Pacific Northwest
NATIONAL LABORATORY

PNNL-33703

Validation and functional characterization of transcription factors in wheat using cell-free protein expression and high-throughput sequencing technologies

November 2022

Irina V Novikova
James E Evans
Meng Markillie
Hugh Mitchell

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-33703

Validation and functional characterization of transcription factors in wheat using cell-free protein expression and high-throughput sequencing technologies

November 2022

Irina V Novikova
James E Evans
Meng Markillie
Hugh Mitchell

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Transcription factors (TFs) are critical biomolecules that control and regulate gene expression in every organism. In wheat, there are 3,606 genes annotated as putative TFs which are primarily uncharacterized. In this work, using a DAP-seq approach, we have validated and characterized a select number of TFs (**5bl** and **5dl**) belonging to GeBP family of proteins, responsible for plant cellular growth, development, and differentiation. We found that top DNA binding motifs for **5bl** and **5dl** are (G/T)N(T/G)GTGGT and (C/G)AA(C/G)AA respectively. These motifs/peaks are enriched in the promoter regions of the genes, which are associated with rRNA biogenesis and protein maturation pathways. This study is also a first demonstration that DAP-seq can be applied for such a complex and large genome such as wheat.

Summary

This project was a pilot study with a 1-year timeframe to establish an in-house workflow that can help to validate and further characterize transcription factors. The science focus has been centered on a small family of transcription factors from wheat that has never been characterized previously. As DAP-seq protocols were never tested on a wheat genome and taking into account its high complexity and repetitiveness, we were successfully able to characterize a select number of transcription factors, providing a first demonstration that DAP-seq technology can be successfully used for such complex and large genomes such as wheat.

Acknowledgments

The experimental design, DNA library preparation, transcription factor synthesis, pulldown assays and data processing was supported by the **Earth and Biological Sciences Mission Seed Investment**, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). Access to the NextSeq550 instrument for DNA sequencing and to data archival was supported by the capacity proposal 60640 of Environmental Molecular Sciences Laboratory. PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Contents

Figures

Tables

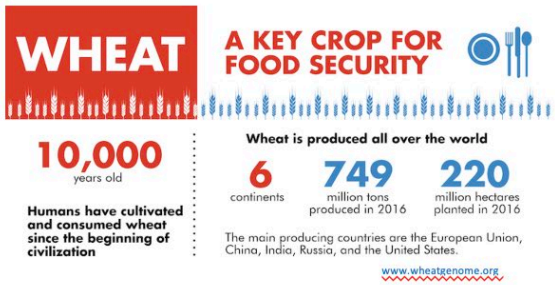
1.0 Introduction

In multicellular organisms, transcription factors (TFs) account for a significant fraction of protein coding genes (from 6 to 10%)[1]. They are critical macromolecules, which bind and alter chromatin architecture, regulating the first step of genome expression – transcription. As they are primary regulators of gene expression, their functional impacts are tremendously diverse. Just to highlight a few of them, in plant and fungal organisms, they control growth and stress response (such as drought) critical for bioenergy/biomass applications[2,3]. Despite the importance of these proteins and significant efforts of the research in this area, the vast number of them remain rather uncharacterized. The focus of this study has been centered on wheat transcription factors. As climate changes and global supply deficit, a need to bioengineer high-yield and drought-resistant crops or any other plants is crucial. The knowledge of such gene regulatory control (which includes characterization of these proteins) is the key to success for any biosystem engineering efforts in the future.

1.1 Transcription factors in wheat

There are 3,606 genes mainly annotated as putative TFs in wheat (http://planttfdb.gao-lab.org/index.php?sp=Tae), which are further classified into 56 families. Sadly, less than 10 out of 3,606 have been validated and characterized experimentally today[1]. While some of wheat TF families have very close homologs in *A. thaliana* and their function and DNA binding preferences can be somewhat predicted from there, a plant-specific TF family known as GeBP[4] represents unconventional Leu-zipper TF proteins that remain very poorly characterized in any plant. This TF family have potential roles in the regulation of cytokinin response5 – a process in plants that influences plant, axillary bud, and cellular growth, development, differentiation, as well as affecting apical dominance and leaf senescence. Wheat has 12 GeBP TFs that have been catalogued as of 2021, all annotated as uncharacterized (Fig.1b and 1c). These genes have been chosen as pilot targets for the proposed research efforts below since they are a small but complete family of proteins and with high scientific importance.

a)



b)

Species	TF ID	Description
Triticum aestivum (12)	TRAES3BF060500360CFD_t1	GeBP family protein
	Traes_2AS_8C5E619C6.1	GeBP family protein
	Traes_2BS_EEE045515.2	GeBP family protein
	Traes_2DS_6BAF1825F.1	GeBP family protein
	Traes_3AS_4446297C5.1	GeBP family protein
	Traes_4AS_928691B39.1	GeBP family protein
	Traes_4BS_1281828D6.2	GeBP family protein
	Traes_4BS_96AE357C6.1	GeBP family protein
	Traes_4DL_21C202E37.1	GeBP family protein
	Traes_5BL_4CD61BF76.1	GeBP family protein
	Traes_5DL_386343EA1.1	GeBP family protein
	Traes_6DL_BA85EE462.1	GeBP family protein

c)

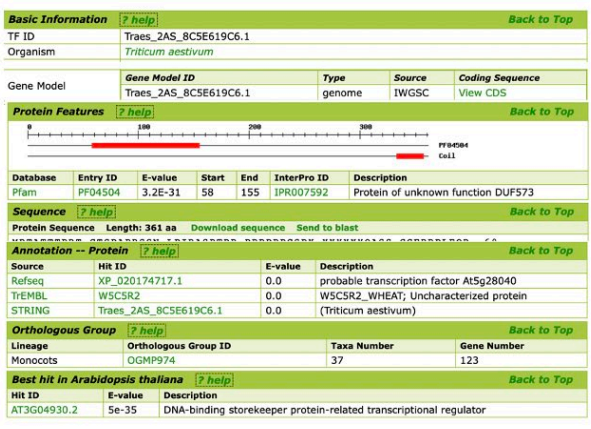


Figure 2. Wheat and its transcription factors. a) The importance of wheat. The image is sourced from wheat genome sequencing consortium. b) 12 genes from *T. aestivum* that classified as GeBP family TFs. c) Current annotation of Taes_2AS_8C5E619C6.1 gene

as an example. The information has been pulled out from PlantTFDB v5.0 (http://planttfdb.gao-lab.org/family.php?sp=Tae&fam=GeBP).

1.2 DAP-seq technology

Despite a long-standing interest of researchers in TF biology, the validation and experimental characterization of DNA-binding proteins such as TFs remains rather challenging. Gold-standard experimental approaches include chromatin immunoprecipitation (ChIP)-based assays to determine the specific genomic binding sites of TFs and their relationships to transcriptional regulation. A low throughput of ChiP method is caused by its requirement to raise specific antibodies and its dependency on the conditions used during crosslinking which limits identifying global binding profiles.

Based on the vast number of uncharacterized TF, there has been a shift to use alternative and high-throughput TF validation and functional characterization methods to overcome the shortcomings of ChIP. In particular, the DAP-seq method[5] for TF characterization stands out as a powerful alternative with increased specificity as compared to other well-known high-throughput methods. The method combines affinity-purified TFs synthesized via cell-free protein methodologies with next generation sequencing using genomic DNA library.

For the proposed study, the plan was to establish the following workflow (Figure 2). First, high-quality wheat berries will be purchased from Palouse Brand foods, which is a local WA state farm. The genomic DNA will be extracted using standard procedures with ThermoFisher's PureLink Plant Total DNA Purification kit. Extracted DNA material will be further fragmented and ligated with Illumina-based sequencing adaptors to generate a final DNA library for binding experiments. In parallel, TF genes will be synthesized by custom gene services from Genscript, and the transcription factors will be further expressed and purified using the Protemist DTII automated wheat germ cell-free protein synthesis platform at EMSL. Purified cell-free TFs will then be incubated with the generated DNA library, and TF-bound DNA will be pulled down by the affinity tag on the TF. The bound DNA will be further released, PCR amplified and sequenced on NextSeq550 instrument located at EMSL.

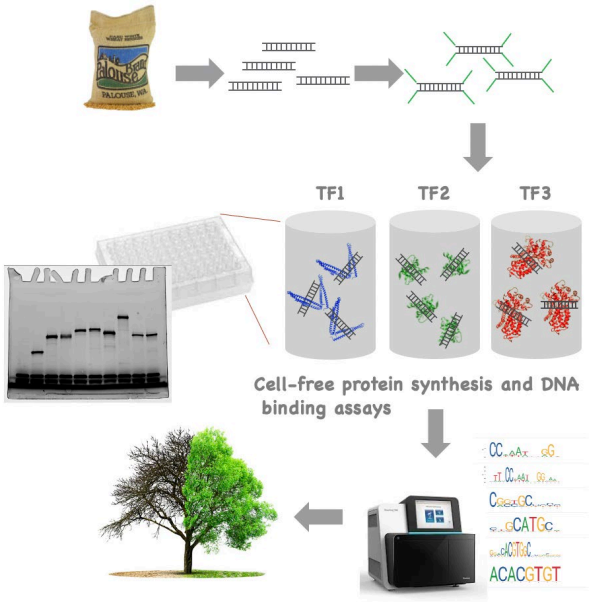


Figure 2. A simplified workflow of DAP-seq. TF1, TF2 and TF3 stand for transcription factors 1, 2 and 3. PDB IDs (1GD2, 4AYA, 2VEQ) were used as TF models.

2.0 Results

2.1 Sourcing of genomic DNA from wheat

Wheat berries were purchased from Palouse Brand foods, which is a local WA state farm. The genomic DNA was extracted using standard procedures with ThermoFisher's PureLink Plant Total DNA Purification kit. The quality of the resulting genomic DNA was validated using agarose gels (Figure 3). Extracted DNA was further fragmented using a Covaris S2 Ultrasonicator, further repaired using End-it DNA End-repair kit (Lucigen) and A-tailed using Klenow fragment (NEB) using the guidelines described in the DAP-seq method paper[5]. The resulting DNA was then ligated with the Y adaptor for PCR amplification. The quality of the resulting library was confirmed by qPCR amplification to verify adaptor ligation.

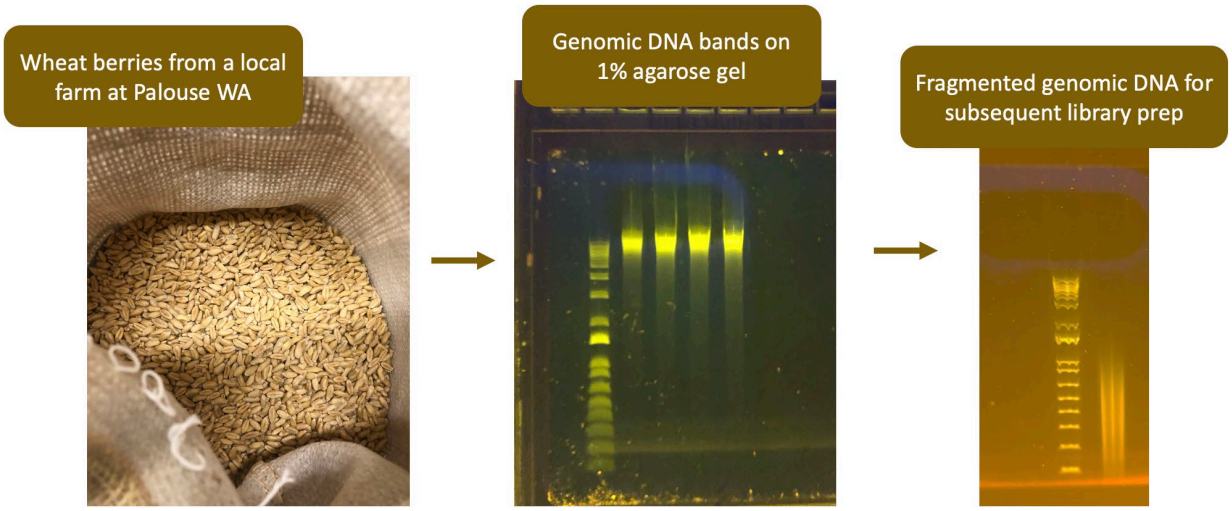


Figure 3. Genomic DNA extraction and fragmentation.

2.2 Cell-free protein synthesis of wheat TFs

Gene synthesis and cloning into a pEU vector (designed for WGCF platform) was outsourced to Genscript. A 3XFLAG-tag to the N-terminus of DNA constructs was incorporated for purification purposes. Obtained plasmids were used as DNA templates in the transcription, where synthesized mRNA is then fed into translation reactions. Translation reactions were conducted first overnight at mini scale (~20 ul volume) in the presence of fluorophore-labelled lysine-charged tRNA (Promega), which allows incorporation of fluorophore-labeled lysines (<1%) into newly synthesized proteins. The latter allows rapid in-gel detection by fluorescence of the protein product in the crude mixture with no need for its purification[6]. The products were analyzed by SDS-PAGE and GE Typhoon FLA 9000 Gel Imaging Scanner (Figure 4). The above experiments provided a rapid screening technology to verify protein expression and integrity and all target protein showed good expression levels with no evidence of premature termination. The translation reactions were then repeated in the absence of fluor-tRNA to get native (unmodified) TFs for subsequent DNA binding assays. After the completion of protein synthesis, TFs were purified using ANTI-FLAG M2 Magnetic Beads (Sigma) and washed several times with binding buffer to remove nonspecific binders. Then, each TF was incubated with 100 ng of genomic DNA library obtained in 2.1 and washed again with to remove nonspecifically bound DNA. The bound DNA was further released by

heat and PCR amplified using illumina Index primers (from Invitrogen). As each sample was amplified using a unique Index primer, DNA was then pooled together and routed for sequencing.

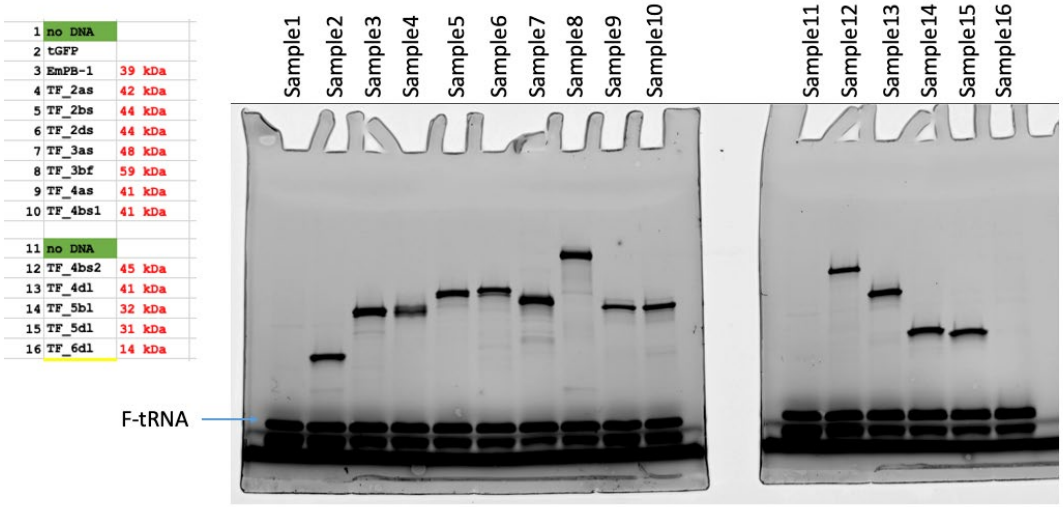


Figure 4. Cell-free protein expression screening of newly synthesized TFs in crude mixtures by fluorescence. F-tRNA stands for fluorophore-labeled lysine-charged tRNA. tGFP and EmBP-1 are used as positive controls. TF_* designate 12 chosen proteins from *T. aestivum* that classified as GeBP family TF. The expression of TF_6dl (sample 16) is not clearly visible due to its small size and a mobility overlap with F-tRNA.

2.3 DNA sequencing

An initial DAP-seq run (Run #1) was targeting around 20-30M of raw reads per sample, so a single sequencing run of NextSeq550 (200-400M) per cell was conducted first to analyze all TFs at once. Sample-specific reads were distinguished by the Illumina barcode. Based on a very large size of wheat genome (17Gb) and its high percentage of repetitive sequences, the data processing of this run showed that, unfortunately, we don't achieve enough sequencing depth for the genome size of this complexity (see below). Due to limited budget if this proposal, we had to narrow further sequencing runs only to a select number of target proteins with target ~200M reads per sample. Run #2 was focused TF_5bl and TF_5dl and Run #3 was on negative_control(no DNA) and EmBP. All raw reads are archived and can be found on NEXUS under these **Upload IDs** (2516155, 2635542, 2635703).

2.4 DAP-seq data analysis

The data processing and quality metrics for DAP-seq are similar to ChIP-seq data analysis. So ChIP-seq bioinformatics tools are applied for data processing. The ChIPQC report for Run #1 showed that we don't reach enough sequencing depth but proteins TF_5bl and TF_5dl look the most promising (Table 1 and Figure 5).

ID	Tissue	Factor	Condition	Replicate	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%
nothing	NA	NA	NA	NA	220509	0	137	279	0.28	2.7	NA
Sample11			4bs2	1	166560	0	128	268	0.077	0.94	NA
Sample13			5bl	1	386668	0	137	279	0	5.3	1.4
Sample14			5dl	1	359819	0	136	274	0.13	2.7	0.85
Sample15			6dl	1	157853	0	130	261	0.088	1.4	0
Sample3			pos	1	70479	0	127	299	0.44	0.76	0
Sample6			2ds	1	97991	0	129	260	0.93	0.95	0
Sample8			3bf	1	307775	0	124	251	0.038	1.7	0

Table 1. Summary of ChIP-seq filtering and quality metrics for Run #1.

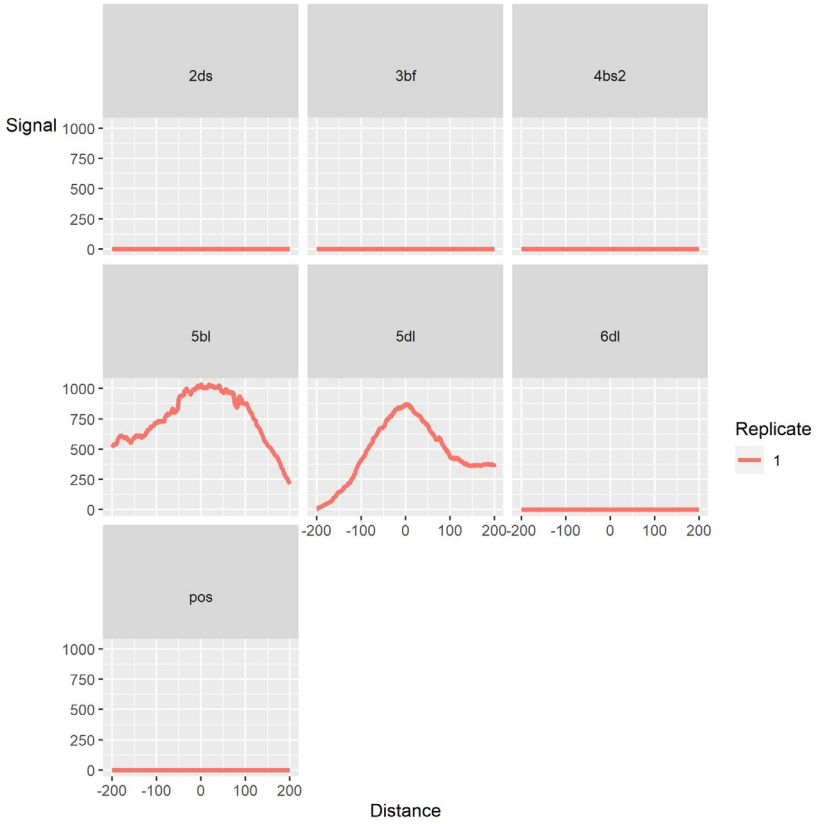


Figure 5. Plot of the average signal profile across peaks. TF_5bl and TF_5dl look the most promising targets for further sequencing.

The ChIPQC report for Runs #2 and #3 (Table 2 and Figure 6):

ID	Tissue	Factor	Condition	Replicate	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%
bigNeg	NA	NA	NA	NA	2281640	0	133	272	0.1	6.9	NA
Sample1	wheat		DAPSEQ13	1	3680711	0	132	265	-0.024	14	4
Sample2	wheat		DAPSEQ14	1	3430425	0	135	274	-0.07	11	1.7
Sample3	wheat		emBP	1	505067	0	120	246	0.17	0.7	0

Table 2. Summary of ChIP-seq filtering and quality metrics for Runs #2 and #3.

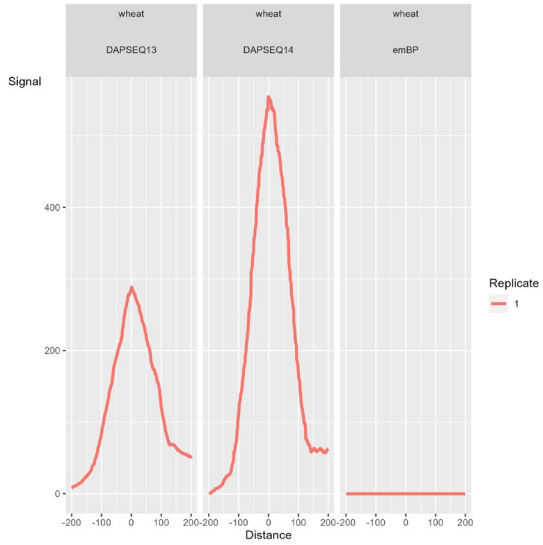


Figure 6. Plot of the average signal profile across peaks.

2.5 Overall results

Identified DAP-seq peaks for TF_5bl and TF_5dl were significantly enriched in the promoter regions of genes in comparison with a negative control, which is expected for a successful DAP-seq experiment. Enriched peaks were further analyzed by a MEME motif discovery software, the results of which are shown in Figures 8 and 9. Both TFs demonstrate very specific DNA binding profiles with very high E-values. We also analyzed the locations of those peaks to identify genes that they might regulate, and those gene lists are summarized in Table 3. Both transcription factors are the closest homologs based on phylogeny, so overlapping function is expected. Interestingly, both of them involve in rRNA biogenesis. While TF_5bl involve in the regulation of all rRNA subunits, TF_5dl is specific to 18S rRNA.

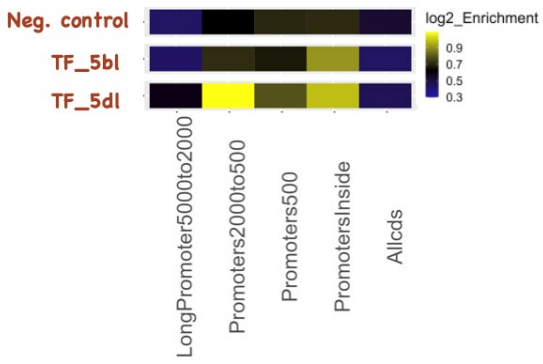


Figure 7. Heatmap of log2 enrichment of reads in genomic features.

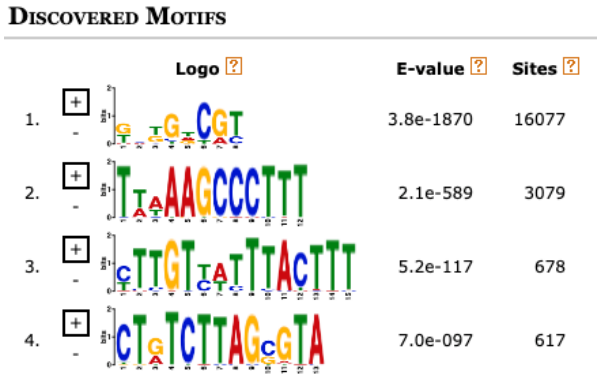


Figure 8. Discovered motifs for TF_5bl.

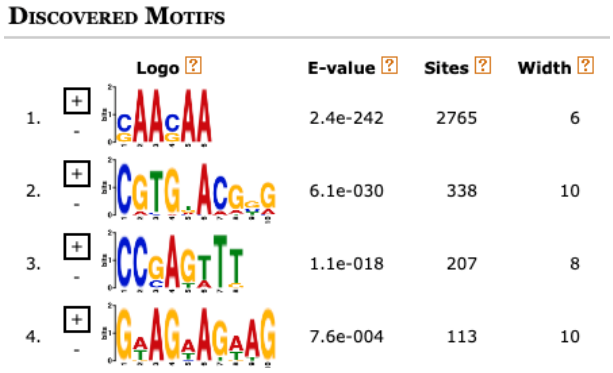


Figure 9. Discovered motifs for TF_5dl.

Transcription factors	Gene ID	Gene annotation
TF_5bl	LOC123073733	18S ribosomal RNA
	LOC123073958	18S ribosomal RNA
	LOC123087585	*uncharacterized protein
	LOC123141118	GATA zinc finger domain-containing protein 15-like
	LOC123076302	*uncharacterized protein
	LOC123150351	5.8S ribosomal RNA
	LOC123137730	*lncRNA
	LOC123152556	18S ribosomal RNA
	LOC123152693	28S ribosomal RNA
	LOC123152675	28S ribosomal RNA
	LOC123183523	5S ribosomal RNA
	LOC123158638	*uncharacterized protein
	LOC123181840	alpha-glucosidase 2-like
	LOC123041082	*lncRNA
	LOC123052285	putative polyol transporter 1
	LOC123058560	*uncharacterized protein

	LOC123059340	*lncRNA
	LOC123089030	18S ribosomal RNA
	LOC123088691	5.8S ribosomal RNA
	LOC123084114	*uncharacterized protein (similarities with Phosphatases)
	LOC123091979	*lncRNA
	LOC123100832	5S ribosomal RNA
	LOC123116914	glycosyltransferase-like protein gnt13
	LOC123116915	glycosyltransferase-like protein gnt13
	LOC123113624	mRNA (no alphafold model)
	LOC123140164	5.8S ribosomal RNA
	LOC123140203	U1 spliceosomal RNA
	LOC123166325	3'(2'),5'-bisphosphate nucleotidase-like
	LOC123171834	5.8S ribosomal RNA
	LOC123171826	28S ribosomal RNA
	LOC123172547	BTB/POZ and MATH domain-containing protein 1-like
	LOC123174687	18S ribosomal RNA
	LOC123174684	5.8S ribosomal RNA
	LOC123174677	18S ribosomal RNA
	LOC123174678	18S ribosomal RNA
	LOC123175213	28S ribosomal RNA
	LOC123175593	18S ribosomal RNA
	LOC123175686	18S ribosomal RNA
	LOC123175756	5S ribosomal RNA
	LOC123176443	18S ribosomal RNA
	LOC123176563	5S ribosomal RNA
	LOC123176769	18S ribosomal RNA
	LOC123177239	*lncRNA
	LOC123177319	5.8S ribosomal RNA
	LOC123178007	5.8S ribosomal RNA
	LOC123178029	5.8S ribosomal RNA
	LOC123178503	5.8S ribosomal RNA
	LOC123178642	5.8S ribosomal RNA
	LOC123178805	5.8S ribosomal RNA
	LOC123178862	5.8S ribosomal RNA
TF_5dl	LOC123073777	18S ribosomal RNA
	LOC123073703	18S ribosomal RNA
	LOC123073733	18S ribosomal RNA
	LOC123073739	18S ribosomal RNA
	LOC123073958	18S ribosomal RNA

	LOC123137730	*lncRNA
	LOC123152527	18S ribosomal RNA
	LOC123059340	*lncRNA
	LOC123065121	*uncharacterized protein
	LOC123089030	18S ribosomal RNA
	LOC123087447	serine/threonine-protein phosphatase PP1-like
	LOC123105709	Bowman-Birk type proteinase inhibitor PVI-3(2)-like
	LOC123155118	small nucleolar RNA R24
	LOC123170386	cilia- and flagella-associated protein 251-like
	LOC123174677	18S ribosomal RNA
	LOC123174678	18S ribosomal RNA
	LOC123175334	disease resistance protein RGA5-like
	LOC123175593	18S ribosomal RNA
	LOC123175686	18S ribosomal RNA
	LOC123176769	18S ribosomal RNA

Table 3. Identified genes near Dap-seq peaks.

3.0 References

1. UniProt, Consortium. "Uniprot: The Universal Protein Knowledgebase in 2021." Nucleic Acids Res 49, no. D1 (Jan 8 2021): D480-D89. https://doi.org/10.1093/nar/gkaa1100. https://www.ncbi.nlm.nih.gov/pubmed/33237286.

2. Xie, Z., T. M. Nolan, H. Jiang, and Y. Yin. "Ap2/Erf Transcription Factor Regulatory Networks in Hormone and Abiotic Stress Responses in Arabidopsis." Front Plant Sci 10 (2019): 228. https://doi.org/10.3389/fpls.2019.00228. https://www.ncbi.nlm.nih.gov/pubmed/30873200.

3. Chen, L., Q. Tong, C. Zhang, and K. Ding. "The Transcription Factor Fgcrz1a Is Essential for Fungal Development, Virulence, Deoxynivalenol Biosynthesis and Stress Responses in Fusarium Graminearum." Curr Genet 65, no. 1 (Feb 2019): 153-66. https://doi.org/10.1007/s00294-018-0853-5. https://www.ncbi.nlm.nih.gov/pubmed/29947970.

4. Chevalier, F., D. Perazza, F. Laporte, G. Le Henanff, P. Hornitschek, J. M. Bonneville, M. Herzog, and G. Vachon. "Gebp and Gebp-Like Proteins Are Noncanonical Leucine-Zipper Transcription Factors That Regulate Cytokinin Response in Arabidopsis." Plant Physiol 146, no. 3 (Mar 2008): 1142-54. https://doi.org/10.1104/pp.107.110270. https://www.ncbi.nlm.nih.gov/pubmed/18162594.

5. Bartlett, A., R. C. O'Malley, S. S. C. Huang, M. Galli, J. R. Nery, A. Gallavotti, and J. R. Ecker. "Mapping Genome-Wide Transcription-Factor Binding Sites Using Dap-Seq." [In English]. Nature Protocols 12, no. 8 (Aug 2017): 1659-72. https://doi.org/10.1038/nprot.2017.055. <Go to ISI>://WOS:000405918500003.

6. Novikova, I. V., M. Zhou, C. Du, M. Parra, D. N. Kim, Z. L. VanAernum, J. B. Shaw, et al. "Tunable Heteroassembly of a Plant Pseudoenzyme-Enzyme Complex." ACS Chem Biol 16, no. 11 (Nov 19 2021): 2315-25. https://doi.org/10.1021/acschembio.1c00475. https://www.ncbi.nlm.nih.gov/pubmed/34520180.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov