Characteristics of 29 novel atypical solute carriers of major facilitator superfamily type: evolutionary conservation, predicted structure and neuronal co-expression

Emelie Perland, Sonchita Bagchi, Axel Klaessson and Robert Fredriksson

Article citation details
Open Biol. 7: 170142. http://dx.doi.org/10.1098/rsob.170142

Review timeline
Original submission: 12 June 2017
Revised submission: 28 July 2017
Final acceptance: 31 July 2017

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSOB-17-0142.R0 (Original submission)

Review form: Reviewer 1

Recommendation
Accept with minor revision (please list in comments)

Are each of the following suitable for general readers?

a) Title
   No

b) Summary
   Yes

c) Introduction
   Yes
Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Is it clear how to make all supporting data available?
Yes

Is the supplementary material necessary; and if so is it adequate and clear?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
This is a very interesting manuscript from a highly-respected transporter research group. The results and conclusions will be of interest to a broad spectrum of readers, from evolutionary biologists through molecular physiologists to clinicians and pharmacologists studying human diseases and their treatments. On the other hand, at present the text is rather dense and abbreviation-packed and close inspection reveals some important points that require clarification, as listed below in broadly descending order of importance.

1. Classifications and abbreviations
   (a) L.140, L275 – 276 and elsewhere. Does your analysis actually confirm the assignment of these “orphan” atypical SLC transporters to the MFS clan, given they have less than 20% AA identity to any other MFS protein? Might they have a closer relationship to other CLS clans, noting that some of the analyses only include MFS transporters (see Figure 2)? This needs to be confirmed and stated explicitly. (b) Is there any underlying logic to the ordering of the proposed AMTF number classification? Would it not be simpler to propose additional SLC numbering instead of a whole new classification nomenclature for these transporters (granted that this might need to be confirmed by the HUGO database curators)? (c) The text switches between atypical SLC, MFSD and AMTF classifications to a confusing extent as the manuscript unfolds:- I suggest some editing to reduce and / or clarify the rationale behind such switches. (d) A general list of abbreviations (including e.g. MFSD, AMTF, Pfam, PLA) would be helpful. (e) L.13 and elsewhere. Given likely broad readership, it should be stressed that the SLC classification is based on the human genome. The Transporter Classification Database identifiers for each transporter should also be shown at some point, especially given that inter-species analyses are included in various Figures and Tables. Similarly, a clearer indication of protein vs cDNA sequence comparisons in text and Figure legends should be made (e.g. L.143-144, amino acid identity?). (f) Supplementary Table 1. MFS matrix is actually “Atypical MFS matrix” and SLC matrix is actually “MFS matrix”. (g) L59. What are the other 2 atypical SLC? In which clans do they cluster?

2. Phylogeny and evolutionary origins
   (a) The MFS clan includes only about 1/3 of SLC transporters, yet almost all atypical SLCs. Might that be related to factors such as evolutionary age / structural advantages that favour evolutionary divergence of function? On the other hand, there are remarkably few atypical SLC in yeast (see Table 2). Might this reflect comparatively recent evolution? Unfortunately it appears that bacteria were not included in the phylogenetic analyses so any prokaryotic origins of atypical SLCs cannot be distinguished. (b) L.1. Title states HUMAN but the co-expression and regulation studies use mouse tissues and cell-lines (c) L.102 “several animal kingdoms”? There is only one. Similarly L.51. states “Several organisms”, which might be better as “widely distributed through the animal kingdom”. (d) L.38. “Passive diffusion”? I presume this refers to uniprot. L.75. CLN3 has 6 predicted TMS domains, whereas P.104 states 11 TMS. This is clarified in the Discussion but perhaps needs raising earlier to avoid confusion. L.119. Type of anaesthetic
used should be included at this point. I suggest moving relevant section in L204 – 206 to here. L.168 “grammatical constrains”? Is this referring to steric constraints? L.189-191. Repetitious L.192-194. Appears to be repetition of bootstrapping information L.219. Clarify as SELECTED antibody pairs (referring to Table 3). Text gives the initial impression that all pairs (29 x 29?) were tested. L.321 “all”? Not all, see also Figure 1 legend. L.345. MFSD6 is included in the analysis, so I presume it is detected. L.382. Is it 28 or 29? L.406 “nod” should be “node” I think? L.409 nm not um I think. L.469 THE most common structure? For different clans? Please clarify. Figure 2. Possibly add the proposed ATMF classification as an outer ring to this web Figure 3. Legend. Amount or number of proteins / variants?

Review form: Reviewer 2

Recommendation
Accept with minor revision (please list in comments)

Are each of the following suitable for general readers?

a) Title
   Yes

b) Summary
   Yes

c) Introduction
   Yes

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Is it clear how to make all supporting data available?
Yes

Is the supplementary material necessary; and if so is it adequate and clear?
Not Applicable

Do you have any ethical concerns with this paper?
No

Comments to the Author
This is an important documentation of a group of putative transporters. The function is hard to predict at this stage, particularly in the MFS family, which is very diverse. Line 182 onwards: Please confirm that the identified transcript are full length. Pseudogenes sometimes have spurious transcripts, it would be important to clarify that these members a fully expressed proteins. Discussion: It might be worthwhile to look at catalytically important residues in the MFS fold, if the alignments are precise enough. This could help with identifying possible substrates.
Decision letter (RSOB-17-0142)

24-Jul-2017

Dear Dr Perland,

We are pleased to inform you that your manuscript RSOB-17-0142 entitled "Characteristics of 29 novel human atypical SLCs of MFS type: evolutionary conservation, predicted structure and neuronal co-expression" has been accepted by the Editor for publication in Open Biology. The reviewer(s) have recommended publication, but also suggest some minor revisions to your manuscript. Therefore, we invite you to respond to the reviewer(s)' comments and revise your manuscript.

Please submit the revised version of your manuscript within 14 days. If you do not think you will be able to meet this date please let us know immediately and we can extend this deadline for you.

To revise your manuscript, log into https://mc.manuscriptcentral.com/rsob and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.

You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, please revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you will be able to respond to the comments made by the referee(s) and upload a file "Response to Referees" in "Section 6 - File Upload". You can use this to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the referee(s).

Please see our detailed instructions for revision requirements https://royalsociety.org/journals/authors/author-guidelines/.

Before uploading your revised files please make sure that you have:

1) A text file of the manuscript (doc, txt, rtf or tex), including the references, tables (including captions) and figure captions. Please remove any tracked changes from the text before submission. PDF files are not an accepted format for the "Main Document".

2) A separate electronic file of each figure (tiff, EPS or print-quality PDF preferred). The format should be produced directly from original creation package, or original software format. Please note that PowerPoint files are not accepted.

3) Electronic supplementary material: this should be contained in a separate file from the main text and meet our ESM criteria (see http://royalsocietypublishing.org/instructions-authors#question5). All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rsob.2016[last 4 digits of e.g. 10.1098/rsob.20160049].
4) A media summary: a short non-technical summary (up to 100 words) of the key findings/importance of your manuscript. Please try to write in simple English, avoid jargon, explain the importance of the topic, outline the main implications and describe why this topic is newsworthy.

Images
We require suitable relevant images to appear alongside published articles. Do you have an image we could use? Images should have a resolution of at least 300 dpi, if possible.

Data-Sharing
It is a condition of publication that data supporting your paper are made available. Data should be made available either in the electronic supplementary material or through an appropriate repository. Details of how to access data should be included in your paper. Please see http://royalsocietypublishing.org/site/authors/policy.xhtml#question6 for more details.

Data accessibility section
To ensure archived data are available to readers, authors should include a ‘data accessibility’ section immediately after the acknowledgements section. This should list the database and accession number for all data from the article that has been made publicly available, for instance:

• DNA sequences: Genbank accessions F234391-F234402
• Phylogenetic data: TreeBASE accession number S9123
• Final DNA sequence assembly uploaded as online supplemental material
• Climate data and MaxEnt input files: Dryad doi:10.5521/dryad.12311

Once again, thank you for submitting your manuscript to Open Biology, we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Sincerely,

The Open Biology Team
mailto:openbiology@royalsociety.org

Author's Response to Decision Letter for (RSOB-170142)
See Appendix A.

Decision letter (RSOB-17-0142.R1)

31-Jul-2017

Dear Mrs Perland

We are pleased to inform you that your manuscript entitled "Characteristics of 29 novel atypical SLCs of MFS type: evolutionary conservation, predicted structure and neuronal co-expression" has been accepted by the Editor for publication in Open Biology.
You can expect to receive a proof of your article from our Production office within approx. 5 working days. Please let us know if you are likely to be away from e-mail contact during this period. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

Article processing charge
Please note that the article processing charge is immediately payable. A separate email will be sent out shortly to confirm the charge due. The preferred payment method is by credit card; however, other payment options are available.

Thank you for your fine contribution. On behalf of the Editors of Open Biology, we look forward to your continued contributions to the journal.

Sincerely,

The Open Biology Team
mailto: openbiology@royalsociety.org
Appendix A

Response to reviewers

Thank you for your kind and very helpful comments. We have adjusted the manuscript according to all suggested changes, to make the manuscript as good as possible. We must have had different line numbering then you, so we have sometimes deduced what parts you referred to and answered accordingly. But we think we found it all. For specific details, see further down. The revised submission is adjusted according to Open Biology’s requirements.

Referee: 1

1. Classifications and abbreviations

(a) Does your analysis actually confirm the assignment of these “orphan” atypical SLC transporters to the MFS clan, given they have less than 20% AA identity to any other MFS protein? Might they have a closer relationship to other CLS clans, noting that some of the analyses only include MFS transporters (see Figure 2)? This needs to be confirmed and stated explicitly.

The 20% aa identity also applies to the SLC families named by HGNC, and not to the Pfam clans. If we compare the sequences of the already known SLCs of MFS type, there is low identity as well. This can thus not be used as a parameter for the clan belonging. We found the atypical SLCs (see previous publication [1]) by HMM searches, where the models were built on SLC proteins belonging to the MFS clan (we also searched for proteins from the APC clan). The proteins studied here were found by this search. They are also listed as MFS proteins by Pfam. So their belonging to MFS is recognized, and we say that they are also SLC proteins. It is stated already that they were found in the MFS Pfam clan, and we have clarified it by adding the following to the introduction:

These proteins were identified as possible SLCs by searching the human proteome using Hidden Markov models (HMM) composed of SLC sequences originating from the MFS Pfam clan [1].

(b) Is there any underlying logic to the ordering of the proposed AMTF number classification? Would it not be simpler to propose additional SLC numbering instead of a whole new classification nomenclature for these transporters (granted that this might need to be confirmed by the HUGO database curators)?

When we started with the manuscript we had the idea to give the families SLC names. However, we finally decided to give them AMTF naming since we have not proven that they actually are transporters when it comes to their functions. HGNC, who provide the SLC names, are usually reluctant to assess names to proteins without known function. Some of the
proteins have been studied and are indeed transporters (e.g. MFSD2A, 4B, 5 and 10) whereas others have more debated functions, as for the SV2 proteins. It could exist MFS proteins in humans that are not functionally transporters, and if we suggested using the SLC families we think it will be assumed as confirmed that they transport molecules, even though we do not state that. To give them the AMTF naming show they are related with each other, but not to forget that their functions still need to be determined. However, whenever their functions are established, they should be given family names according to the SLC root system.

We have added some text about this in the manuscript to clarify it.

*The atypical SLCs are probably SLC proteins, but most are still orphan regarding function. Therefore they were divided into AMTF families instead of using the SLC nomenclature. In this way it is clear they are possible transporters, but that their function remains to be elucidated. Whenever their functions are determined, they could however be renamed according to the SLC root system.*

(c) The text switches between atypical SLC, MFSD and AMTF classifications to a confusing extent as the manuscript unfolds:- I suggest some editing to reduce and / or clarify the rationale behind such switches.

Sorry for the confusion, we have clarified this throughout the manuscript.

(d) A general list of abbreviations (including e.g. MFSD, AMTF, Pfam, PLA) would be helpful.

A very good idea! We have done so.

(e) L.13 and elsewhere. Given likely broad readership, it should be stressed that the SLC classification is based on the human genome. The Transporter Classification Database identifiers for each transporter should also be shown at some point, especially given that interspecies analyses are included in various Figures and Tables.

We have added the TC numbers to the summary table. However, some of the proteins do not have a TC-number, which is why we had used the HGNC identifier.

Similarly, a clearer indication of protein vs cDNA sequence comparisons in text and Figure legends should be made (e.g. L.143-144, amino acid identity?).

In general, when we discuss proteins we use capitalised letters, whereas RNA is in italic in lowercase.

(f) Supplementary Table 1. MFS matrix is actually “Atypical MFS matrix” and SLC matrix is actually “MFS matrix”.

This has been updated.
TMEM104 belongs to the APC clan and OCA2 cluster with the IT clans, which now is stated in the introduction.

2. Phylogeny and evolutionary origins
(a) The MFS clan includes only about 1/3 of SLC transporters, yet almost all atypical SLCs. Might that be related to factors such as evolutionary age / structural advantages that favour evolutionary divergence of function? On the other hand, there are remarkably few atypical SLC in yeast (see Table 2). Might this reflect comparatively recent evolution? Unfortunately it appears that bacteria were not included in the phylogenetic analyses so any prokaryotic origins of atypical SLCs cannot be distinguished.

The SLCs of MFS type is the largest group of phylogenetically related SLCs, and it is probably because the MFS family is so large and diverse. It is relatively easy to use MFS proteins to search for related ones, but there are probably more atypical SLCs in the proteome apart from the known. However, it is hard to search for them. We can build reliable HMM for SLCs from both the MFS and APC clan since there are enough known members to include in the models. However, in addition to these, it is only the CPA/AT and DMT clan that includes more than one known family, meaning we cannot build HMM models to search for more atypical SLCs.

If we are allowed to guess, we would say the proteins emerged late to exert specific functions, as several of them are not found in all species investigated. But yes, more species must be studied to make this as a conclusion.

(b) L.1. Title states HUMAN but the co-expression and regulation studies use mouse tissues and cell-lines

We removed Human from the title to better describe the content.

(c) L.102 “several animal kingdoms”? There is only one. Similarly L.51. states “Several organisms”, which might be better as “widely distributed through the animal kingdom”. Correct, we have changed this.

3. Specific minor points
Table 4 / Figure 7. Please highlight table rows with statistically significant changes. The differences seem relatively small compared to archetypal AA adaptive response genes. Do any of the regulated genes include AA response elements? Are the SLC mentioned in L.93-94 altered in your data?
We have highlighted the significant changes in the table. We have not studied the AA response elements as we guess most are not amino acid transporters. I think I have different line numbers than you, so I cannot recall which SLCs you refer to. But most genes were affected by the diet, even though the changes were small.

The divergence between responses of Cln3 and Unc93a “duplicates” needs to be discussed. L.249 implies gene duplication, but this is not entirely clear.

This is probably not because of gene duplication but rather because these genes are present under different accession numbers in the database used to define the genes on the chip. These two occurrences of Unc93 and Cln3 represent different splice variants present under different (unique) accession numbers. We have not analysed this in detail, but because the two spots (or rather group of spots) on the array generally follows the same trend in changes of expression we assume that both spice variants are regulated the same way. This is now included in the manuscript.

L.38. “Passive diffusion”? I presume this refers to uniport.

Yes, have changed this.

L.75. CLN3 has 6 predicted TMS domains, whereas P.104 states 11 TMS. This is clarified in the Discussion but perhaps needs raising earlier to avoid confusion.

We added a sentence about this in the result section as well.

L.119. Type of anaesthetic used should be included at this point. I suggest moving relevant section in L204 – 206 to here.

Have added this.

L.168 “grammatical constrains”? Is this referring to steric constraints?

No, this is (in general terms) a way to formally define rules for an alphabet. In this context it refers to the amino acids used in the HMM models, to be able to us these strings of amino acid symbols to calculate probabilities. It is not necessary to include this information in the context of this paper so we have removed the section to avoid confusion.

L.189-191. Repetitious

Corrected.
L.192-194. Appears to be repetition of bootstrapping information

We have removed the repetition.

L.219. Clarify as SELECTED antibody pairs (referring to Table 3). Text gives the initial impression that all pairs (29 x 29?) were tested.

Corrected, now we state they were selected.

L.321 “all”? Not all, see also Figure 1 legend.

This has been clarified.

L.345. MFSD6 is included in the analysis, so I presume it is detected.

Yes, it was MFSD6L that was not found.

L.382. Is it 28 or 29?

29.

L.406 “nod” should be “node” I think?

Correct, have been updated.

L.409 nm not um I think.

Correct again.

L.469 THE most common structure? For different clans? Please clarify

Has been clarified.

Figure 2. Possibly add the proposed ATMF classification as an outer ring to this web

We had this in the first draft of the figure, but it turned out very messy. So we restricted the information to make the figure readable.

Figure 3. Legend. Amount or number of proteins / variants?

Corrected.
Referee: 2

Line 182 onwards: Please confirm that the identified transcripts are full length. Pseudogenes sometimes have spurious transcripts, it would be important to clarify that these members are fully expressed proteins.

We use the ENSEMBLE “all protein” dataset as the source for our mining, which is supposed to, according to the ENSEMBLE definition, to not contain pseudogenes. In additions we tried to assure this by:

- Making sure that all transcripts translates into an ORF
- That all part of the predicted ORF is aligning to the most similar protein
- That the longest transcript is used
- That all genes used in the analysis are expressed, i.e. are represented by ESTs or cDNA clones.

If a sequence full fills all these criteria it is for our analysis considered a non-pseudogene.

One additional control we make is to inspect the trees for unexpectedly long branches. A pseudogene usually have an altered evolutionary pressure resulting in sequence evolution that strongly diverges from the rest of the sequences in the tree, which results in long branches in the ML/Bayesian analysis.

Discussion: It might be worthwhile to look at catalytically important residues in the MFS fold, if the alignments are precise enough. This could help with identifying possible substrates.

This is a very good suggestion; however, there is little similarity between the sequences so the overall alignments are not precise enough. It will probably be better to do this when focusing on fewer proteins at a time.

References