ABC-finder: A containerized web server for the identification and topology prediction of ABC proteins

Poonam Vishwakarma[1#], Naveen Kumar Meena[1#], Rajendra Prasad[2]*, Andrew M. Lynn[1]* and Atanu Banerjee[2]*

[1]*School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India*

[2]*Amity Institute of Biotechnology and Amity Institute of Integrative Sciences and Health, Amity University Haryana, Gurugram, India*

 # contributed equally

**corresponding authors:*

Atanu Banerjee; abanerjee1@ggn.amity.edu

Andrew M. Lynn; andrew@jnu.ac.in

Rajendra Prasad; rprasad@ggn.amity.edu

ABSTRACT

In view of the multiple clinical and physiological implications of ABC transporter proteins, there is a considerable interest among researchers to characterize them functionally. However, such characterizations are based on the premise that ABC proteins are accurately identified in the genome, and their topology is correctly predicted. With this objective, we have developed ABC-finder, i.e., a Docker-based package for the identification of ABC proteins in all organisms, and visualization of the topology of ABC proteins using an interactive web browser. ABC-finder is built and deployed in a Linux container, making it scalable for many concurrent users on our servers and enabling users to download and run it locally. Overall, ABC-finder is a convenient, portable, and platform-independent tool for the identification and topology prediction of ABC proteins.

ABC-finder is accessible at http://abc-finder.osdd.jnu.ac.in

KEYWORDS

ABC proteins, Nucleotide-binding domains, Transmembrane domains, Docker, Reproducibility

RUNNING TITLE

ABC-finder web server

1 INTRODUCTION

ABC proteins comprise one of the largest and most important protein families, the majority of proteins being involved in active transport [1]. A typical ABC transporter is composed of a pair of nucleotide-binding domains (NBDs) and transmembrane domains (TMDs) [2]. While the NBDs are involved in fueling the transport process by means of ATP hydrolysis, the TMDs are responsible for substrate recognition and form the translocation channel [3]. NBDs have several conserved motifs, namely, Walker A, Walker B, Signature sequence (C-motif), H-loop, D-loop, and Q-loop [4]. Contrarily, the TMDs display poor conservation across different subfamilies. HUGO Gene Nomenclature Committee divided the ABC superfamily into 7 subfamilies from ABCA to ABCG. This classification was later extended for non-mammalian proteins with the addition of the ABCH subfamily found in insects [5] and fishes [6,7], and the ABCI subfamily found in plants [8]. ABC transporters can function both as importers as well as exporters, however, the former is restricted only to bacteria and plants and mediate the uptake of various nutrients, micronutrients, and phytohormones, etc [9]. Furthermore, the Energy-Coupling Factor (ECF) transporters have been recently categorized as Type III ABC importers [9]. On the other hand, the ABC exporters facilitate the extrusion of a wide spectrum of substrates including ions, lipids, peptides, toxic xenobiotics, etc [10]. Numerous studies have suggested the role of ABC transporters in various human diseases and chemoresistance [1]. Besides their role as transporters, some ABC proteins harbor just the NBDs, for instance, the ABCE and ABCF family representatives, and have implications in ribosome biogenesis, translation control, etc. [11,12], further adding to the immense biological relevance of this superfamily. The primary requirement for in-depth investigations pertaining to ABC proteins is their accurate identification from the genome/proteome and analysis of its topology. Even though programs performing certain steps in isolation exist, there is a need for a unified package to do the job with lesser hassles and more emphasis on reproducibility, keeping in view that reproducibility of research is the key element in modern science [13]. In an effort to make a simpler program available to the biology researcher community, we herein present “ABC-finder” as a simple and fast tool for identification and topology prediction of ABC proteins based on our previously established pipeline which led to the inventorization of ABC proteins in a number of yeast species [14–16]. ABC-finder combines stand-out methodologies, namely profile-HMM and TOPCONS for homology detection, and prediction of the topology of membrane proteins, respectively in a

seamless manner whereby the users need to provide only the organism's name or the proteome file as an input. Our analysis with reference organisms shows a strong correlation with the ABC protein inventories available in the literature.

2 METHODOLOGIES AND WORKFLOW

2.1 Data submission

Protein sequence data can be submitted in the following manner to the ABC-finder platform. The user can either specify the organism name (*Homo sapiens*, *Arabidopsis thaliana*, or *Candida glabrata*) or can upload a raw FASTA file (uncompressed) of the proteome. The file size is, however, limited to 95 MB. In the last step, the user can provide an email address to get the result files directly delivered.

2.2 Method

2.2.1 Prediction of potential ABC proteins based on NBD HMM

To extract the putative ABC proteins from the sequence input, we utilized a protocol that was developed previously by us for inventorization of ABC proteins in yeasts [14–16]. Briefly, the HMM profile of the ABC-tran model (accession PF00005) obtained from the Pfam database [17] is used as a query to match against the proteome of an organism with the help of the "hmmsearch" function within the HMMER package [18] using the default settings. Positive hits above the default threshold are then further filtered based on a cutoff defined from the plot of domain-score and E-value. The cutoff is interpreted from the point of a sudden drop in the domain score in the plot.

2.2.2 Clustering of sequences using Cd-hit

Next, ABC-finder uses the Cluster Database at High Identity with Tolerance (CD-HIT) [19]. This clustering algorithm allows the grouping of all the sequences based on their sequence similarities. Given that Cd-hit clusters the highly similar (redundant) sequences, it allows the grouping of different ABC protein sequence variants that usually populate the database in case of higher eukaryotes. The user has the option to optimize the input parameters for the

particular program from the “advanced options” menu.

2.2.3 Prediction of the TMD and protein topology

Since most of the families under the ABC superfamily include membrane transporters, it is relevant to detect their presence among putative proteins. Following the CD-hit run, ABC-finder utilizes the stand-out program TOPCONS [20], which is a widely used program for consensus prediction of membrane protein topology. Once the TM helices are defined using TOPCONS, ABC-finder uses a set of in-house written Python scripts to demarcate the NBDs and TMDs in each of the putative ABC proteins.

2.2.4 Visualization

E-value and domain score values obtained from 2.2.1 and the TMD and NBD arrangement obtained from 2.2.2 are used to generate the E-value v/s domain score plot and the final topology diagram of the proteins, respectively. Dynamic and Static plots are generated with Plotly (https://plot.ly/) and Orca (https://plot.ly/python/orca-management/). All the results files and several other log files are generated and zipped as the output data for downloading.

The overall workflow implemented in the ABC-finder server is shown in **Figure 1**. The help page provides a step-by-step guide to ABC-finder. All results are kept on the server for ten days.

3 IMPLEMENTATION

ABC-finder server is implemented in Python 3.7.6 (http://www.python.org/) using the Web framework Django 3.0.3 (http://www.djangoproject.com/). The system is containerized using Docker 19.03.3 [21] and docker-compose 1.25.0. Job queuing is carried out by the asynchronous task queue Celery 3.0.19 that uses the distributed message passing system Redis 3.3.11.

Users can also install the ABC-finder web server directly from the source code or build and run it from within Docker container using the instructions provided in the **Supplementary information**. The **Supplementary information** also contains FAQs and some useful links.

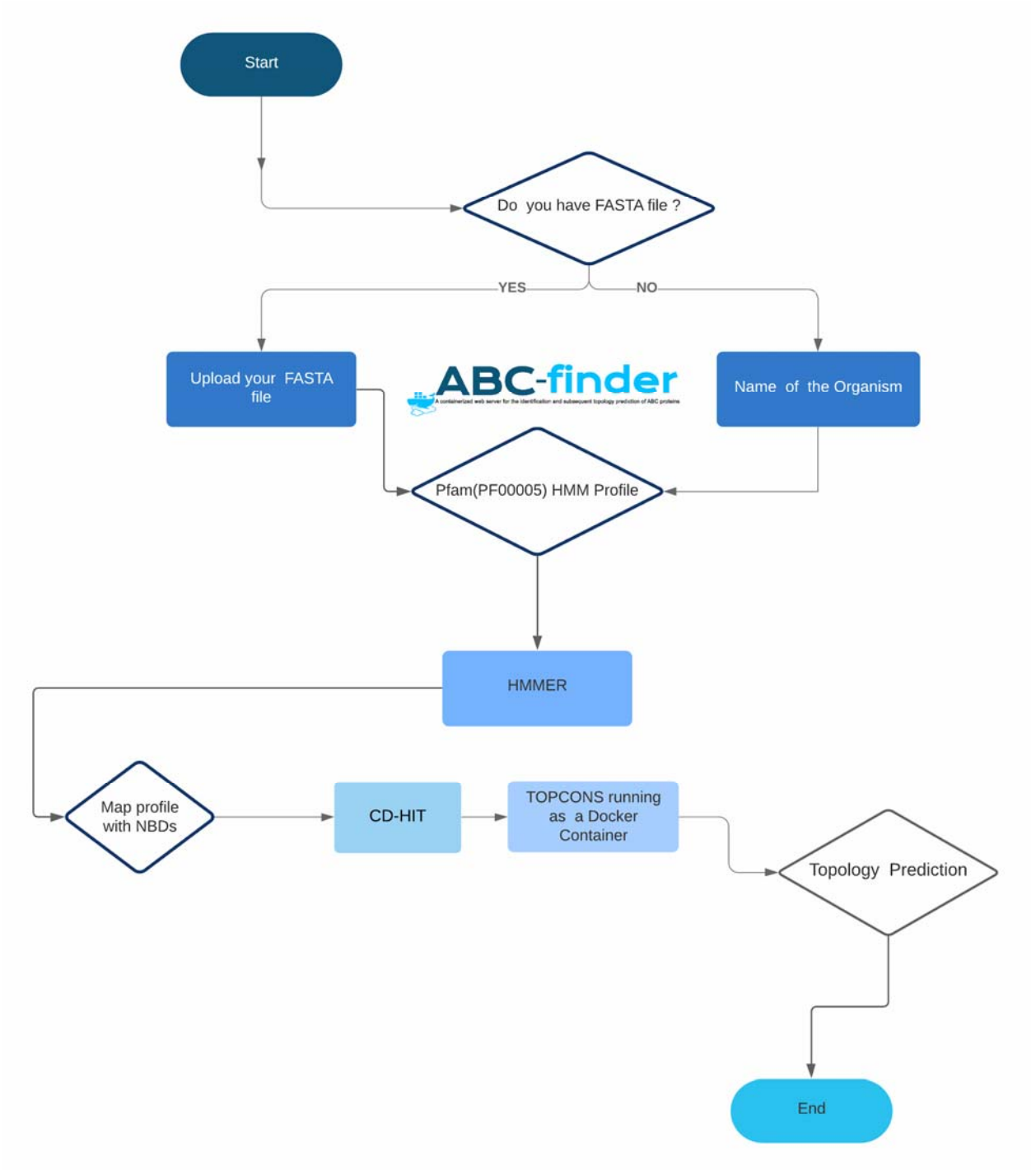

Figure 1: Generalized workflow of ABC-finder

4 RESULTS

Once the input proteome file is uploaded or the organism is specified to ABC-finder, the server provides a zipped result folder which includes the following files in different formats (indicated within the parenthesis):

1. **search(.faa)** → it is the proteome file that is submitted to ABC-finder for analysis. It is either uploaded by the user or retrieved by the web server depending upon the organism specified by the user.
2. **search_faa0(.clstr)** → it is the output file obtained after running Cd-hit. It contains the different protein clusters based on sequence similarity.
3. **search_faa1(.faa)** → FASTA file utilized as the input for TOPCONS.
4. **hmm_output** → it is the output obtained after running the "hmmsearch" function with the user-specified proteome file using profile.hmm as the query.
5. **threshold(.txt)** → this file lists the IDs of the putative ABC protein candidates along with their E-values. These positive hits are obtained based on a cutoff defined from the plot of domain-score and E-value.
6. **profile(.hmm)** → it is the profile HMM derived from the PF00005 seed alignment.
7. **query.result(.txt)** → it is the final output file obtained from TOPCONS.
8. **E-value_Domain_Score(.html/.pdf/.PNG/.SVG/.JPEG)** → it is the plot of E-value v/s domain score for the sequences above the inclusion threshold as obtained from the hmm_output file. A representative plot for the analysis of *S. cerevisiae* proteome is provided in **Figure 2**.

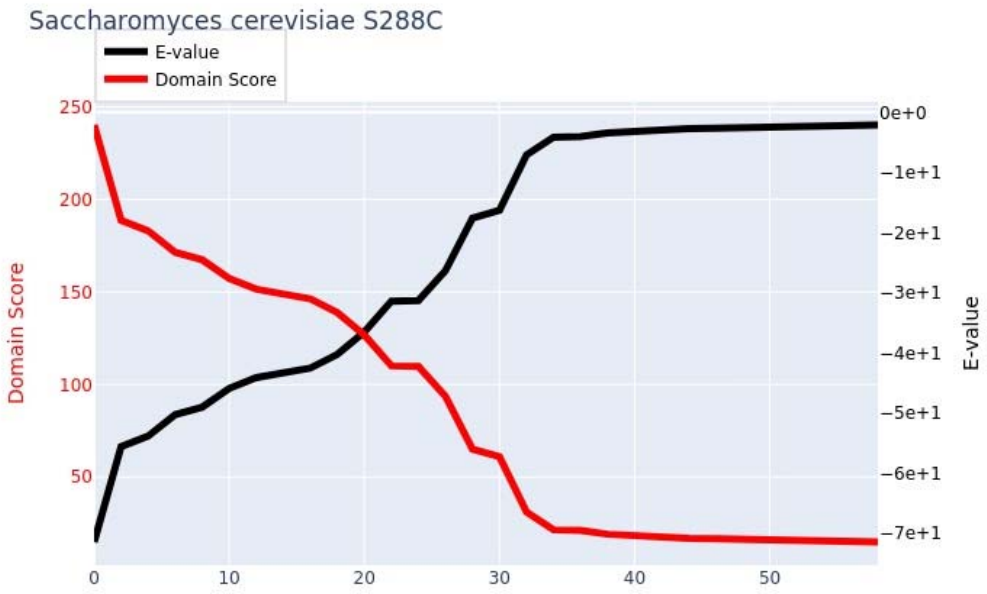

Figure 2: Plot of E-value v/s domain score of the sequences above the inclusion threshold value obtained after running the "hmmsearch" function with the *S. cerevisiae* S288C proteome.

9. **Topology Plot(.html/.pdf/.PNG/.SVG/.JPEG)→** this plot represents the topology of the identified ABC proteins. TOPCONS is utilized for the prediction of the transmembrane domains. A representative topology plot for the predicted ABC proteins in the *S. cerevisiae* S288C strain is provided in **Figure 3**. The topology plot should be referred alongside the Cd-hit output file "search_faa0(.clstr)" to detect the highly similar sequence clusters, especially in the case of higher eukaryotes, wherein multiple variants are present for each protein in the database.

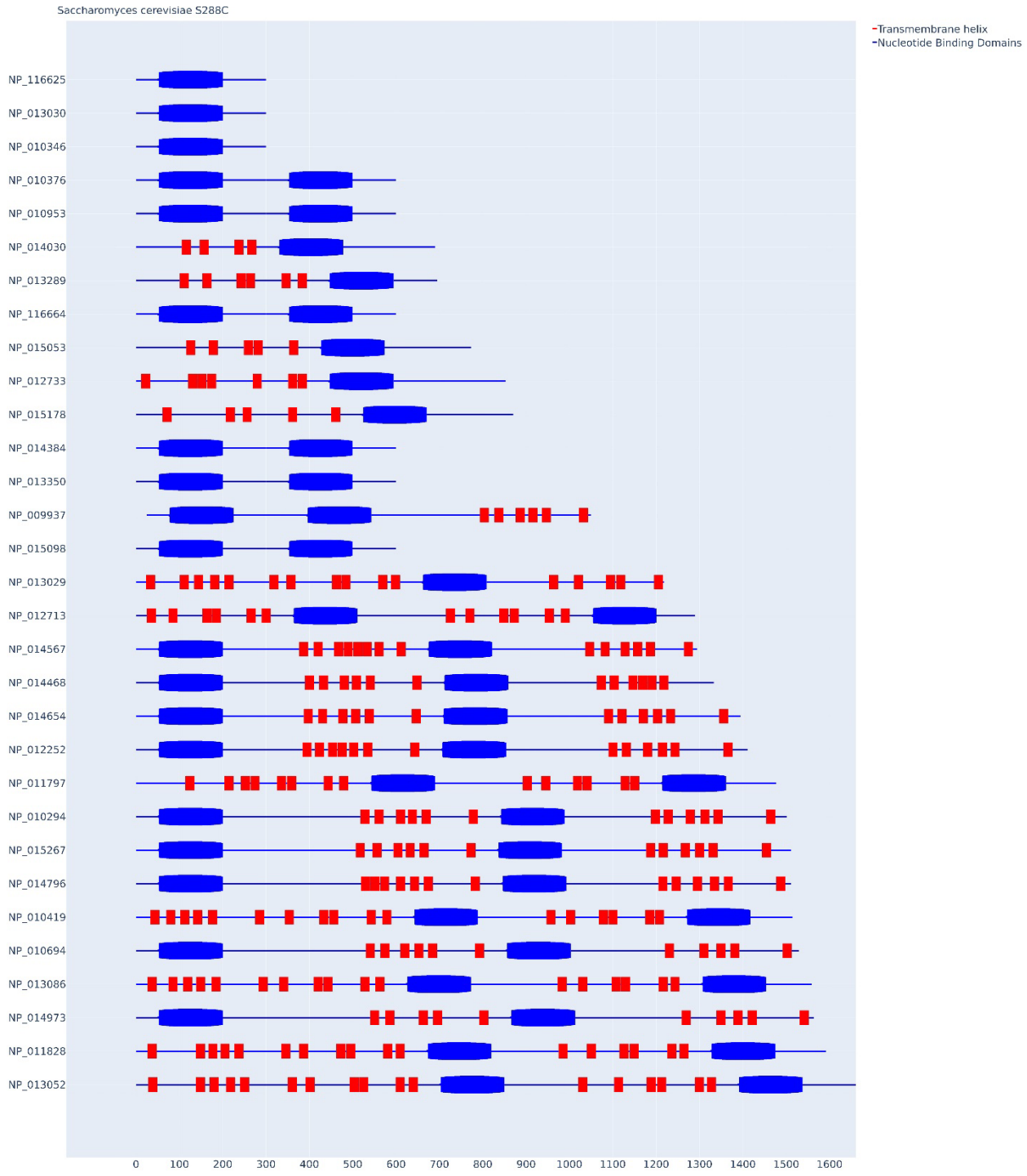

Figure 3: Plot representing the topology of the predicted ABC proteins in the *S. cerevisiae* S288C strain. TOPCONS is utilized for the prediction of the transmembrane domains. Red columns represent transmembrane domains and blue spheres represent the nucleotide binding domains.

To test ABC-finder, we utilized already reported inventories of ABC proteins from prokaryotic as well as eukaryotic systems. The results are summarized in **Table 1**.

Table 1: A comparison between the number of ABC proteins reported in the literature v/s predicted by ABC-finder. cd-hit parameter used: **-c =** sequence identity threshold; **-n =** word length; **Default =** (-c = 0.9; -n = 5). Result files for all the below mentioned organisms are provided in a zipped folder as supplementary information.

S. no.	**Organism name**	**ABC proteins reported [literature reference]**	**ABC proteins predicted by ABC-finder**	**Cd-hit parameter used**	**Match percent (between the literature reported and ABC-finder reported numbers)**
1	*Homo sapiens*	48-49 [12,22]	48	-c 0.6 -n 4	100
2	*Drosophila melanogaster*	56 [23]	55	Default	98
3	*Escherichia coli*	79 [24]	77	Default	97
4	*Saccharomyces cerevisiae*	30 [25]	31	Default	97
5	*Candida glabrata*	25 [14]	25	Default	100
6	*Candida auris*	28 [16]	28	Default	100

7	*Debaryomyces hansenii*	30 [15]	30	Default	100
8	*Magnaporthe oryzae*	50 [26]	50	-c 1 -n 5	100
9	*Caenorhabditis elegans*	60 [27]	56	Default	93
10	*Leishmania major*	42 [28]	41	-c 1 -n 5	98
11	*Bombyx mori*	53-55 [23]	53	Default	100
12	*Oryza sativa*	121 [29]	111	-c 1 -n 5	92
13	*Arabidopsis thaliana*	120 [29]	121	Default	99

5 SOFTWARE AVAILABILITY

The scripts we used to deploy the containers in our study are publicly available on GitHub and the Docker images are available on DockerHub.

GitHub URL (project home page for source code): https://github.com/lynngroup/abcfinder

DockerHub URL: https://hub.docker.com/r/lynngroup/abcfinder

6 ACKNOWLEDGEMENTS

We are grateful to all our colleagues involved in testing the ABC-finder web server. AB acknowledges funding support from SERB (SRG/2019/000514). PV and NKM are thankful to funding agencies ICMR and Department of Biotechnology (DBT/JRF/15/AL, DBT/2016/JNU/695), Govt. of India and JNU for financial support and fellowships. The authors acknowledge the High-Performance Computational Facility (HPCF) at SCIS, JNU.

7 CONFLICT OF INTEREST

None

8 REFERENCES

1. Higgins CF. ABC transporters: physiology, structure and mechanism--an overview. Res. Microbiol. 2001; 152:205–210

2. Locher KP. Mechanistic diversity in ATP-binding cassette (ABC) transporters. Nat. Struct. Mol. Biol. 2016; 23:487–493

3. Prasad R, Banerjee A, Khandelwal NK, et al. The ABCs of Candida albicans Multidrug Transporter Cdr1. Eukaryot. Cell 2015; 14:1154–1164

4. Moreno A, Banerjee A, Prasad R, et al. PDR-like ABC systems in pathogenic fungi. Res. Microbiol. 2019; 170:417–425

5. Dean M, Annilo T. Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. Annu. Rev. Genomics Hum. Genet. 2005; 6:123–142

6. Luckenbach T, Fischer S, Sturm A. Current advances on ABC drug transporters in fish. Comp. Biochem. Physiol. C. Toxicol. Pharmacol. 2014; 165:28–52

7. Popovic M, Zaja R, Loncar J, et al. A novel ABC transporter: the first insight into zebrafish (Danio rerio) ABCH1. Mar. Environ. Res. 2010; 69 Suppl:S11–3

8. Verrier PJ, Bird D, Burla B, et al. Plant ABC proteins--a unified nomenclature and updated inventory. Trends Plant Sci. 2008; 13:151–159

9. Lewinson O, Livnat-Levanon N. Mechanism of Action of ABC Importers: Conservation, Divergence, and Physiological Adaptations. J. Mol. Biol. 2017; 429:606–619

10. Beis K. Structural basis for the mechanism of ABC transporters. Biochem. Soc. Trans. 2015;

43:889–893
11. Murina V, Kasari M, Takada H, et al. ABCF ATPases Involved in Protein Synthesis, Ribosome Assembly and Antibiotic Resistance: Structural and Functional Diversification across the Tree of Life. J. Mol. Biol. 2019; 431:3568–3590
12. Vasiliou V, Vasiliou K, Nebert DW. Human ATP-binding cassette (ABC) transporter family. Hum. Genomics 2009; 3:281–290
13. Baker M. 1,500 scientists lift the lid on reproducibility. Nature 2016; 533:452–454
14. Kumari S, Kumar M, Khandelwal NK, et al. ABC transportome inventory of human pathogenic yeast Candida glabrata: Phylogenetic and expression analysis. PLoS One 2018; 13:e0202993
15. Wasi M, Khandelwal NK, Vishwakarma P, et al. Inventory of ABC proteins and their putative role in salt and drug tolerance in Debaryomyces hansenii. Gene 2018; 676:227–242
16. Wasi M, Khandelwal NK, Moorhouse AJ, et al. ABC Transporter Genes Show Upregulated Expression in Drug-Resistant Clinical Isolates of Candida auris: A Genome-Wide Characterization of ATP-Binding Cassette (ABC) Transporter Genes. Front. Microbiol. 2019; 10:1445
17. Finn RD, Coggill P, Eberhardt RY, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44:D279–85
18. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009; 23:205–211
19. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006; 22:1658–1659
20. Tsirigos KD, Peters C, Shu N, et al. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 2015; 43:W401–7
21. Boettiger C. An introduction to Docker for reproducible research. Oper. Syst. Rev. 2015; 49:71–79
22. Dean M, Rzhetsky A, Allikmets R. The human ATP-binding cassette (ABC) transporter superfamily. Genome Res. 2001; 11:1156–1166
23. Dermauw W, Van Leeuwen T. The ABC gene family in arthropods: comparative genomics and role in insecticide transport and resistance. Insect Biochem. Mol. Biol. 2014; 45:89–110
24. Feng Z, Liu D, Wang L, et al. A Putative Efflux Transporter of the ABC Family, YbhFSR, in

Escherichia coli Functions in Tetracycline Efflux and Na+(Li+)/H+ Transport. Front. Microbiol. 2020; 11:556
25. Paumi CM, Chuk M, Snider J, et al. ABC transporters in Saccharomyces cerevisiae and their interactors: new technology advances the biology of the ABCC (MRP) subfamily. Microbiol. Mol. Biol. Rev. 2009; 73:577–593
26. Kim Y, Park S-Y, Kim D, et al. Genome-scale analysis of ABC transporter genes and characterization of the ABCC type transporter genes in Magnaporthe oryzae. Genomics 2013; 101:354–361
27. Sheps JA, Ralph S, Zhao Z, et al. The ABC transporter gene family of Caenorhabditis elegans has implications for the evolutionary dynamics of multidrug resistance in eukaryotes. Genome Biol. 2004; 5:R15
28. Leprohon P, Légaré D, Girard I, et al. Modulation of Leishmania ABC protein gene expression through life stages and among drug-resistant parasites. Eukaryot. Cell 2006; 5:1713–1725
29. Garcia O, Bouige P, Forestier C, et al. Inventory and comparative analysis of rice and Arabidopsis ATP-binding cassette (ABC) systems. J. Mol. Biol. 2004; 343:249–265