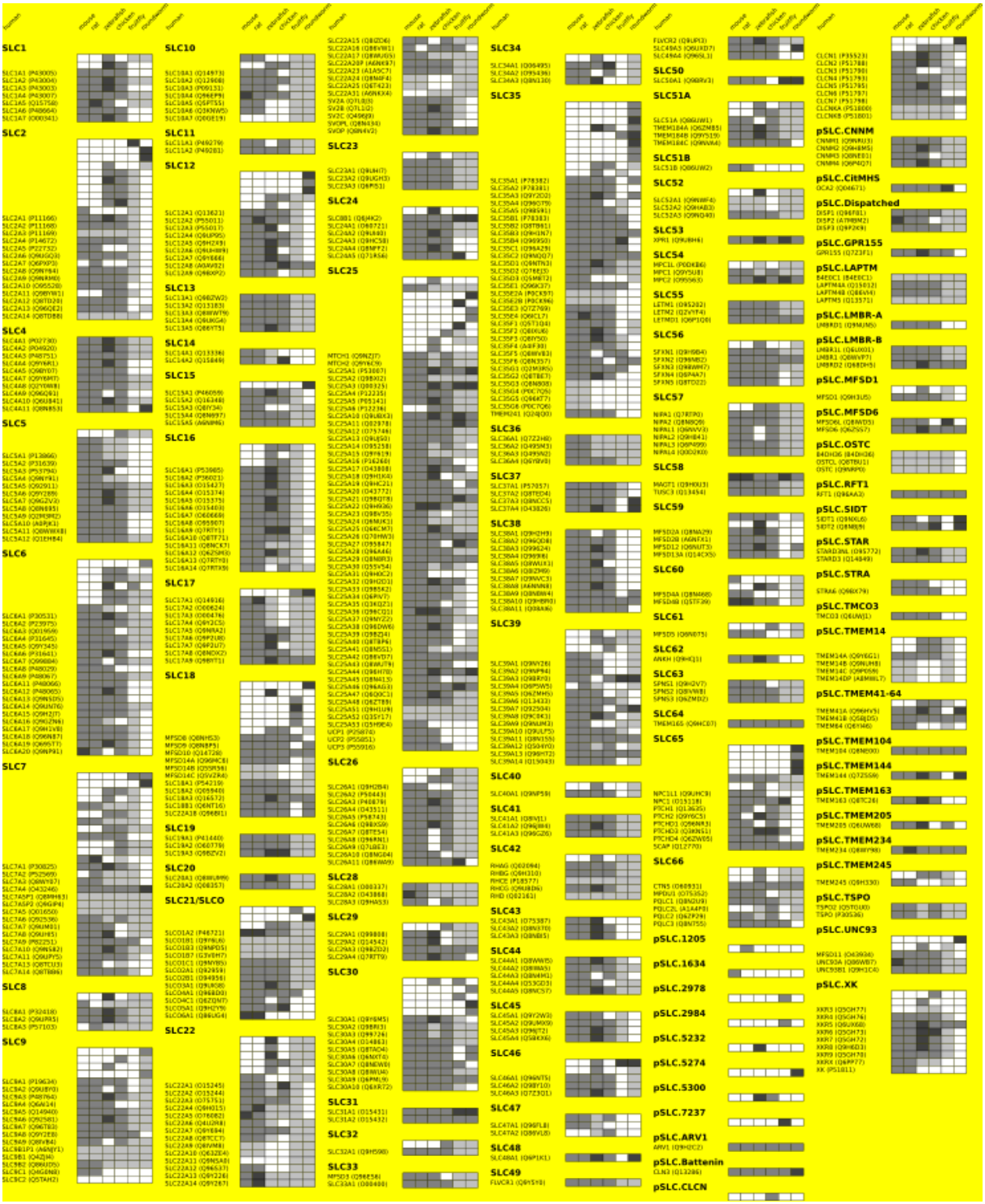
394 Detecting remote homology to proteins with a known transporter fold can also support their transporter
395 function as well as give clues about their transport mechanism. Fold families that are well-characterized and
396 typically host transporters with an alternating access mechanism encode potential transporter-like
397 structures. Out of the 18 fold families to which novel SLC-like proteins show similarity to (S4 Table),
398 representatives of 7 have been observed in various conformational states that hint at an alternating access
399 mechanism, while for a further 7 fold families, a transport mechanism has been suggested (S4 Table). These
400 fold families cover 87 of the novel SLC-like transporters, suggesting a possible transporter-like transport
401 mechanism for these proteins (S4 Table).

402 Phylogenetic analysis

403 Model organisms can be useful to study the biological function of various proteins, including solute
404 transporters. In order to relate the results to human, however, knowledge of orthologous gene pairs is
405 necessary. To this end, we performed phylogenetic analysis on each family of SLC-like proteins
406 corresponding to our clustering. In brief, unrooted phylogenetic trees for each SLC-like protein family from

407 all organisms have been generated and reconciled with the species tree of the 7 organisms in our study to
408 identify gene duplication and speciation events in their evolutionary histories (see Methods). The resulting
409 evolutionary trees are deposited in S1 File. Based on these trees, we carried out orthology analysis focusing
410 on human proteins and the human lineage. The resulting data is presented in Fig 2, showing relationships
411 between human genes and their orthologs in the other 6 organisms in our study. In addition, gene clusters
412 are presented that have arisen through gene duplication events in the evolutionary history of the human
413 lineage, but where the corresponding human genes have likely been lost.

414 **Figure 2. Orthology analysis of SLC-like proteins based on the human lineage.** Each line shows a gene
415 cluster that has evolved through gene duplication events along the human lineage. The human gene in the
416 cluster, if present, is noted in the text label. Normal grey boxes in each column indicate a 1:1 orthology
417 relationship between the human gene and a corresponding gene in the organism specified. Dark grey boxes
418 indicate that the human gene has several orthologs, and light grey boxes indicate that several human genes
419 share a common ortholog in that organism. White boxes denote no ortholog in the organism specified, while
420 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 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

427 Table 3. While examining the available literature of these proteins, we have also extracted information
428 related to our criteria for SLC-likeness, especially with a focus on structure, function and transport
429 mechanism. In total, 53 proteins could be assigned a small-molecule substrate and therefore fulfill our
430 criterion for substrate size (2), and for a further 6, putative small-molecule substrates have been suggested.
431 Nevertheless, 74 proteins, while showing sequence similarity to transporters, are still "orphans" with no
432 indication of a substrate. We have not found any indication in the literature that any of the proteins found in
433 our search would have nucleoside-driven transport activity, therefore all of these proteins likely also fulfill
434 our criterion 4. Interestingly, 9 proteins have been found that have described roles other than
435 transmembrane solute transport, while showing similarity to SLC-like transporter proteins. These proteins
436 potentially do not fulfill our functional criteria (5, 6, 7), they are marked with an asterisk in Table 3 and the
437 details are further evaluated in S4 Table. In total, 3669 potentially SLC-like proteins have been found, and
438 out of the 520 human proteins, 134 proteins that are not classic SLC transporters showed similarity to SLC-
439 like proteins. Ruling out proteins with suggested other functions and a channel-like mechanism leaves 119
440 proteins, out of which 79 show similarity to transporters with a proposed transport mechanism, while 47
441 could be assigned endogenous small-molecule substrates.

442 Discussion

443 Our curation and search efforts have revealed a surprising 134 human proteins that are SLC-like but have not
444 been officially part of the SLC nomenclature before the start of our study. Interestingly, around 30 of them
445 were addressed in an earlier study and referred to as "atypical SLC transporters" [10,11]. All of these atypical
446 transporters have also been identified in our study, together with many others, including several that are not
447 part of the major facilitator superfamily (MFS) or the amino acid-polyamine-cation (APC) transporter
448 superfamily.

449 Our HMM fingerprint-based classification method yielded 102 protein families. Interestingly, in order to
450 reproduce some of the classic SLC families, we had to introduce clustering constraints to artificially split or
451 merge predicted families. In turn, increasing or decreasing the clustering threshold by 0.1 changes the

452 number of families to 98 and 107, respectively. Researchers have reported similar problems with the
453 classification of the highly diverse protein set of the Pfam database [29]. In particular, Finn et al. note that
454 closely related Pfam families may have artificially high thresholds to prevent them from overlapping, while
455 divergent families cannot always be covered by a single model [29]. Thus, we believe that however protein
456 families are shaped, the discovery of evolutionary relationships between them will always be necessary to
457 define higher-order superfamilies, similarly to Pfam clans [29]. Similarly, the analysis of the evolutionary
458 relationship of proteins within each family is equally validated, and several classic SLC families indeed
459 contain established subfamilies based on sequence and/or functional comparisons [37–40]. In this regard,
460 we believe that our work provides correlation between transporter proteins in the studied organisms with
461 the hierarchy of families and subfamilies in the TCDB database on the one hand, and a scale-free framework
462 for protein grouping and classification through HMM fingerprints on the other hand. Still, more work is
463 needed to examine the evolutionary relatedness of different SLC-like families using more robust
464 phylogenetic approaches.

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

475 The interesting cases of apparent outlier sequences in several families (SLC5A7, SLC10A7, SLC25A46,
476 SLC30A9, SLC39A9, MPDU1/SLC66A5 as well as the SLC9B and SLC35 families) have previously been partially
477 documented. These members show striking sequence and/or functional divergence from other members of

478 their SLC families. SLC5A7 diverges from other SLC5 proteins in phylogenetic trees [42,43] and shares only
479 20-25% sequence identity with them [44]. The sequence of SLC10A7 is more similar to its bacterial relatives
480 than to other SLC10 members, and its genetic structure, particularly its number of exons, is also different
481 from other proteins in the SLC10 family [45,46]. SLC10A7 also appears to have diverged at the functional
482 level, as it has been suggested to be a regulator of Ca^{2+} influx [47], while having no documented transport
483 activity. SLC25A46 turns out to be an outer mitochondrial membrane protein [48], in contrast to most other
484 members of the SLC25 family, which are generally located on the inner mitochondrial membrane. Consistent
485 with this, its yeast ortholog Ugo1, referred to as "a degenerate member of the mitochondrial metabolite
486 carrier family", was reported to be part of the mitochondrial outer membrane fusion machinery [49].
487 Similarly, MTCH1 and MTCH2, which also seem to be distantly related to other SLC25 members according to
488 our dendrogram (Fig 1), are also outer mitochondrial membrane proteins [50,51], and together with
489 SLC25A46, are collectively referred to as "peculiar" members of the family [52]. SLC30A9 is found in the
490 cytosol and nuclear fractions and functions as a coactivator of nuclear receptors after hormonal stimulation
491 [53], in contrast to other members of the family, which are Zn^{2+} transporters. SLC39A9, the only member of
492 "subfamily I" of Zrt/Irt-like proteins (ZIPs) [37] might function as an androgen hormone receptor [54], while
493 most other family members are primarily transporters of divalent metal ions. SLC9B subfamily proteins
494 NHA1 and NHA2 of the SLC9 Na^+/H^+ exchanger family are more similar to their bacterial homologs than to
495 SLC9A members [55].

496 In the following sections, we discuss the non-classic set of human SLC-like proteins resulting from our search,
497 with a particular focus on the identities of the transported substrates and, if known, the structural and
498 mechanistic aspects of transport. In certain cases, the proteins have already been officially included in the
499 SLC nomenclature as per our initiative and as a result of the precursor of this work, following approval by the
500 Human Gene Nomenclature Committee (HGNC). Where existing information about substrate and function is
501 not available, we have speculated on these aspects using available information in the literature and
502 considerations of sequence similarity. For several proteins, however, their classification requires further
503 consideration. We would also like to articulate open questions about what further work is needed in order to

504 identify additional transporters. We believe our approach has been useful to pinpoint proteins that have a
505 high probability of being novel transporters. Thus, specific biological, biochemical and structural efforts
506 could focus on these specific targets highlighted in our work, all of which would contribute to a complete
507 assessment of the SLC-ome in human cells.

508 Transporters with existing evidence for transport function

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

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

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

523 Interestingly, the **TMEM163** protein clustered together with SLC30 zinc transporter (ZnT) proteins, since the
524 "Cation_efflux" Pfam model, representative of the SLC30 family, was present in its sequence, although at a
525 low score and non-significant e-value (2.9e-3). In the TCDB, TMEM163 is also classified under subfamily
526 #2.A.4.8, sharing a common family with SLC30 transporters (#2.A.4.2). Multiple sequence alignment as well
527 as pairwise alignments with existing SLC30 members reveal very low sequence identity with SLC30 proteins
528 (4.2-14.4%), albeit these numbers are similar to those of SLC30A9 (6.5-13.4%). Given the marginal similarity

529 to the “Cation_efflux” domain, it is tempting to assume that SLC30 proteins and TMEM163 are distantly
530 related. Indeed, TMEM163 has been shown to bind [61] and transport Zn^{2+} [62–64], and substitution of its
531 proposed substrate-binding residues with alanine abolished Zn^{2+} efflux activity [64]. Transport has been
532 demonstrated to be H^{+}-coupled, and the protein functioning as a dimer [62], while extruding Zn^{2+} from the
533 cell [64]. Intracellularly, TMEM163 was originally shown to be expressed in synaptic vesicles [65]. In
534 overexpression systems, it is localized to both the plasma membrane and intracellular membrane
535 compartments [64]. TMEM163 has been linked to Parkinson’s disease (PD) [66], even though the opposite
536 conclusion has also been drawn [67]. TMEM163 has also been reported to be upregulated by olanzapine, a
537 psychotropic drug prescribed for PD patients [68]. In addition, TMEM163 was also shown to be highly
538 expressed in insulin secretory vesicles in human pancreas [69], and has been identified as a risk factor in type
539 2 diabetes [70,71]. Disruption of TMEM163 expression might impair insulin secretion at high glucose stimuli
540 [69].

541 The **TMEM165** protein clustered into its own family and is the only protein in human containing the
542 “UPF0016” Pfam domain and showing similarity to TCDB family #2.A.106.2. TMEM165 is a member of a
543 highly conserved family of transmembrane proteins that is present in many species of eukaryotes and
544 bacteria [72]. Initially, TMEM165 and its yeast homolog, Gdt1p, have been hypothesized to be Ca^{2+}/H^{+}
545 exchangers [73,74]. However, recently, evidence has been mounting about its involvement in manganese
546 (Mn^{2+}) homeostasis [74], and both Ca^{2+} and Mn^{2+} transport activity has directly been shown [75]. TMEM165
547 is localized to the trans-Golgi in human cells [72], and is proposed to play a crucial role in regulating Mn^{2+}
548 uptake into the Golgi apparatus [72,74]. In line with this, its homologs in other organisms, also containing
549 the UPF0016 domain, are also annotated as Mn^{2+} transporters [74]. Manganese plays an important role as a
550 co-factor for enzymes involved in glycosylation, and impairment of TMEM165 function results in
551 glycosylation defects. Indeed, mutations of TMEM165 found in patients with congenital disorder of
552 glycosylation (CDG) type II hamper the transport function or localization of TMEM165 [75]. Due to the
553 importance of TMEM165 in lactate biosynthesis [76], it has also been suggested that TMEM165 could be a
554 transporter importing both Ca^{2+} [77] and Mn^{2+} into the Golgi in exchange for protons [74]. TMEM165

555 proteins contain two copies of the UPF0016 domain, and each domain contains a signature motif, E-φ-G-D-
556 (K/R)-(T/S), where φ denotes a hydrophobic amino acid. The glutamic acid of the second motif, E248, has
557 been shown to be crucial for affecting the glycosylation function of the Golgi but not the expression of the
558 protein [78], and so can be speculated to form part of a binding site for transporter function. However, in
559 the absence of an experimentally determined structure, further investigation will be required to understand
560 the transport mechanism of TMEM165.

561 Proteins with sequence similarity to existing transporters

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

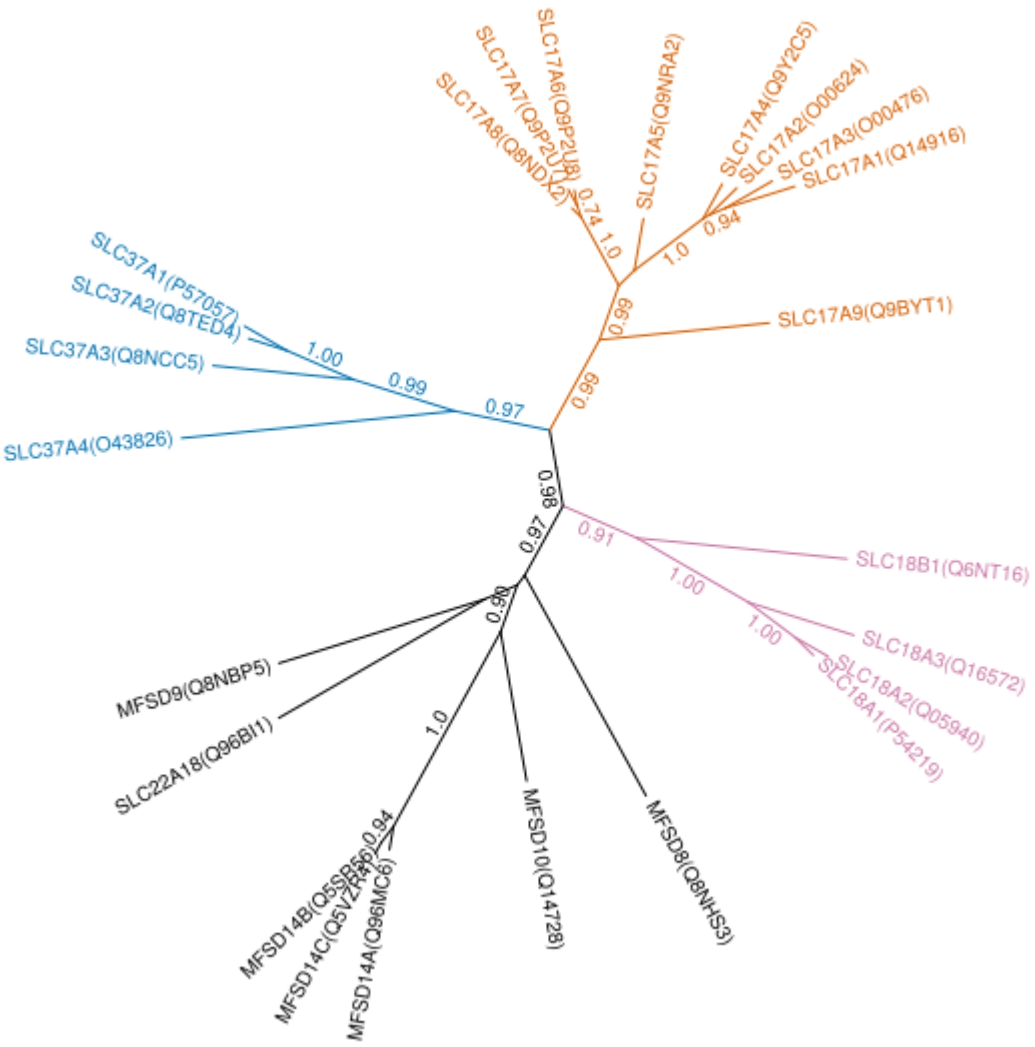
567 Atypical transporters

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

577 Some of the "atypical" SLC-like transporters (e.g. MFSD8, MFSD9, MFSD10 and MFSD14 proteins) clustered
578 together with members of the classical SLC18 family, which prompted us to examine the relationship of
579 these and neighboring proteins in more detail. We constructed multiple alignments and a phylogenetic tree

580 of the proteins one level above these proteins in our clustering dendrogram (i.e., members of the SLC17,
581 SLC18, SLC37 families as well as MFSD8, MFSD9, MFSD10, MFSD14A-C and SLC22A18 proteins, Fig 3). As
582 expected, the phylogenetic tree gives a better separation of these very similar proteins than the HMM
583 fingerprint-based dendrogram, and the branch support values suggest a clear separation of the SLC17, SLC18
584 and SLC37 families. In addition, the phylogenetic tree highlights that the atypical SLC proteins MFSD8,
585 MFSD9, MFSD10, MFSD14A-C as well as SLC22A18, while being more divergent, are likely to have evolved
586 from a single common ancestor. The relationship between the MFSD9, MFSD10 and MFSD14A-B proteins
587 also agrees with earlier studies [1,2]. Similarly, the evolutionary dendrogram created using all 7 organisms in
588 our study for the SLC18 family (S1 File) suggests that MFSD9, MFSD10, MFSD14A-B and SLC22A18 likely
589 share a common evolutionary origin and are thus more closely related to each other than to SLC18 proteins,
590 while MFSD8 is more distantly related. Further studies may be required to elucidate the particular
591 evolutionary relationship between these proteins.

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

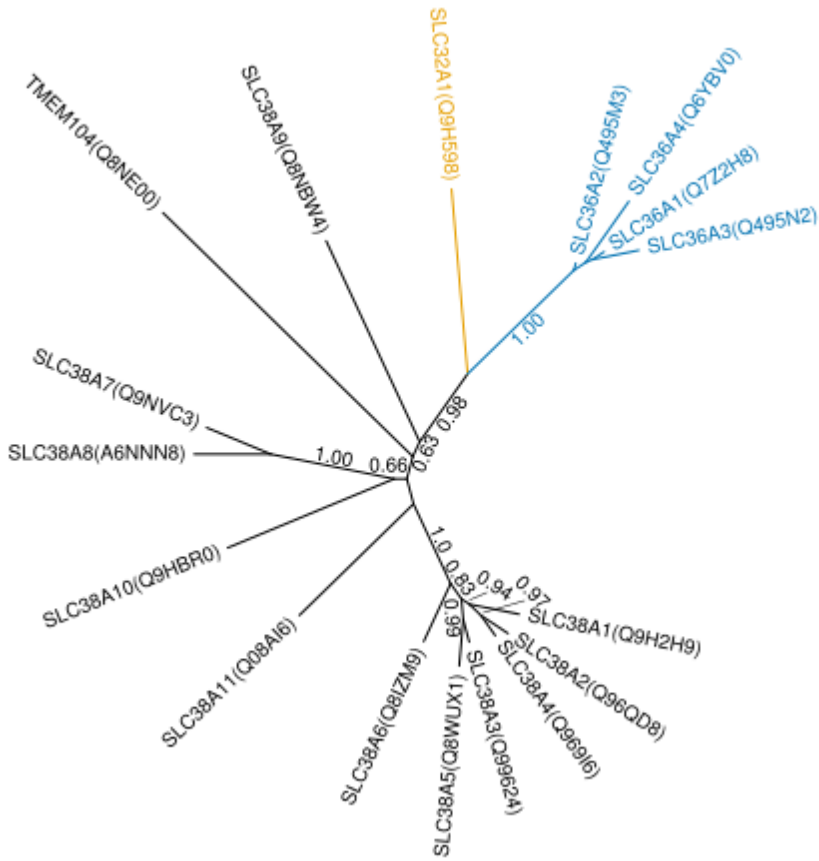


Interestingly, **MFSD3** has clustered together with the SLC33A1 protein in our HMM fingerprint-based clustering analysis. Indeed, the “Acatn” (Acetyl-coenzyme A transporter 1) Pfam domain is present in MFSD3, albeit with a relatively low score, but with significant e-value (5.6e-12). Sequence alignment 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

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

618 **Figure 4. Phylogenetic tree of proteins clustering in the SLC38-36-32 families and TMEM104.** Branch
619 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 model 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

633 been linked to bovine milk cholesterol levels [94], the hypothalamic-gonadal axis and testosterone response
634 [95]. It is also highly expressed in the hypothalamus [95].

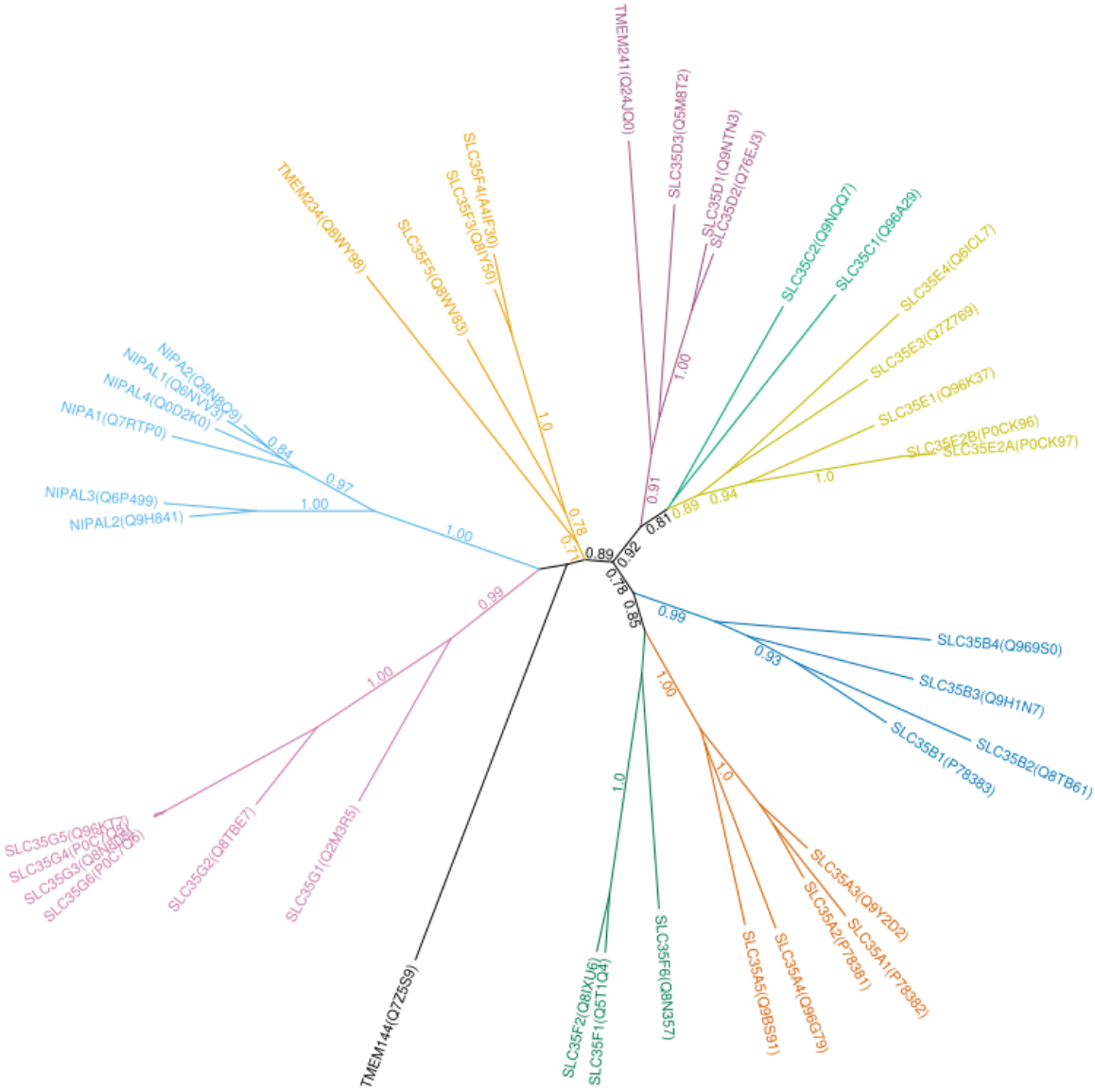
635 **TMEM234** is classified in the TCDB #2.A.7.32 family and also contains a corresponding "TMEM234" Pfam
636 domain, which is a member of the "DMT" clan of Pfam domains as well. The physiological role of TMEM234
637 is not known. However, in zebrafish, its homolog might play a role in the formation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

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

646 Due to the sequence diversity of the SLC35 family, we were interested in the relationships between
647 individual proteins. To this end, we have built a phylogenetic tree of human proteins that showed similarity
648 to existing SLC35 transporters (Fig 5). In the tree, most SLC35 subfamilies could be resolved as a single clade,
649 while TMEM241 and TMEM234 form clades with SLC35D and SLC35F3-5 proteins with a support value of
650 0.91 and 0.71, respectively. The relationship of TMEM241 with the SLC35D subfamily is also supported by
651 our HMM fingerprint-based clustering results. TMEM241 shows 12.0-21.4% sequence identity with SLC35D
652 proteins. In contrast, our phylogenetic tree with SLC35 proteins from all 7 organisms (S1 File) indicated that
653 TMEM241 is most closely related to SLC35E4. On the other hand, TMEM234 only weakly associated with
654 SLC35F proteins, with sequence identities 3.6-9.2%. TMEM144 appears to be only distantly related to SLC35
655 proteins. The elucidation of evolutionary relationships between proteins in the SLC35 thus likely requires
656 further investigation.

657 **Figure 5. Phylogenetic tree of proteins with HMM fingerprints overlapping with the SLC35 family.** Branch

658 support values ≥ 0.7 are shown. The different subfamilies are colored in various colors.



659

660

661 ## Others

662 **GPR155** is an enigmatic protein that seems to be a concatenation of a membrane transporter domain (Pfam:

663 "Mem_trans") and a G-protein coupled receptor (GPCR) domain, which might be the reason why it is

664 annotated as a GPCR. The membrane transporter part seems to be most similar to TCDB #2.A.69.3 subfamily

665 proteins, which are annotated as malate/malonate transporters in the Auxin Efflux Carrier (AEC) family

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

668 The “Mem_trans” domain is only present in GPR155 from all human proteins analyzed, and matches the first
669 10 TMHs of the protein in a 5+5 arrangement. In our structural search, the second half of this transporter
670 domain (TMHs 6-10) of GPR155 exhibits similarity to the N-terminal half of sodium/bile transporters of the
671 AsbT fold (SLC10 family). On the other hand, the last 7 TMHs of GPR155 (TMHs 11-17) indeed show
672 similarity to GPCR-fold (7-TM) proteins with known structure, with highest similarity to structures of the
673 human Smoothened receptor homolog (PDB ID: 6OT0). Interestingly, in our search, the N-terminal half of the
674 transporter domain of GPR155 did not show any structural homologues.

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

682 The **RFT1** protein was originally thought to be a scramblase of lipid-linked origosaccharides [106]. However,
683 these molecules have at least 12-14 sugar moieties, so given their size, it is unlikely that a single transporter
684 could catalyze their flipping. Later studies refuted the scramblase concept and suggested instead that RFT1
685 could serve as an accessory protein to a flippase, but would not act as a flippase itself [107–109].

686 Nevertheless, the corresponding “Rft1” Pfam model shows similarity to multidrug and toxic compound
687 extrusion (MATE) transporters (SLC47 family) and belongs to a clan of Pfam domains (“MviN_MATE”) that
688 contains transporters as well. In line with this, human RFT1 showed significant similarity to MATE
689 transporters in our search for structural homologs, indicating likely structural similarity. Some members of

690 the corresponding TC# 2.A.66.3 subfamily also contain weak hits of the "MatE" Pfam domain. Thus, while
691 RFT1 shows similarity to existing transporters, its biological function is still unclear.

692 The C-terminal half of the **TMEM245** protein shows weak similarity to TC# 2.A.86 proteins (Autoinducer-2
693 Exporter/AI-2E family), which contain both small-molecule exporters [110,111] as well as Na^+/H^+ antiporter
694 proteins [112–114]. Accordingly, TMEM245 also has weak similarity to the corresponding Pfam model ("AI-
695 2E_transport").

696 The HMMs of TC# 2.A.86 and #2.A.86.1 match residues 444-866 of human TMEM245, which are the last 6
697 TMHs according to UniProt predictions. The last 5 TMHs are separated from the previous ones by a slightly
698 larger loop. This architecture is similar to the 3+5 arrangement of the previously described bacterial
699 $Na^+(Li^+)/H^+$ antiporter TC# 2.A.86.1.14 according to UniProt predictions (accession code: NLHAP_HALAA). This
700 bacterial protein also matched the full-length "AI-2E_transport" domain from Pfam, while only the last 5
701 TMHs of TMEM245 match with C-terminal region of "AI-2E_transport". The human TMEM245 protein
702 contains 14 TMHs in total according to UniProt predictions. Thus, TMEM245 might have a transporter-like
703 domain at the C-terminus. In terms of the structure, we have not found any similarities to proteins with
704 known structure. Therefore both TMEM245 and bacterial exporters and antiporters in the TC# 2.A.86 family
705 are likely to have a yet uncharacterized tertiary structure. Functionally, the TMEM245 protein also remains
706 elusive.

707 The **TMEM41A**, **TMEM41B**, and **TMEM64** proteins clustered to the same family in our results. These are the
708 only proteins in human that show any similarity to the "SNARE_assoc" Pfam domain, as well as to the TCDB
709 family #9.B.27. While no protein with this domain or from this family has direct evidence for transport
710 activity, Pfam reports SCOOP-based similarity [31] of the "SNARE_assoc" domain with "Sm_multidrug_ex",
711 which is a domain encoding transporter proteins. Some members of the family in the TCDB have been
712 proposed to be "cation:proton importers" (#9.B.27.2.2) or "selenite transport proteins" (#9.B.27.2.3).

713 The most characterized member of the human protein family is TMEM41B. Interestingly, a recent study
714 reported a putative structure generated *ab initio* using evolutionary covariance-derived information [115].

715 Strikingly, this structural model shows features reminiscent of secondary transporters, such as a tandem
716 internal repeat with two-fold rotational symmetry, and the authors suggest a H^+ antiporter activity as a
717 mechanism of transport [115].

718 While the exact function of TMEM41B is still unclear, it forms a complex with vacuole membrane protein 1
719 (VMP1), also harboring the "SNARE_assoc" domain, and both are required for autophagosome formation
720 [116,117]. Tmem41b localized to mitochondria-associated ER membranes [118–120]. Interestingly,
721 TMEM41B seems to be an absolutely required factor for SARS-CoV-2 [121], and probably also flaviviral [122]
722 infection, possibly by facilitating a membrane curvature that is beneficial for viral replication [122].

723 The proteins **TMEM184A**, **TMEM184B**, **TMEM184C** clustered together with **SLC51A** (family of transporters of
724 steroid-derived molecules) in our analysis. While human TMEM184B and SLC51A are included in the TCDB as
725 members of family #2.A.82, TMEM184A and C are not. Independently, the "Solute_trans_a" Pfam model was
726 found in all four proteins with high scores and significance, but not in other human proteins. Therefore, it is
727 likely that the four proteins, TMEM184A-C and SLC51A, are homologous. In spite of this, sequence identity
728 between TMEM184 proteins and SLC51A is low (12.3%-13.6%), but moderate among TMEM184 proteins
729 (26.5%-62.0%). All four proteins are predicted to harbor 7 TM helices according to UniProt, yet our search
730 has found no similar proteins with known structures.

731 TMEM184A was identified as a heparin receptor in vascular cells [123], but no transport activity has been
732 reported. Interestingly, while SLC51A is known to function as a bile acid transporter [124–126], TMEM184B
733 has been proposed to be responsible for ibuprofen uptake [127]. This is interesting in view of the partial
734 chemical similarity between steroid acids and ibuprofen, both harboring a hydrophobic hydrocarbon part
735 and a carboxyl moiety. TMEM184C resides in a genetic locus that has been suggested to be responsible for
736 the pathogenicity of X-linked congenital hypertrichosis syndrome [128], but no transport activity has been
737 suggested.

738 ## Putative transporters

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

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

754 A family of 4 lysosomal-associated transmembrane proteins (**LAPTM4A**, **LAPTM4B**, **LAPTM5** and sequence
755 **B4E0C1**) turned up in our search, corresponding to the TCDB subfamily #2.A.74.1 and Pfam model "Mtp"
756 (mouse transporter protein). The family also includes an uncharacterized transcript with the UniProt
757 accession "B4E0C1". Originally, the mouse transporter protein (Mtp, ortholog of LAPTM4A) was
758 characterized as a transporter mediating the transport of nucleosides and nucleobases between the
759 cytoplasm and intracellular compartments [138], and was later also associated with multidrug-resistance
760 (MDR) in yeast, where its expression changed the subcellular compartmentalization of a heterogenous group
761 of compounds [139,140]. LAPTM4A was shown to be involved in glycosylation and glycolipid regulation
762 [141,142]. All three proteins seem to be lysosomal [138,143,144]. Nevertheless, these proteins appear to
763 interact with other characterized transporters, such as SLC22A2 (hOCT2), SLC7A5/SLC3A2 (LAT1/4F2hc) and

764 MDR-related ABC transporters [145–147]. But it has been claimed that they are not *per se* transporters, but
765 rather regulatory factors, either assisting the localization and targeting or the function of other transporters
766 [147]. Interestingly, the transcript "B4E0C1" appears to have 4 TMHs at its N-terminus, which is identical to
767 human LAPTM5 apart from a ~40-amino acid insertion between TMH3 and TMH4. This region also shows
768 significant similarity to both the "Mtp" Pfam model and the TC# 2.A.74.1 subfamily. However, the C-terminal
769 region of the transcript is identical to the C-terminal segment of "actin filament-associated protein 1-like 1"
770 protein (UniProt accession Q8TED9). We did not find any similar fusion sequences in the other organisms we
771 analyzed. The "Mtp" Pfam domain, which is the hallmark of the family, belongs to the "Tetraspannin" Pfam
772 clan, which has no other domains with annotated transporter function and no similarity to existing
773 transporters. Structural information on "Tetraspannin" proteins is also not available. In summary, the
774 transport function of these proteins requires further investigation.

775 Our search identified four proteins in human (**LMBR1**, **LMBR1L**, **LMBD1/LMBRD1**, **LMBRD2**) bearing the
776 "LMBR1" Pfam domain, which clustered into two families in our analysis. These proteins correspond to TCDB
777 family #9.A.54. The LMBD1 protein, encoded by the LMBRD1 gene, was suggested to function as a vitamin
778 B12 (cobalamin) transporter, exporting vitamin B12 from the lysosomes into the cytoplasm [148]. However,
779 it was later shown that LMBD1 actually interacts with ABCD4 and assists in its lysosomal trafficking [149],
780 and that ABCD4 transports vitamin B12 even in the absence of LMBD1 [150]. Therefore, it is likely that
781 LMBD1 itself is not a vitamin B12 transporter. LMBD1 was originally coined as having "significant homology"
782 to lipocalin membrane receptors [148], and indeed the LIMR (lipocalin-1-interacting membrane receptor)
783 protein, encoded by the LMBR1L gene, is responsible for binding lipocalin 1 (LCN-1) with high affinity
784 [151,152]. LMBRD2 was proposed to be a regulator of β2-adrenoceptor signaling [153], while the first
785 protein identified in the family, LMBR1, was associated with polydactyly and limb malformations [154,155].
786 However, its physiological role is still elusive. The proteins seem to contain 9 TMHs in a 5+4 arrangement
787 according to UniProt predictions, but the tertiary structure of the proteins is still unknown, and no homologs
788 with a known structure were found in our search.

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

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

811 **TMEM205** is a 4-TMH protein according to UniProt annotations, which was linked to cisplatin resistance
812 [170]. The protein is expressed mostly in liver, pancreas and adrenal glands, and is present on the plasma
813 membrane [170]. TMEM205-mediated resistance was shown to be selective towards platinum-based drugs,
814 such as cisplatin and oxaliplatin, but not carboplatin [171]. While structural information about the protein is

815 not available, mutagenesis studies of TMEM205 showed that mutating sulfur-containing residues, especially
816 in TMH2 and TMH4, diminishes the effect of cisplatin resistance [171]. Nevertheless, neither the biological
817 function nor the physiological substrates of TMEM205 are known.

818 Transporters with hydrophobic substrates

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

829 Conclusions

830 Our study represents the first systematic correlation of the SLC and TCDB nomenclature schemes. Many of
831 the transport proteins discovered in our search are underexplored and there is limited information about
832 them, although they often have important physiological roles and/or potentially represent new therapeutic
833 targets. Even with proteins that have been studied for their physiological involvement, it was often not taken
834 into account that they could have a transport function. Numerous proteins uncovered in our search have
835 similarity to proteins with transport function in other organisms, but their physiological substrates remain
836 unknown. These proteins will be interesting targets for deorphanization studies in order to reveal their
837 natural substrates. In our work, we also highlight proteins for which transport activity has been
838 controversial, and more specific analyses are required to clarify their biological function. In addition, our
839 search reveals new SLC-like proteins that have no structural information. This hinders a deeper

840 understanding of their transport mechanism. Future structure determinations would be of crucial
841 importance to accelerate validation of the identified proteins. The combination of all these efforts would
842 greatly facilitate the completion of the SLC-ome in human cells. Thus, our study points out important
843 directions in which future studies could help resolve the lack of information about SLC transporters, which
844 will help unlock their therapeutic potential.

845 Acknowledgements

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

852 Methods

853 HMM building for TCDB families and subfamilies

854 Sequences of selected TCDB families and subfamilies were aligned using PSI-Coffee 11.00 [172] and NCBI
855 BLAST+ 2.6.0 [173] using the "nr" BLAST database of 2018-02-12. PSI-Coffee was run with BLAST mode
856 "LOCAL" and otherwise with default options. Altogether, 4221 different sequences from the TCDB were
857 present in the subfamily and family alignments, and the profiles used for alignments contained 1-1070
858 sequences. In total, 130 sequences from the TCDB returned zero hits from the "nr" database in the iterative
859 BLAST searches, meaning their profiles just contained the query sequence itself. PSI-Coffee uses the profiles
860 to guide the generation of multiple alignments of the query sequences within each subfamily and family. The
861 alignments were subsequently turned into HMMs using the "hmmbuild" command of HMMER 3.1b2 [30]
862 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

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

908 Structural homolog search

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

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

916 Phylogenetic trees

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

925 Data Availability

926 Scripts and data generated during the preparation of the manuscript are available at
927 https://github.com/bioparadigms/SLCAtlas. In addition, information resulting from this work is also available
928 in a browsable format at https://www.bioparadigms.org/.

929 Supplementary information

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

933 **S2 Table. False positive hits or incorrectly annotated human sequences found in our search.** Gene symbol
934 and protein name shown based on UniProt data. Most similar (highest scoring) TCDB families/subfamilies
935 and Pfam families are shown, and reason for being false positive are indicated.

936 **S3 Table. Table of existing (classic) SLC transporters, indicating the correlation with TCDB**
937 **families/subfamilies, Pfam families, PDB structures and structural folds.** Gene symbol and protein name
938 based on UniProt information are shown. Most similar (highest-scoring) TCDB families and subfamilies, as
939 well as Pfam families are shown for each protein. The PDB structure from the pdb70 dataset that is most

940 similar (highest-scoring) to each protein is shown, along with its fold family assignment by us, partly based
941 on TCDB family names.

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

951 **S4 Table. Non-classic " SLC-like proteins of Table 3 found in our search, extended with additional**
952 **information.** Protein sequence identifiers, most similar (highest-scoring) PDB structure from the pdb70
953 dataset, chemical identifiers, PubMed links to references, and comments have been included. The table also
954 contains information about whether individual SLC-likeness criteria (numbered 1-8) have been satisfied or
955 violated.

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

References

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.