DATA NOTE

Open Access

Phylogenetic trees, conserved motifs and predicted subcellular localization for transcription factor families in pearl millet

Yingwei Qu[1], Ambika Dudhate[2], Harshraj Subhash Shinde[3], Tetsuo Takano[1] and Daisuke Tsugama[1*]

Abstract

Objectives Pearl millet (*Pennisetum glaucum*) is a cereal crop that is tolerant to a high temperature, a drought and a nutrient-poor condition. Characterizing pearl millet proteins can help to improve productivity of pearl millet and other crops. Transcription factors in general are proteins that regulate transcription of their target genes and thereby regulate diverse processes. Some transcription factor families in pearl millet were characterized in previous studies, but most of them are not. The objective of the data presented was to characterize amino acid sequences for most transcription factors in pearl millet.

Data description Sequences of 2395 pearl millet proteins that have transcription factor-associated domains were extracted. Subcellular and suborganellar localization of these proteins was predicted by MULocDeep. Conserved domains in these sequences were confirmed by CD-Search. These proteins were classified into 85 families on the basis of those conserved domains. A phylogenetic tree including pearl millet proteins and their counterparts in *Arabidopsis thaliana* and rice was constructed for each of these families. Sequence motifs were identified by MEME for each of these families.

Keywords Pearl millet, Transcription factor, Phylogenetic analysis, Protein family, Subcellular localization, Protein domain, Motif

Objective

Pearl millet (*Pennisetum glaucum*) is a staple cereal crop that is tolerant to a high temperature, a drought and a poor-nutrient condition and that is produced in semi-arid regions [1]. Characterization of pearl millet genes can help to better understand pearl millet stress tolerance and to improve productivity of pearl millet and other crops. The whole genome sequence of pearl millet was released previously [2]. On the basis of this sequence, pearl millet gene or protein families such as a WRKY transcription factor (TF) family, an NAC (NAM, ATAF and CUC) TF family, a GRAS TF family and a MYB TF family have been identified and characterized [3–6]. However, most pearl millet protein families are

*Correspondence:
Daisuke Tsugama
tsugama@g.ecc.u-tokyo.ac.jp
[1] Asian Research Center for Bioresource and Environmental Sciences (ARC-BRES), Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Midori-cho, Nishi-tokyo-shi, 188-0002 Tokyo, Japan
[2] Stowers Institute for Medical Research, 1000 East 50th Street, 64110 Kansas City, issouri, USA
[3] University of Kentucky, 40506 Lexington, Kentucky, USA

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	seqID_family.txt	Tab-delimited text file (.txt)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]
Data file 2	family_seqID_w_ol.txt	Tab-delimited text file (.txt)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]
Data file 3	CD-Search_full_all.txt	Tab-delimited text file (.txt)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]
Data file 4	CD-Search_wo_domain_define_all.txt	Tab-delimited text file (.txt)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]
Data file 5	MULocDeep_subcellular_localization_prediction_all.txt	Tab-delimited text file (.txt)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]
Data file 6	MULocDeep_suborganellar_localization_prediction_all.txt	Tab-delimited text file (.txt)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]
Data file 7	methods_notes.txt	Text file (.txt)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]
Data set 1	phylogenetic_trees.zip	Zip archive file (.zip)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]
Data set 2	MEME_results.zip	Zip archive file (.zip)	figshare (https://doi.org/10.6084/m9.figshare.21623829) [26]

uncharacterized. TFs in general regulate transcription of multiple genes and thus can act as hubs for diverse processes. TFs can therefore be useful as either a transgene in genetic modification or a target of genome editing for improving plant performance. The objective of the data presented was to characterize amino acid sequences of most pearl millet TFs.

Data description

Amino acid sequences for all pearl millet proteins deduced from its whole genome sequence [2] were downloaded from the International Pearl Millet Genome Sequencing Consortium website [7]. Hidden Markov models (HMMs) for protein families in the Pfam database [8] were downloaded from an InterPro website [9]. HMMs in those amino acid sequences were detected by the hmmscan program in HMMER (version 3.3) [10, 11]. On the basis of the detected HMMs, 2395 sequences were regarded as the sequences for putative pearl millet TFs and these were classified into 85 families. Conserved domains in these TFs were confirmed by Batch CD-Search [12, 13]. Subcellular and suborganellar localization of these TFs was predicted by MULocDeep [14, 15]. Amino acid sequences of rice (*Oryza sativa* ssp. *japonica*) and *Arabidopsis thaliana* TFs were downloaded from a PlantTFDB website [16–18]. For the families that were not available in PlantTFDB, amino acid sequences of all rice (*O. sativa* ssp. *indica*) and Arabidopsis proteins were downloaded from an Ensembl Plants website [19, 20] and used for hmmscan as described above to identify proteins in those families. For each of these families except the 13 families which contain less than five members, the sequences from pearl millet, rice and Arabidopsis were aligned by ClustalW [21] and a phylogenetic tree file was obtained with the neighbor-joining method on the MEGA X software [22]. The phylogenetic tree was visualized on the Interactive Tree of Life (iTOL) online tool (version 6) [23, 24]. For each of the 84 families identified, motifs in the pearl millet amino acid sequences were identified *de novo* by the MEME program (version 5.5.0) [25]. Data obtained by these analyses were deposited in the figshare repository (Table 1) [26].

Limitations

- Previous studies on protein family characterization [e.g., 3, 4, 5, 6] were not integrated in the data presented.
- Most protein families other than the TF families in pearl millet are still uncharacterized.

Abbreviations

TF	Transcription factor
HMM	Hidden Markov model

Acknowledgements

The authors appreciate data and advice from Dr. Shashi Kumar Gupta and his colleagues in International Crops Research Institute for the Semi-Arid Tropics (ICRISAT).

Authors' contributions

All authors collected data and wrote the manuscript.

Funding

This work was supported by JSPS (Japan Society for the Promotion of Science) Kakenhi Grant (Grant Number: 19KK0155 and 19H02928).

Data availability

The data described in this Data note can be freely and openly accessed on figshare under https://doi.org/10.6084/m9.figshare.21623829. Please see Table 1 and references [26] for details and links to the data.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 27 November 2022 / Accepted: 6 March 2023
Published online: 20 March 2023

References

1. Basavaraj G, Rao PP, Bhagavatula S, Ahmed W. Availability and utilization of pearl millet in India. J SAT Agrirc Res. 2010;8:1–6.
2. Varshney RK, Shi C, Thudi M, Mariac C, Wallace J, Qi P, et al. Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. Nat Biotechnol. 2017;35:969–76.
3. Chanwala J, Satpati S, Dixit A, Parida A, Giri MK, Dey N. Genome-wide identification and expression analysis of WRKY transcription factors in pearl millet (*Pennisetum glaucum*) under dehydration and salinity stress. BMC Genomics. 2020;21:231.
4. Dudhate A, Shinde H, Yu P, Tsugama D, Gupta SK, Liu S, Takano T. Comprehensive analysis of NAC transcription factor family uncovers drought and salinity stress response in pearl millet (*Pennisetum glaucum*). BMC Genomics. 2021;22:70.
5. Jha DK, Chanwala J, Sandeep IS, Dey N. Comprehensive identification and expression analysis of GRAS gene family under abiotic stress and phytohormone treatments in pearl millet. Funct Plant Biol. 2021;48:1039–52.
6. Chanwala J, Khadanga B, Jha DK, Sandeep IS, Dey N. MYB transcription factor family in pearl millet: genome-wide identification, evolutionary progression and expression analysis under abiotic stress and phytohormone treatments. Plants (Basel). 2023;12:355.
7. Varshney RK, Shi C, Thudi M, Mariac C, Wallace J, Qi P et al. International Pearl Millet Genome Sequencing Consortium (IPMGSC). https://cegresources.icrisat.org/data_public/PearlMillet_Genome/v1.1/. Accessed 27 Nov 2022.
8. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49:D412–9.
9. InterPro. https://www.ebi.ac.uk/interpro/download/Pfam/. Accessed 27 Nov 2022.
10. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41:e121.
11. HMMER. http://hmmer.org/. Accessed 27 Nov 2022.
12. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48:D265–8.
13. Batch CD-Search. https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi. Accessed 27 Nov 2022.
14. Jiang Y, Wang D, Yao Y, Eubel H, Künzler P, Møller IM, Xu D, MULocDeep. A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. Comput Struct Biotechnol J. 2021;19:4825–39.
15. MULocDeep. https://mu-loc.org/. Accessed 27 Nov 2022.
16. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45(D1):D1040–5.
17. PlantTFDB. http://planttfdb.gao-lab.org/index.php?sp=Osj. Accessed 27 Nov 2022
18. PlantTFDB. http://planttfdb.gao-lab.org/index.php?sp=Ath. Accessed 27 Nov 2022
19. Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, et al. Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res. 2022;50:D996–D1003.
20. EnsemblPlants. https://plants.ensembl.org/info/data/ftp/index.html. Accessed 27 Nov 2022.
21. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
22. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol. 2018;35:1547–9.
23. Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49:W293–6.
24. Interactive Tree of Life. https://itol.embl.de/. Accessed 27 Nov 2022.
25. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39–49.
26. Tsugama D, Qu Y, Dudhate A, Shinde HS, Takano T. Pearl millet transcription factor family characterization data. figshare. 2022. https://doi.org/10.6084/m9.figshare.21623829

Publisher's note