microPublication BIOLOGY

8/6/2022 - *Open Access*

Gossypium hirsutum gene of unknown function, Gohir.A02G044702.1, encodes a potential B3 Transcription Factor of the REM subfamily

Michael Allen[1], Amanda M. Hulse-Kemp[2,3§], Amanda R. Storm[1§]

[1]Department of Biology, Western Carolina University, Cullowhee, NC

[2]Genomics and Bioinformatics Research Unit, The Agricultural Research Service of U.S. Department of Agriculture, Raleigh, NC

[3]Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC

[§]To whom correspondence should be addressed: amanda.hulse-kemp@usda.gov; arstorm@wcu.edu

Abstract

A gene of unknown function, Gohir.A02G044702.1, identified in *Gossypium hirsutum* was studied using sequence and structure bioinformatic tools. The encoded protein (UniProt A0A1U8MGX4) was predicted to localize to the nucleus, was found to retain the B3 transcription factor domain with conserved DNA-binding residues and to most closely cluster with REM subfamily members of B3-domain containing proteins.

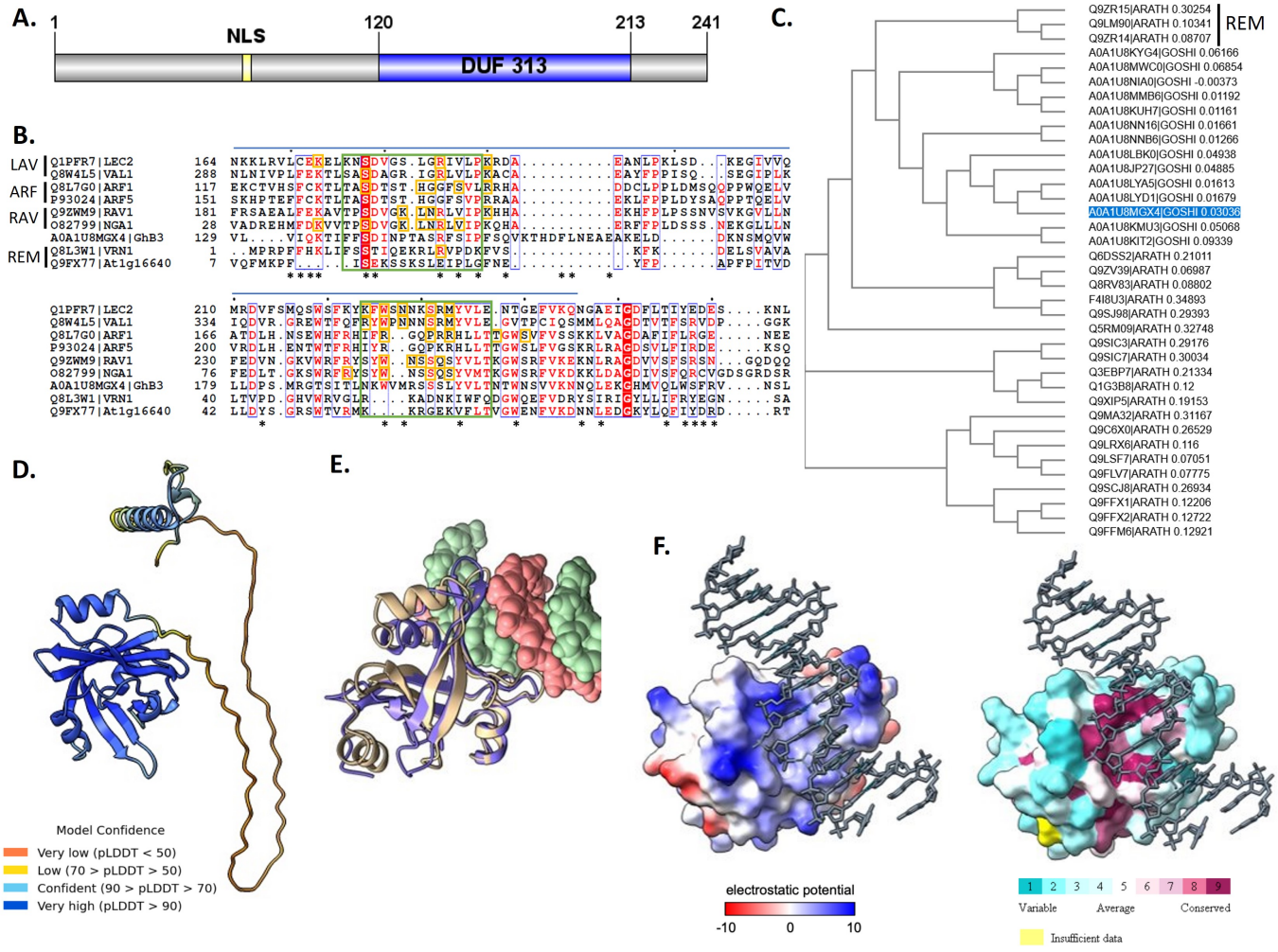

Figure 1. Sequence and Structure Characterization of GhB3-A0A1U8MGX4

(A) Domain architecture of GhB3-A0A1U8MGX4 indicating the location of predicted sequence features: Created using IBS (Wenzhong et al., 2015) based on predictions from InterPro (Blum et al., 2020), BUCSA (Savojardo et al., 2018), and LOCALIZER (Sperschneider et al., 2017), NLS - nuclear localization sequence, DUF 313 domain (PF03754). (B) Multi-

Sequence Alignment of DUF 313 domain region of GhB3-A0A1U8MGX4 and B3 domain-containing proteins with known structures from each subfamily : LAV (LEC2 - PDB 6J9C, VAL1 - PDB 6J9A, Tao et al., 2019), ARF (ARF1 - PDB 4LDX, ARF5 - PDB 4LDU, Boer et al., 2014), RAV (RAV1 - PDB 1WID, Yamasaki et al., 2004; NGA1 - PDB 5OS9, Sasnauskas et al., 2018), REM (VRN1 - PDB 4I1K, Gordon et al., 2013; At1g16640 - PDB 1YEL, Waltner et al., 2005), Created with ClustalOmega (Madeira et al., 2019) and ESPript3 (Robert and Gouet, 2014) and annotated to show location of the DUF 313 domain (blue bar), ConSurf-identified highly conserved sequences in GhB3-A0A1U8MGX4 (*), DNA-interacting N- and C-arm loops (green boxes) and residues implicated in DNA binding either by crystal structures or mutagenesis (yellow boxes). (C) Phylogenetic tree of cotton (*Gossypium hirsutum*) and *Arabidopsis thaliana* GhB3-A0A1U8MGX4 homologs: homologs identified using PhyloGenes (Zhang et al., 2020) and tree created using ClustalOmega, the GhB3-A0A1U8MGX4 sequence is highlighted (blue) and the 3 Arabidopsis homologs identified as belonging to the REM subfamily are indicated. (D) AlphaFold model structure for GhB3-A0A1U8MGX4 with coloration based on model confidence. (E) Structure overlay of GhB3-A0A1U8MGX4 model (beige ribbon) and LEC2-DNA complex (purple ribbon and space-filling DNA, PDB 6J9C). (F) GhB3-A0A1U8MGX4 model with DNA showing (left) electrostatic surface calculated by ChimeraX and (right) surface coloring based on ConSurf conservation scores.

Description

Introduction

In a recent sequencing of the genome of upland cotton, *Gossypium hirsutum* (L. accession Texas Marker-1 (TM-1) version 2.0 and annotation version 2.1) (Chen et al., 2020), the gene, LOC107936594 (CottonGen: Gohir.A02G044702_UTX-TM1_v2.1), was identified to encode a conserved protein of unknown function labeled 'B3 domain-containing protein At1g05920-like' (Gohir.A02G044702.1; NCBI: XP_016724843; UniProt: A0A1U8MGX4). Here we present evidence that supports this protein, referred to here as GhB3-A0A1U8MGX4, is part of the B3 domain-containing family, likely within the highly diverse REM subfamily, and contains the sequence and structure features required for nuclear-localization and DNA-binding to function as a transcription factor (TF). Members of the B3 superfamily all contain a B3 domain, a ~100 amino acid DNA-binding domain specific to plants and found in even unicellular green algae (Swaminathan, Peterson, Jack, 2008). It is prevalent in flowering plants with around 100 B3-containing proteins identified in Arabidopsis (Wang et al., 2012). This superfamily is involved in regulating many key processes, including stress and hormone responses (Yamasaki, 2016; Gordon et al., 2013; Boer et al., 2014), embryogenesis (Tao et al., 2019) and development (Waltner et al., 2005). The superfamily has been classified phylogenetically into four subfamilies: LAV (LEC/ABI/VAL), RAV (Related to ABI3/VP1), ARF (Auxin Response Factor) and REM (Reproductive Meristem), which differ in their DNA recognition sequence (Romanel et al., 2009).

Sequence Features

InterPro webserver identified the 241–amino acid GhB3-A0A1U8MGX4 protein as a member of the ‘DNA-binding pseudobarrel domain’ superfamily (IPR015300) and the ‘B3 domain-containing protein At2g31720-like’ family (IPR005508) with a domain of unknown function DUF 313 (PF03754). This plant-specific family of unknown function includes At2g31720, Auxin Response Factor 70 (ARF70, UniProt Q8RV83), a protein linked to stress response through yeast one-hybrid screening of TF-promoter interactions (Ikeuchi et al., 2018). Sequence analysis of GhB3-A0A1U8MGX4 by subcellular localization programs BUSCA and LOCALIZER predicted a nuclear localization sequence (NLS), KRKR, in the N-terminal region, in agreement with the function as a transcription factor. A domain architecture was created to visualize these sequence features (**Figure 1A**). ConSurf (Ashkenazy et al., 2016) was used to calculate the evolutionary conservation of each residue and the most highly conserved residues were found within the DUF 313 domain region (amino acids 120-213), these residues are indicated in the multi-sequence alignment by asterisks (**Figure 1B**). The full ConSurf results are available as Extended Data.

Homology

Of the 40 plant and 10 non-plant genomes used by the PhyloGenes webserver, homologs of GhB3-A0A1U8MGX4 were found in 25 species, confined to magnoliophyta (angiosperms), which agrees with the species listed by pfam for DUF 313 domain (PF03754) containing proteins. Arabidopsis ARF70 (Q8RV83) is a listed homolog but none of the other proteins have been functionally characterized or named. This indicates that these proteins belong to a set of B3 domain-containing proteins distinct from any currently characterized member of this family. PhyloGenes identified 14 cotton paralogs and 23 Arabidopsis orthologs of GhB3-A0A1U8MGX4. ClustalOmega was used to create a multi-sequence alignment (available as Extended Data) and a phylogenetic tree (Neighbor-joining) of these sequences (**Figure 1C**). The cotton paralogs separate into two distinct clusters with the most closely aligned Arabidopsis sequences to GhB3-A0A1U8MGX4 (Q9ZR15, Q9LM90, Q9ZR14) being proteins of unknown function but listed as belonging to the REM subfamily of B3 domain-containing proteins (Wang et al., 2012).

Structure Features

The AlphaFold tool in UCSF ChimeraX (Version 1.3, Pettersen et al., 2021; Jumper et al., 2021) was used to predict a structure model (available as Extended Data) for the GhB3-A0A1U8MGX4 protein (**Figure 1D**). The model showed a high level of confidence across amino acids 111-241, which includes the entirety of the DUF 313 domain region. This domain showed a seven-stranded, beta-sheet open barrel flanked by two short alpha helices characteristic of B3 DNA binding domains (Yamasaki et al., 2016). The lower confidence N-terminal 110 amino acids were deleted from the structure for remaining structure analyses.

A structure analog search using the DALI server (Holm, 2020) identified structures of B3 domain-containing proteins from each of the four subfamilies (LAV, RAV, ARF and REM) that closely matched the GhB3-A0A1U8MGX4 model (Z-scores between 9.7-11.1 and r.m.s.d values of 2.3-3.1 across 90-109 residues) even though sequence similarity was low (10-27%). Two structures were chosen from each subfamily and their sequences were aligned with GhB3-A0A1U8MGX4 using ClustalOmega and ESPript3 to create a multi-sequence alignment (**Figure 1B**). The multi-sequence alignment (MSA) shows the domain region where the greatest similarity was seen. The GhB3-A0A1U8MGX4 sequence retains many of the residues conserved across subfamilies (K134, S139, D140, D181, V212, G220) along with a number of residues identified as involved in DNA binding in other homologs (yellow boxes) which ConSurf also identified as evolutionarily conserved (asterisks).

The aligned sequences separated based on their known subfamilies with the GhB3-A0A1U8MGX4 sequence clustering most closely with the REM representative sequences (VRN1 and At1g16640). Subfamilies have been reported to have characteristic motifs within the DNA-interacting C-arm loop (RAV: 'WN/RSSQS'; ARF: 'RGQPK/RR'; LAV: 'WPNNKSR'; REM: deletion of 3-4 residues) (Swaminathan, Peterson, Jack, 2008). These motifs can be seen for each subfamily within the MSA but it is interesting that the GhB3-A0A1U8MGX4 C-arm loop sequence doesn't match any of these motifs. Differences in this region have been attributed to DNA-binding specificity so this could indicate that GhB3-A0A1U8MGX4 binds to a distinct DNA sequence. This is still in agreement with belonging to the REM subfamily as it is known to be the largest and most diverse subfamily with highly diverged DNA binding (Romanel et al., 2009).

DNA was modeled into the predicted binding site using a structure overlay between the GhB3-A0A1U8MGX4 model and a LEC2-DNA structure (PDB 6J9C) using UCSF ChimeraX MatchMaker (**Figure 1E**). There is good alignment (r.m.s.d. of 1.014 angstroms over 53 of 96 pruned atom pairs) with DNA-interacting N- and C-arm loops associating with the major groove. A characteristic feature of B3 DNA-binding domains is a large positively charged area around the DNA binding site (Yamasaki et al., 2004). The GhB3-A0A1U8MGX4 model contains a similar distinct and large area of positively charged residues in the modeled binding site. Mapping of ConSurf conservation scores onto the structure indicates that these residues are also highly conserved (**Figure 1F**).

Conclusion

Evidence from the sequence, homology and structure analyses support GhB3-A0A1U8MGX4 being a B3 domain-containing protein within the REM subfamily, with a retained DNA-binding site. The nuclear subcellular localization and conserved and charged binding site features indicate a DNA binding function such as a transcription factor, similar to other members of this family, although potentially with a unique DNA recognition sequence. With a distinct binding motif and largely uncharacterized homologs, this protein could belong to a new subset of REM proteins involved in regulating a distinct plant process, likely specific to angiosperm species.

Acknowledgements: This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D. The authors thank the WCU Biology Department for supporting the Course-based Undergraduate Research Experience that was the basis for data collection.

Extended Data

Description: ConSurf sequence conservation results for A0A1U8MGX4 . Resource Type: Dataset. File: A0A1U8MGX4_ConSurf_seq_results.pdf. DOI: 10.22002/D1.20239

Description: AlphaFold modeled structure of A0A1U8MGX4. Resource Type: Model. File: A0A1U8MGX4_AlphaFold_model.pdb. DOI: 10.22002/D1.20240

Description: ClustalOmega MSA of PhyloGenes phylogenetic tree. Resource Type: Dataset. File: A0A1U8MGX4_ClustalO_MSA.aln. DOI: 10.22002/D1.20241

References

Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N. 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 44: W344-50. PubMed ID: 27166375

Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD. 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49: D344-D354. PubMed ID: 33156333

Boer DR, Freire-Rios A, van den Berg WA, Saaki T, Manfield IW, Kepinski S, López-Vidrieo I, Franco-Zorrilla JM, de Vries SC, Solano R, Weijers D, Coll M. 2014. Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. Cell 156: 577-89. PubMed ID: 24485461

Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, Ding M, Ye W, Kirkbride RC, Jenkins J, Plott C, Lovell J, Lin YM, Vaughn R, Liu B, Simpson S, Scheffler BE, Wen L, Saski CA, Grover CE, Hu G, Conover JL, Carlson JW, Shu S, Boston LB, Williams M, Peterson DG, McGee K, Jones DC, Wendel JF, Stelly DM, Grimwood J, Schmutz J. 2020. Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. Nat Genet 52: 525-533. PubMed ID: 32313247

King GJ, Chanson AH, McCallum EJ, Ohme-Takagi M, Byriel K, Hill JM, Martin JL, Mylne JS. 2013. The Arabidopsis B3 domain protein VERNALIZATION1 (VRN1) is involved in processes essential for development, with structural and mutational studies revealing its DNA-binding surface. J Biol Chem 288: 3198-207. PubMed ID: 23255593

Holm L. 2020. Using Dali for Protein Structure Comparison. Methods Mol Biol 2112: 29-42. PubMed ID: 32006276

Ikeuchi M, Shibata M, Rymen B, Iwase A, Bågman AM, Watt L, Coleman D, Favero DS, Takahashi T, Ahnert SE, Brady SM, Sugimoto K. 2018. A Gene Regulatory Network for Cellular Reprogramming in Plant Regeneration. Plant Cell Physiol 59: 765-777. PubMed ID: 29462363

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596: 583-589. PubMed ID: 34265844

Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, Lahrmann U, Zhao Q, Zheng Y, Zhao Y, Xue Y, Ren J. 2015. IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics 31: 3359-61. PubMed ID: 26069263

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47: W636-W641. PubMed ID: 30976793

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci 30: 70-82. PubMed ID: 32881101

Robert X, Gouet P. 2014. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res 42: W320-4. PubMed ID: 24753421

Romanel EA, Schrago CG, Couñago RM, Russo CA, Alves-Ferreira M. 2009. Evolution of the B3 DNA binding superfamily: new insights into REM family gene diversification. PLoS One 4: e5791. PubMed ID: 19503786

Tao Z, Hu H, Luo X, Jia B, Du J, He Y. 2019. Embryonic resetting of the parental vernalized state by two B3 domain transcription factors in Arabidopsis. Nat Plants 5: 424-435. PubMed ID: 30962525

Sasnauskas G, Manakova E, Lapėnas K, Kauneckaitė K, Siksnys V. 2018. DNA recognition by Arabidopsis transcription factors ABI3 and NGA1. FEBS J 285: 4041-4059. PubMed ID: 30183137

Savojardo C, Martelli PL, Fariselli P, Profiti G, Casadio R. 2018. BUSCA: an integrative web server to predict subcellular localization of proteins. Nucleic Acids Res 46: W459-W466. PubMed ID: 29718411

Sperschneider J, Catanzariti AM, DeBoer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM. 2017. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Sci Rep 7: 44598. PubMed ID: 28300209

Swaminathan K, Peterson K, Jack T. 2008. The plant B3 superfamily. Trends Plant Sci 13: 647-55. PubMed ID: 18986826

Waltner JK, Peterson FC, Lytle BL, Volkman BF. 2005. Structure of the B3 domain from Arabidopsis thaliana protein At1g16640. Protein Sci 14: 2478-83. PubMed ID: 16081658

Wang Y, Deng D, Zhang R, Wang S, Bian Y, Yin Z. 2012. Systematic analysis of plant-specific B3 domain-containing proteins based on the genome resources of 11 sequenced species. Mol Biol Rep 39: 6267-82. PubMed ID: 22302388

Yamasaki K, Kigawa T, Inoue M, Tateno M, Yamasaki T, Yabuki T, Aoki M, Seki E, Matsuda T, Tomo Y, Hayami N, Terada T, Shirouzu M, Osanai T, Tanaka A, Seki M, Shinozaki K, Yokoyama S. 2004. Solution structure of the B3 DNA binding domain of the Arabidopsis cold-responsive transcription factor RAV1. Plant Cell 16: 3448-59. PubMed ID: 15548737

Yamasaki K. 2016. Chapter 4 - Structures, Functions, and Evolutionary Histories of DNA-Binding Domains of Plant-Specific Transcription Factors. Plant Transcription Factors. Academic Press, 57-72 DOI: 10.1016/B978-0-12-800854-6.00004-X

Zhang P, Berardini TZ, Ebert D, Li Q, Mi H, Muruganujan A, Prithvi T, Reiser L, Sawant S, Thomas PD, Huala E. 2020. PhyloGenes: An online phylogenetics and functional genomics resource for plant gene function inference. Plant Direct 4: e00293. PubMed ID: 33392435

Funding: This research was funded in part by the U.S. Department of Agriculture Agricultural Research Service (USDA-ARS) project number 6066-21310-005-00D and Cotton Incorporated project 18-274 to AMH-K.

Author Contributions: Michael Allen: investigation, formal analysis, visualization, writing - original draft. Amanda M. Hulse-Kemp: conceptualization, data curation, writing - review editing, funding acquisition, project. Amanda R. Storm: conceptualization, data curation, formal analysis, methodology, supervision, validation, writing - original draft, writing - review editing, investigation, visualization, project.

Reviewed By: Anonymous

History: Received February 3, 2022 **Revision Received** July 28, 2022 **Accepted** July 26, 2022 **Published Online** August 6, 2022 **Indexed** August 20, 2022

Citation: Allen, M; Hulse-Kemp, AM; Storm, AR (2022). *Gossypium hirsutum* gene of unknown function, Gohir.A02G044702.1, encodes a potential B3 Transcription Factor of the REM subfamily. microPublication Biology. 10.17912/micropub.biology.000574