A Rice Promoter Protein Binding Microarray for Cis-Acting Elements for Rice Transcription Factors

JOUNG SUG KIM
Myongji University
SongHwa Chae
Myongji University
Kyong Mi Jun
GreenGene BioTech Inc.
Gang-Seob Lee
National institute of argicultural sciences
Jong-Seong Jeon
Kyung Hee University
Kyungdo Kim
Myongji University
Yeon-Ki Kim (kim750a11@gmail.com)
https://orcid.org/0000-0002-4507-341X

Short communication

Keywords:

Posted Date: May 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-28489/v1

Abstract

Background

Transcription factors (TFs) regulate the expression of genes at the transcriptional level by binding a specific DNA sequence. Thus, predicting the DNA-binding motifs of TFs is one of the most important areas for the functional analysis of TFs in the postgenomic era. Although many methods have been developed for this challenge, there are still many TFs with unknown DNA-binding motifs.

Findings

In this paper, we designed an rice (*Oryza sativa*)-specific protein binding microarray (RPBM), and its probes are 40 bp long with 20 bp of overlap; there are 49 probes spanning the 1 kb promoter region before the translation start site of each gene. To confirm the efficiency of RPBM technology, we selected two TFs, *OsWOX13* and *OsSMF1*. We identified the ATTGATTG DNA-binding sequence and 635 putative target genes of *OsWOX13*. OsSMF1 bound to GCTGACTCA and GGATGCC sequences and bound especially strongly to CCACGTCA. A total of 932 putative target genes were identified for *OsSMF1*.

Conclusions

RPBM can be applicable in the analysis of DNA-binding motifs for TFs where binding is evaluated in extended natural promoter regions. The analysis can also be applicable to TFs that have single or multiple binding motifs. The technology might even be expanded for application to TFs that are heterodimers or form higher-order complexes.

Introduction

Transcription factors (TFs) play a pivotal role in the regulation of gene expression by binding to their cognate motifs in the promoter regions. For many years, This binding activity has been investigated by biochemical assays such as electrophoretic mobility shift assays (EMSAs), nitrocellulose filter binding assays, footprinting assays, and yeast one-hybrid system assays (Hellman and Fried 2007; Helwa and Hoheisel 2010). However, such approaches are generally laborious and slow, and many TFs still remain uncharacterized.

High-throughput methods such as chromatin immunoprecipitation (ChIP)-chip, ChIP followed by sequencing (ChIP-seq), and protein binding microarrays (PBM) have been developed with the availability of whole-genome sequences and advances in microarray technology (Barski et al. 2007; Ren et al. 2000; van Steensel et al. 2001; Wang et al. 2008). PBM has some advantages compared to ChIP. For example, PBM is not dependent on the availability of highly specific antibodies and does not need to use cross-linking reagents, eliminating the risk of cross-linking artifacts. Protein binding microarrays (PBMs) were introduced to conveniently determine protein-DNA interactions in vitro (Berger and Bulyk 2009). PBMs were improved by adapting de Bruijn sequences and in situ synthesis of DNA oligonucleotides on slides

(Berger et al. 2006). The de Bruijn sequences represent not only all contiguous 10-mers but also all 10-mers with a gap size of 1 nucleotide. The whole-genome yeast intergenic microarray for PBM was prepared by spotting double-stranded DNA (Zhu et al. 2009). Recently, in efforts to characterize the DNA-binding activity of transcription activator-like effectors (TALEs), which are secreted by the bacteria Xanthomonas via their Type III secretion system function and function as virulence factors, a custom PBM was developed (Anderson et al. 2020; Rogers et al. 2015). TALE–DNA interactions were comprehensively assayed in this PBM in which ~ 5,000–20,000 unique DNA sequences per effector protein were spotted.

Identification of genomic regulatory elements has led to the construction of the databases TRANSFAC (Wingender et al. 1996), GRASSIUS (Yilmaz et al. 2009), PlnTFDB (Perez-Rodriguez et al. 2010), UniPROBE (Hume et al. 2015), and PlantTFDB (Jin et al. 2017). In particular, PlantTFDB was constructed based on a collection of 156 plant species with sequenced genomes. Recent advances in ChIP-seq have provided powerful ways to identify genome-wide profiling of DNA-binding proteins and histone modifications, leading to databases such as ChEA, CistromeMap, and ChIPBase (Lachmann et al. 2010; Qin et al. 2012; Yang et al. 2013).

Previously, we designed a PBM, denoted Q9-PBM, in such a way that target probes are quadruples of all possible 9-mer combinations (Kim et al. 2009). A total of 131,072 features were selected from the 262,144 reads after consideration of the reverse complimentary sequences because double-stranded DNA has a bidirectional aspect. The quadruple sequences can provide highly consistent and concrete results for consensus binding motifs. Q9-PBM employs DsRed fluorescent protein, which eliminates multiple wash and hybridization steps. Q9-PBM confirmed the well-known DNA-binding sequences of Cbf1 and CBF1/DREB1B, and it was also applied to elucidate the unidentified cis-acting elements of the OsNAC6, MYB44, and OsSMF1 rice TFs (Kim et al. 2009). These PBMs can identify binding motifs but could be limited by the number of designed nucleotide sequences in terms of oligomer length (9 or 10). It also opens the possibility that the binding sites of TFs can be searched in gene-specific promoters.

To overcome the limitations from the number of nucleotides and investigate the binding activity in the promoter region, we designed a rice (*Oryza sativa)*-specific PBM (RPBM) in such a way that the 1 kb gene-specific promoter region was covered by overlapping 40 nt long probes. The single oligomers on the microarray were subjected to polymerase chain reaction (PCR) to form double strands, and then the binding sites of the TFs OsWOX13 and OsSMF1 were tested. OsWOX13 is known to preferentially bind to a ATTGATTG DNA-binding motif, while OsSMF1 has multiple DNA-binding motifs such as GCN4 [TGA(G/C)TCA], ACGT [CCACGT(C/G)], and ATGA [GGATGAC] (Kim et al. 2017; Minh-Thu et al. 2018). Using this RPBM, we confirmed the DNA-binding motifs and identified putative target genes and of OsWOX13 and OsSMF1.

Results

Design of a RPBM

Probes for the RPBM were designed from promoters of genes deposited in the IRGSP RAP2 database (http://rapdb.lab.nig.ac.jp). A probe is 40 bp long, covering a gene-specific region, with 20 bp for an annealing site for PCR. Each gene-specific region overlapped 20 bp, and 49 probes spanned the 1 kb promoter region before the translation start site of each gene (Fig. 1). Considering the ambiguity of annotation, the first probe of genes without 5'-UTR or with a 5'-UTR longer than 200 bp was designed from the 5' upstream region including methionine. In this way, 954,520 probes were designed from 19,480 genes among 31,439 genes. Each target probe was followed by a sequence complementary to a primer sequence (5'-CGGAGTCACCTAGTGCAG-3') and was connected by a 5 nt thymidine linker on the microarray.

Analysis of signal intensities

The full-length *OsWOX13* and *OsSMF1* cDNAs were fused at the N-terminus to the *DsRed* fluorescent protein gene. Purified recombinant OsWOX13 and OsSMF1 proteins fused with DsRed:6xHis were hybridized to the RPBM as described in the Methods section. Then, the consensus binding motifs were determined based on signal strength (Kim et al. 2017; Kim et al. 2009).

A rank-ordered signal distribution showed a steep slope on the left followed by a heavy right tail for RPBM. As the probes in the steep slope region differed in only one base, we assumed that the signal distribution was due to specific interactions between the proteins and features on the microarray. Two independent linear models, $y = ax + b$, were applied in the steep and heavy right tail regions using the R statistical language. In OsWOX13, the slope and y-axis intercept of the steep sloping region were − 14.7 and 66,570.6, respectively, while those of the heavy tail region were − 0.0043 and 3,144, respectively (Additional file 1: Figure S1a). The number of strong binding probes from the deep slope was 34,778 (Additional file 2: Table S1).

OsSMF1 gave a similar rank-ordered signal distribution, showing a steep slope on the left followed by a heavy right tail. The slope and y-axis intercept of the steep slope region were − 25.1 and 64,928.8, respectively, while those of the heavy tail region were − 0.0207 and 2283, respectively (Additional file 1: Figure S1b). For OsSMF1, the extrapolated intensity of the heavy right tail was 3,137. The number of target probes for which the intensity was higher than this value was 38,654 (Additional file 3: Table S2). These results suggest that the binding of transcription factors and their cognate binding sites in RPBM as stable as to those found in Q9-PBM. In addition, the probe design from the promoter regions overcome potential complexities due to concatemers of target sites.

Identifying putative target genes of OsWOX13 by RPBM

To find the DNA-binding motif of OsWOX13, a 40 bp probe was split into 9-mers, and each oligomer was given the pseudointensity of the probe. The process was repeated with a base shift, and finally, a probe gave 32 9-mers (Additional file 1: Figure S2a). The strongly binding feature probes (34,778) give 198,384 distinct 9-mers from 1,177,280 of the total frequency (Fig. 2a). We found that 4–5 consecutive G- or C-rich oligomers (3,148) exhibited nonspecific binding and discarded them from the subsequent analysis.

The average intensity and frequency of 9-mers were 21,193.0 and 5.9, respectively. These 9-mers were sorted according to their intensities, and GATTGATTG had the highest intensity of 37,706 with a frequency of 280 (Additional file 4: Table S3). To find a consensus sequence, cluster analysis was performed in such a way that any 9-mer with a 5 nt long sequence matching the template of the highest intensity belonged to a group. The 1,028 9-mers formed a cluster with GATTGATTG as a template. These top 20 9-mers ranked by intensity contained one or more ATTG sequences (Table 1). The occurrences of nucleotides at each position were shown in a position weight matrix (PWM) by clustering of these 9-mers (Fig. 3a). Web logo (weblogo.berkeley.edu) gave ATTGATTG (Fig. 3b). In addition, mutation analysis was conducted by changing bases in each ATTGATTG (Fig. 3c). A base-mutated sequence gave a maximum decrease at the 4th nt, G, and a minimum at the 1st nt, A (10756.4 and 10139.6, respectively). The Wilcoxon-Mann-Whitney test using the ranks with and without the motif clearly showed that the ATTG motif (8-mer) is the binding motif of the OsWOX13 TF. Similarly, oligomer frequency and point mutations at distinct positions were also analyzed with 5-, 6-, 7-, 8-, and 10-mers (Additional file 1: Figure S3). These analyses showed that ATTGATTG is the binding motif of the TF.

Table 1

List of 9-mers highly ranked by intensity and containing the ATTGATTG sequence

Rank a)	9-Mer b)	Intensity_ave c)	Frequency_total d)	Int_ave*Freq_tot e)	Occur_diff_pos f)
1	GATTGATTG	37706.65	280	10557862	31
2	ATTGATTGA	37026.91	270	9997266	31
3	TTGATTGAT	36995.96	343	12689615	31
4	TGATTGATT	36509.81	401	14640432	31
5	ATTGATTGG	36080.41	158	5700705	29
6	TAATTGATT	35690.72	274	9779257	31
7	GATTGACAG	35273.34	41	1446207	17
8	GATTGATTA	35246.55	126	4441065	29
9	ATTGATTGC	35134.35	124	4356660	28
10	TGATTGATG	34743.81	183	6358117	30
11	GTGATTGAT	34695.00	139	4822605	30
12	TGATTGGCG	34639.79	34	1177753	18
13	TATTGATTG	34572.04	95	3284344	23
14	CTGATTGAT	34351.64	121	4156549	26
15	TGATTGATA	34202.91	126	4309567	30
16	AATTGATTG	34058.06	193	6573205	28
17	ATGATTGAC	33941.15	60	2036469	21
18	GACTGATTG	33741.80	35	1180963	17
19	GATTGATGG	33686.23	74	2492781	27

a) Rank order by the intensity

b) 9-Mers were obtained by a base shift on a 40 nt long feature probe, and finally, the probe gave 32 distinct 9-mers.

c) Intensities were averaged over all the feature probes containing the corresponding 9-mer sequence.

d) Total number of frequencies of the 9-mer from the 34,778 strongly binding feature probes.

e) Total intensities for column c * column d

f) Distinct positions of 9-mers in the 40 nt probes. The highest value (near 32) suggests that the 9-mers were obtained from all the positions by a base shift in the probes.

Rank a)	9-Mer b)	Intensity_ave c)	Frequency_total d)	Int_ave*Freq_tot e)	Occur_diff_pos f)
20	ATTGATAGC	33661.22	27	908853	16
a) Rank order by the intensity					
b) 9-Mers were obtained by a base shift on a 40 nt long feature probe, and finally, the probe gave 32 distinct 9-mers.					
c) Intensities were averaged over all the feature probes containing the corresponding 9-mer sequence.					
d) Total number of frequencies of the 9-mer from the 34,778 strongly binding feature probes.					
e) Total intensities for column c * column d					
f) Distinct positions of 9-mers in the 40 nt probes. The highest value (near 32) suggests that the 9-mers were obtained from all the positions by a base shift in the probes.					

An extended motif was constructed using ATTGATTG as a template by adding a base in either the 5’ or 3’ direction (Fig. 3c). For example, GATTGATTG (-1) was chosen from analysis of the 8-mer, which was extended in the 5’ direction with the base G to make GATTGATTG, and repeated analysis showed that T is the farthest in the 5’ direction (-2). Similarly, G and T were added in the 3’ positions of + 1 and + 2, respectively, which gave TGATTGATTGGT. These data were confirmed by counting the actual frequency of nt flanking ATTGATTG. A total of 3,243 genes in rice contained the ATTGATTG motif in the 1 kb promoter regions, and 29,379 genes were retrieved from RAP-DB (http://rapdb.dna.affrc.go.jp/). The preferred nucleotides were searched (Additional file 1: Figure S4) for in flanking sequences around ATTGATTG. A and T were preferable at -3 and − 2, and G and A were preferable at the − 1 position. In contrast, A/G was preferable at the + 1 position, and T was preferable at the + 2 and + 3 positions.

Among 34,778 probes, 646 probes contained the ATTGATTG motif (Fig. 2a, Additional file 5: Table S4). From these probes, we identified 635 putative target genes of *OsWOX13*. Gene ontology (GO)-based functional enrichment analysis of the above candidate genes was performed by the web-based tool AgriGO (http://bioinfo.cau.edu.cn/agriGO/analysis.php). The results revealed that among the 635 genes, 501 were annotated, of which 10 GO terms showed significant differences compared to those in the *Oryza sativa* database as a background reference (Table 2). The most enriched terms of macromolecule metabolic process (GO:0043170) were significantly enriched, including protein (GO:0019538), carbohydrate (GO:0005975), lipid (GO:0006629), and nucleobase (GO:0006139) (Table 2). Categories such as death (GO:0016265) and response to stress (GO:0006950) were also highly enriched. These results were in line with the observation in a previous paper that compared to control plants, rice plants overexpressing *OsWOX13* showed early flowering and drought tolerance (Minh-Thu et al. 2018).

Table 2
Statistical analysis of putative target genes of OsSMF1 and OsWOX13 by AgriGO

GO		**OsWOX13** [a)]		**OsSMF1** [b)]	
ID	Term	Query item	p-value	Query item	p-value
GO:0043170	macromolecule metabolic process	92	1.30E-46	154	2.10E-88
GO:0019538	protein metabolic process	48	2.60E-27	75	5.40E-45
GO:0005975	carbohydrate metabolic process	20	6.10E-15	19	1.50E-11
GO:0006629	lipid metabolic process	9	0.0000014	15	5.30E-11
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	41	9.30E-17	79	2.70E-40
GO:0051171	**regulation of nitrogen compound metabolic process**	27	3.00E-14	**41**	3.70E-22
GO:0045449	regulation of transcription	27	2.60E-14	41	3.00E-22
GO:0006950	**response to stress**	**21**	2.80E-18	17	1.40E-11
GO:0016265	death	7	0.0011	N/A	N/A
GO:0007154	cell communication	6	0.0024	N/A	N/A
GO:0016051	**carbohydrate biosynthetic process**	N/A	N/A	**6**	0.00002
GO:0034660	ncRNA metabolic process	N/A	N/A	6	0.00000013
GO:0051276	chromosome organization	N/A	N/A	18	3.90E-21

The 635 putative target genes of OsWOX13 (a) with the ATTGATTG motif and the 932 putative target genes of OsSMF1 (b) with the GCCACGTCA motif were chosen and subjected to gene ontology analysis using AgriGO (http://bioinfo.cau.edu.cn/agriGO/analysis.php).

To verify putative targets of OsWOX13, we selected *Hd1-3* (Os08g0536300), for which a probe (Os08g0536300_14, AATATAACGAAACATGCAATCAATCAAAATGTTGGGAAGG) contains the ATTG motif (Fig. 3d and Table S1). We assayed its binding specificity to recombinant OsWOX13 by EMSA using carboxyfluorescein (FAM)-labeled double-stranded oligonucleotide probes. The binding of OsWOX13 to the 40 bp probe with the ATTG motif was detected as lagging bands (Fig. 3d). These results confirmed the ATTG motif that has previously been identified using an analysis based on Q9-PBM (Minh-Thu et al. 2018).

.

Identifying the DNA-binding motif of OsSMF1 by RPBM

OsSMF1 reportedly binds multiple cis-elements (Kim et al., 2017). To test this, RPBM was applied to find the binding motif of OsSMF1, and 32 9-mers were extracted from a 40 bp long probe in the same manner as that for OsWOX13. The 15,394 probes gave 178,857 distinct oligomers, and the total frequency was 492,608 (Fig. 2b). The average intensity and frequency of 9-mers were 21,725.2 ± 11,270.6 and 2.75, respectively. In contrast to OsWOX13, several groups were identified by initial cluster analysis, suggesting that OsSMF1 binds several motifs. Thus, the distinct 9-mers with frequencies four times the average frequency (over 11) were sorted according to the value of the intensity multiplied by the frequency, and then the 9-mers were narrowed down to 648 in total (Table 3, Additional file 6: Table S5). This list gave 4 clusters, GCCACGTCA, ACGTAAGCG, GCTGACTCA, and AGGATGCCA, with 335, 24, 31 and 24 9-mers, respectively (Additional file 7: Table S6, Fig. 4A). In addition, these results show that the cluster of GCCACGTCA is predominant and that other clusters were minor but distinct. In a previous paper, Q9-PBM and EMSAs were used to show that OsSMF1 binds the GCN4 (TGA(G/C)TCA), ACGT (CCACGT(C/G)), and ATGA (GGATGAC) motifs with three different affinities (Kim et al. 2017). GCCACGTCA and ACGTAAGCG are part of the ACGT motif, GCTGACTCA is included in the GCN4 motif, and AGGATGCCA is very similar to the ATGA motif.

Table 3
9-Mers highly ranked by intensity and containing the GCCACGTCA sequence

Rank	9-Mer	Intensity_ave	Frequency_total	Int_ave*Freq_tot	Occur_diff_pos
1	GCCACGTCA	24766.52	834	20655280	31
2	TGACGTGGC	23386.91	326	7624133	31
3	CCACGTCAG	24362.08	533	12984991	31
4	TGCCACGTC	23036.39	481	11080503	31
5	CACGTCAGC	24206.7	448	10844602	29
6	CGCCACGTC	23606.49	452	10670132	31
7	GCGCCACGT	22976.78	341	7835082	31
8	GCCACGTGG	20062.16	332	6660638	31
9	CCACGTGGC	19612.18	338	6628917	31
10	ATGCCACGT	22132.63	296	6551258	31
11	CTGCCACGT	21409.31	287	6144472	31
12	CCACGTCAT	23338.93	241	5624681	30
13	TGCCACGTA	23632.13	235	5553550	30
14	CCACGTCAC	22655.01	243	5505168	30
15	GCCACGTAG	21171.72	239	5060042	31
16	TGCCACGTG	20892.83	241	5035173	31
17	TTGCCACGT	24255.62	207	5020914	30
18	GTGCCACGT	21387.37	224	4790770	30
19	CTGACGTGG	23467.43	199	4670019	31
20	TCCACGTCA	21404.54	216	4623380	31
55	GCTGACTCA	17418.66	144	2508287	31
56	TGACTCAGC	17403.51	144	2506105	29
67	CTGACTCAG	18380.93	121	2224092	30
82	GGATGCCAC	24137.48	81	1955136	26
103	GCTGAGTCA	16726.02	99	1655876	27

* Column descriptions same as those for Table 1.

Rank	9-Mer	Intensity_ave	Frequency_total	Int_ave*Freq_tot	Occur_diff_pos
105	AGGATGCCA	23742.69	68	1614503	26
* Column descriptions same as those for Table 1.					

As the GCCACGTCA and ACGTAAGCG clusters have ACGT motifs, they were aligned together and gave a position matrix, and CCACGTCA was a main element (Fig. 4b). The feature probes containing CCACGTCA (932) are listed (Additional file 8: Table S7). The Wilcoxon-Mann-Whitney test was performed as shown for those target probes containing CCACGTCA and those without the sequence, and it gave a p-value of 0, suggesting that CCACGTCA contributed significantly to binding. To test the preferences for any nucleotide flanking CCACGTCA sequences, an extended motif was constructed using CCACGTCA as a template by adding a base in either the 5’ or 3’ direction as with OsWOX13 (Fig. 4c). Mutation analysis was performed as with OsWOX13 by changing the bases in each CCACGTCA (Fig. 4c). Intensities strongly decreased with changes to A at the 3rd position (by 10637.3) and to A at the 7th position (by 8356.0). An extended motif was constructed using CCACGTCAG as a template by adding a base in either the 5’ or 3’ direction, giving TGCCACGTCAGC. Thus, this study showed that CCACGTCA is a DNA-binding motif for OsSMF1, while the flanking sequences of this motif have no significant effect. Similarly, the intensities of the feature probes in terms of the frequency and mutations at each position were also analyzed with 5-, 6-, 7-, 8-, 10-, and 11-mers (Additional file 1: Figure S5).

Among 38,654 probes with 3137 intensity, 932 probes contained the CCACGTCA sequence, from which 890 putative target genes were identified for *OsSMF1* (Fig. 2b, Additional file 8: Table S7). When 687 genes among these candidate genes were subjected to GO analysis using AgriGO, "macromolecule metabolic process" was also highly abundant, similar to the GO analysis of OsWOX13 (Table 2). Several GO terms were enriched, such as "carbohydrate biosynthetic process (GO:0016051)", "regulation of nitrogen compound metabolic process (GO:0051171)", "ncRNA metabolic process (GO:0034660)", and "chromosome organization (GO:0051276)" (Table 2).

To verify putative targets of OsSMF1, we selected two nonapical meristem (NAM) proteins, Os01g0393100 (ONAC026) and Os05g0415400 (ONAC024), from "regulation of nitrogen compound metabolic process (GO:0051171). ONAC026 and ONAC024 were identified as target genes of OsSMF1 in a previous paper (Kim et al. 2017). Probes from the ONAC026 and ONAC024 promoters contain the ACGT and GNC4 motifs, respectively (Fig. 4a). We assayed their binding specificities to recombinant OsSMF1 by EMSA using FAM-labeled double-stranded oligonucleotides corresponding to each probe. The binding of OsSMF1 to the 40 bp probes was detected as lagging bands (Fig. 4d). This result indicated that OsSMF1 directly binds to the promoters of ONAC026 and ONAC024. These results indicate that OsSMF1 has multiple distinct motifs, with OsSMF1 binding to the ACGT (CCACGT(C/G)), GCN4 (TGA(G/C)TCA), and ATGA (GGATGAC) motifs.

Discussion

In this paper, we reported RPBM where the 1 kb promoter region is covered by overlapping 40 bp long probes. The initial signal distribution of RPBM was very similar to that of Q9-PBM, where quadruple 9-mer oligonucleotides were designed as the target probes. These results suggest that the binding of transcription factors and their cognate binding sites in RPBM as stable as to those found in Q9-PBM. The probe design from the promoter regions overcome potential complexities due to concatemers of target sites and the binding is understood in the promoter regions. The analysis of signal intensities of 5–10 oligomers, especially 9 mers, high-lighted putative binding sequences and the comparison of those signals of oligomers with point mutation at each site clearly showed strong binding sequences. Further, it is confirmed the feature probes on RPBM can be directly used in the subsequent EMSA analysis without further modification.

We first applied 9-mer-based analysis and identified the ATTGATTG DNA-binding sequence and 635 putative target genes of *OsWOX13, which has* one dominant binding site. The Plant Transcription Factor Database (Jin et al. 2017) showed that Os01g0818400 (OsWOX8) has a representative motif, CAATCAA, which has a 7 nt sequence of the reverse complement of ATTGATTG. Many homeobox-containing TFs contain ATTGATTG or parts of it in their motifs, and this is also found in the similar homeobox TFs, as shown in Os090528200 and Os03g0170600 in PlantTFDB. We also surveyed the UniPROBE database (Hume et al. 2015) and compared its entries with putative cis-elements of homeo-domain-containing TFs such as UP00615B_1 and UP00158A_1 from humans and mice, respectively. These factors also provided various GA- or AT-rich motifs. In particular, the UP00158A_1 binding site contains AATTAATTA and ATTA repeats and showed a base (A to G) difference with ATTG repeats in the ATTGATTG motif in our analysis (Minh-Thu et al. 2018).

The mode by which OsSMF1 modulates downstream TFs that are bound to GCCACGTCA and ACGTAAGCG, which include the ACGT motif, might be complex. GCTGACTCA is included in the GCN4 motif, GGATGCC is very similar to the ATGA motif, and the cluster near CCACGTCA is predominant, confirming previous results (Kim et al. 2017). Although the cis-elements are not registered in PlantTFDB, the cis-elements representative of the basic leucine zipper in the database are consistent with those found in many basic leucine zipper TFs. These TFs contain an ACGT motif in their representative binding motif. A few examples are Os01g0859500 with GATGACGTCA, Os02g0203000 with TGATGACGTGGC, Os02g0766700 with TGCCACGTGNCC, and Os03g0796900 with TGACGTGG, which is reverse complementary to CCACGTCA (Additional file 9: Table S8). These results suggest that OsSMF1 evolved to have specific functionality involving common DNA-binding activity due to the bZIP domain.

Application of the technology might even be expanded for application to TFs that are heterodimers or form higher-order complexes, as a 40 nt probe could have additional putative cis-elements. In addition, an extended analysis of the databases could be evaluated with other interacting TFs that might be functionally associated in processes such as metabolism and development. For example, the TFs that might be associated with OsWOX13 were sought in PlantTFDB through the elements in the 40 bp flanking

ATTGATTG in the promoter regions (data not shown). Thus, the CAATCA site for Os09g0528200 (homeobox-leucine zipper protein), AAAAAG site for Os02g0707200 (Dof-like protein 34) and CAAGNAA site for Os03g0119966 (NAC-domain protein) are frequently found elements in rice.

Conclusions

These results showed that RPBM is applicable in the analysis of DNA-binding motifs with a TF where binding is evaluated in extended natural promoter regions. The analysis can also be applicable to TFs that have single or multiple binding motifs. The technology might even be expanded for application to TFs that are heterodimers or form higher-order complexes. In addition, the extended analysis of the databases could be evaluated with other interacting TFs that might be functionally associated in processes such as metabolism and development.

Material And Methods

Protein Expression and Purification

All proteins used in this study were expressed as N-terminal fusions to a polyhistidine-tag and the DsRed fluorescent protein. The coding sequence of the DsRed fluorescent protein was amplified from the pDsRed monomer vector (Clontech) by PCR and inserted into the pET32a expression vector (Novagen). Full-length *OsWOX13* and *OsSMF1* were amplified from the cDNA of *O. sativa* and inserted into the pET32a-DsRed recombinant vector. These proteins were expressed in *Escherichia coli* strain BL21-CodonPlus (Stratagene). The overnight-cultured cells were inoculated in fresh liquid LB media, grown at 37 °C to an OD_{260} of 0.6 and induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at 25 °C for 5 h. Cell pellets were resuspended in 5 ml of phosphate-buffered saline (PBS) buffer including protease inhibitor and sonicated to lysis for 5 min with 45 second intervals on ice. Supernatant soluble fractions were retained after centrifugation at 4 °C for 20 min at 14,000 g. Proteins were enriched using Ni-NTA resins (Stratagene) according to the manufacturer's protocols. The purified protein fractions were collected in a volume of 500 µl, and concentrations were determined.

Synthesis of Complementary Strands on Microarray

Complementary DNA strands were synthesized as described in a previous report. The reaction solution contained 40 µM dNTPs (TaKaRa), 1.6 µM CyDye5-dUTP (GE Healthcare), 1 µM 5'-CTGCACTAGGTGACTCCG-3' primer (Bioneer), 1X ThermoSequenase buffer, and 0.5 U/µl ThermoSequenase (USB). A custom-designed PBM (Agilent) was combined with the reaction solution in a hybridization chamber (Agilent) according to the manufacturer's protocol. The assembled hybridization chamber was incubated at 85 °C for 10 min and then 60 °C for 90 min. The microarray was washed in PBS–0.01% (v/v) Triton X-100 at 37 °C for 1 min, PBS–0.01% (v/v) Triton X-100 at 37 °C for 10 min and PBS at room temperature for 3 min and dried by centrifugation at 500 g for 2 min. The doubled-stranded microarray was scanned to verify successful synthesis.

Protein Binding Microarray and Data Analysis

Double-stranded microarrays were washed with PBS containing 0.01% (v/v) Triton X-100 and blocked with PBS containing 2% (wt/v) BSA (Sigma) for 1 h. Then, the microarray was first washed with PBS containing 0.1% (v/v) Tween-20 and then with PBS containing 0.01% (v/v) Triton X-100 for 1 min. The protein binding mixture was prepared containing 200 nM TF in PBS containing 2% (wt/v) BSA, 51.3 ng/µl salmon testes DNA (Sigma), and 50 µM zinc acetate. The prepared protein mixture was incubated to stabilize and bind the microarray at 25 °C for 1 h. The microarray was first washed for 2 min with PBS containing 50 µM zinc acetate and 0.5% (v/v) Tween-20 for 10 min, then with PBS containing 50 µM zinc acetate and 0.01% Triton X-100 for 2 min, and finally with PBS containing 50 µM zinc acetate. Fluorescence images were obtained with a microarray scanner (Axon).

Selection of Promoters Containing ATTGATTG motifs

The 1 kb long promoter regions of 29,379 rice genes were retrieved from RAP-DB (http://rapdb.dna.affrc.go.jp/). The genes containing ATTGATTG were selected by using an in-house Perl script. A total of 1631 genes contained the motif in their promoters. Promoter regions 1 kb long were also retrieved from the same database. To identify cis-elements and TFs that might be associated with OsWOX13, the TFs and their associated cis-elements of *Oryza sativa* were downloaded from the Plant Transcription Factor Database (Jin et al. 2017). The representative cis-elements are extracted by using the nucleotides with higher occupancies than 0.5 at each position in the letter-probability matrix. The motifs with at least 6 distinctive nucleotides and nonconsecutive Ns were chosen for further analysis. With these criteria, 264 TFs and cis-elements were identified.

Electrophoretic Mobility Shift Assay (EMSA)

First, 5' FAM-end labeled and unlabeled oligonucleotides were annealed with each complimentary sequence. Five micrograms of OsWOX13 and OsSMF1 protein was incubated with 40 fmol of FAM-labeled double-stranded oligonucleotides, 1 µg of poly dI-dC, 1X binding buffer, 2.5% (v/v) glycerol and 0.05% (wt/v) NP-40 in a 20 µl reaction volume for 1 h at room temperature according to the manufacturer's instructions (Pierce). The reaction mixture was then analyzed by electrophoresis in a nondenaturing 6% acrylamide gel with 0.5X TBE buffer. The DNA-protein complexes in the gel were detected as fluorescence signals using Fusion SL (Vilber Lourmat).

Abbreviations

PBM
protein binding microarray,
TF
Transcription factors
RPBM

rice (*Oryza sativa*)-specific protein binding microarray
ChIP
chromatin immunoprecipitation
PCR
polymerase chain reaction

Declarations

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (JSK, Grant no. NRF-2018R1D1A1B07049288; YKK, Grant no. NRF-2018R1D1A1B07049348)

Author's contributions

JSK generated the data and wrote the paper. SC and KMJ performed the flanking DNA sequencing analysis. GSL observed the field phenotypes of the rice lines. KK and JSJ analyzed binding motifs in the databases. YKK inspired the overall work and revised the final manuscript. All authors read and approved the final manuscript.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable

References

1. Anderson JT, Rogers JM, Barrera LA, Bulyk ML (2020) Context and number of noncanonical repeat variable diresidues impede the design of TALE proteins with improved DNA targeting. Protein Sci 29:606–616

2. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837
3. Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc 4:393–411
4. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24:1429–1435
5. Hellman LM, Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat Protoc 2:1849–1861
6. Helwa R, Hoheisel JD (2010) Analysis of DNA-protein interactions: from nitrocellulose filter binding assays to microarray studies. Anal Bioanal Chem 398:2551–2561
7. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res 43:D117–D122
8. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 45:D1040–D1045
9. Kim JS, Chae S, Jun KM, Pahk Y-M, Lee T-H, Chung PJ, Kim Y-K, Nahm BH (2017) Genome-wide identification of grain filling genes regulated by the OsSMF1 transcription factor in rice. Rice 10:16
10. Kim M-J, Lee T-H, Pahk Y-M, Kim Y-H, Park H-M, Do Choi Y, Nahm BH, Kim Y-K (2009) Quadruple 9-mer-based protein binding microarray with DsRed fusion protein. BMC Mol Biol 10:91
11. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics 26:2438–2444
12. Minh-Thu P-T, Kim JS, Chae S, Jun KM, Lee G-S, Kim D-E, Cheong J-J, Song SI, Nahm BH, Kim Y-K (2018) A WUSCHEL homeobox transcription factor, OsWOX13, enhances drought tolerance and triggers early flowering in rice. Mol Cells 41:781
13. Perez-Rodriguez P, Riano-Pachon DM, Correa LG, Rensing SA, Kersten B, Mueller-Roeber B (2010) PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res 38:D822–D827
14. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, Wang S, Chen J, Shen L, Duan X, Hu S, Li W, Long H, Zhang Y, Liu XS (2012) CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. Bioinformatics 28:1411–1412
15. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. Science 290:2306–2309
16. Rogers JM, Barrera LA, Reyon D, Sander JD, Kellis M, Joung JK, Bulyk ML (2015) Context influences on TALE-DNA binding revealed by quantitative profiling. Nat Commun 6:7440

17. van Steensel B, Delrow J, Henikoff S (2001) Chromatin profiling using targeted DNA adenine methyltransferase. Nat Genet 27:304–308
18. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40:897–903
19. Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24:238–241
20. Yang JH, Li JH, Jiang S, Zhou H, Qu LH (2013) ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res 41:D177–D187
21. Yilmaz A, Nishiyama MY Jr, Fuentes BG, Souza GM, Janies D, Gray J, Grotewold E (2009) GRASSIUS: a platform for comparative regulatory genomics across the grasses. Plant Physiol 149:171–180
22. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res 19:556–566

Figures

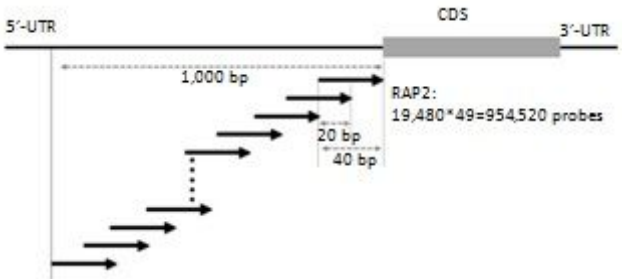

Figure 1

Schematic of the rice promoter protein binding microarray. A probe is 40 bp long, of which 20 bp overlaps. For each gene, 49 probes spanned the 1 kb promoter region before the translation start site. A total of 954,520 probes were designed from 19,480 genes among 31439 genes.

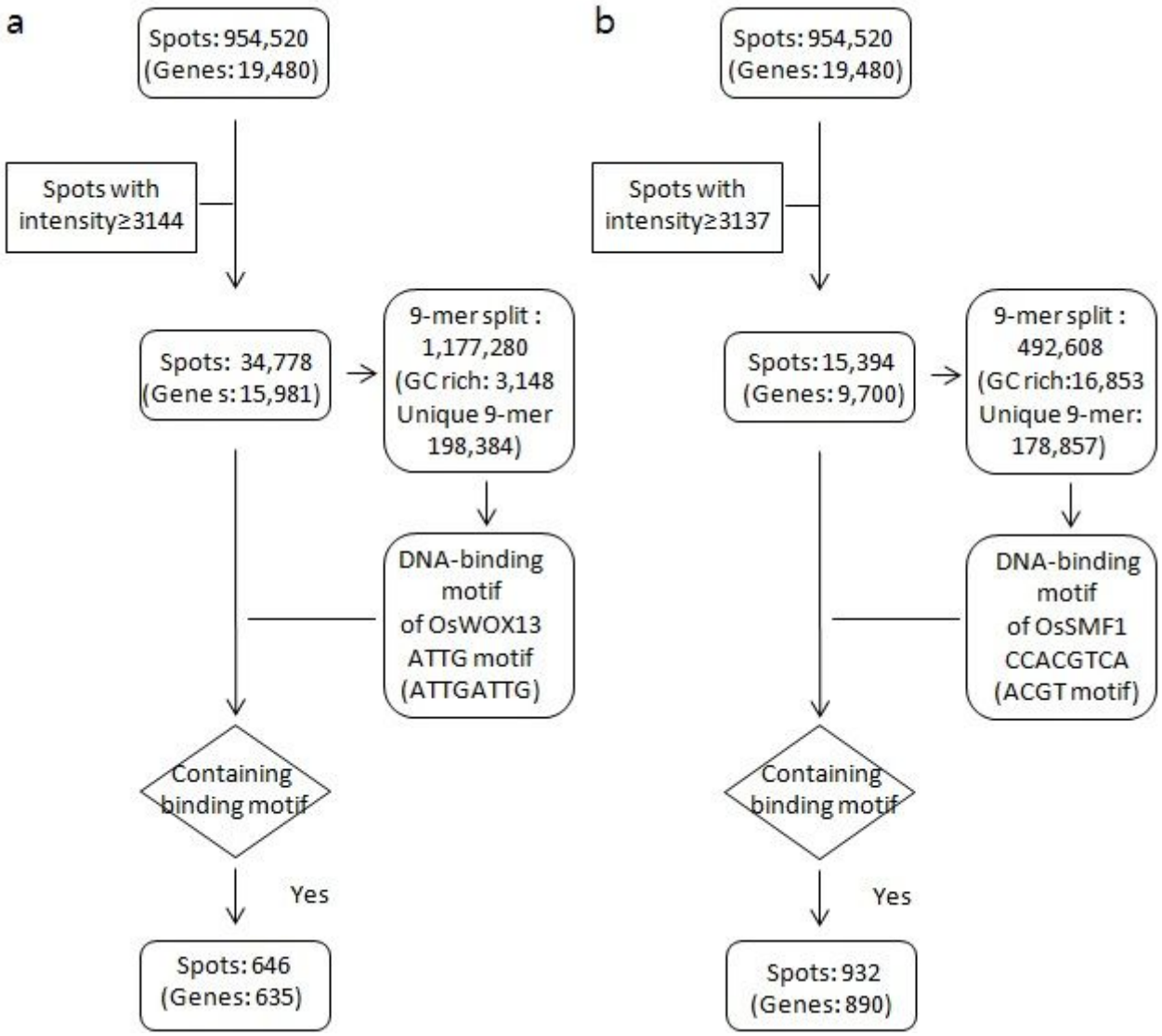

Figure 2

A flowchart diagram of the process for identifying putative target genes of transcription factors OsWOX13 (a) and OsSMF1 (b).

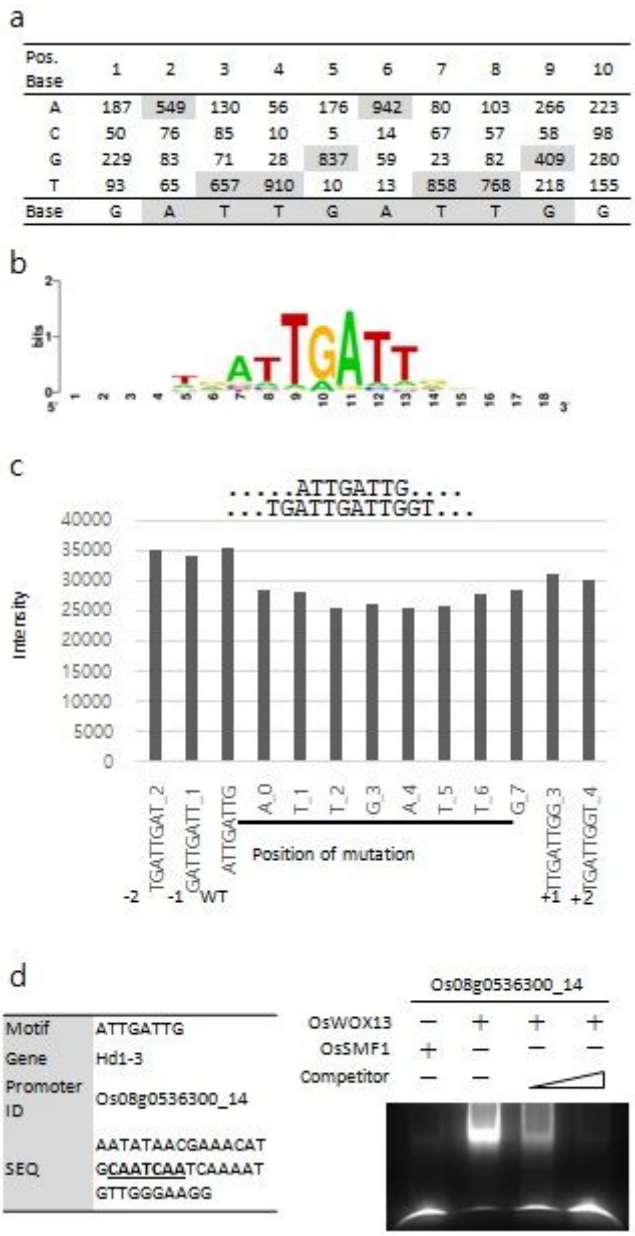

Figure 3

DNA-binding motif analysis of OsWOX13. a) DNA-binding motifs of OsSMF1 determined by clustering of the significant binding sequences. They were visualized with the Web logo program (weblogo.berkeley.edu). b) Position weight matrix from clustering of 9-mers. c) Comparison of the intensities of oligomers with point mutations at distinct positions in ATTGATTG. Binding motif of OsWOX13 from Wilcoxon-Mann-Whitney test, p-value 0. The wild type (WT) has the highest value (37,706), and the intensities of the 9-mer sequences with a point mutation were obtained from the list in Table S1. d) EMSA-based competition analysis of OsWOX13 using the probe Os08g0536300_14, which contains the ATTGATTG motif. The 40 bp sequences used as probes and their competitors are depicted. EMSAs were carried out using the OsWOX13:DsRed protein and a probe 5'-labeled with FAM. Competition for the labeled sequences was tested by adding different concentrations of unlabeled probes.

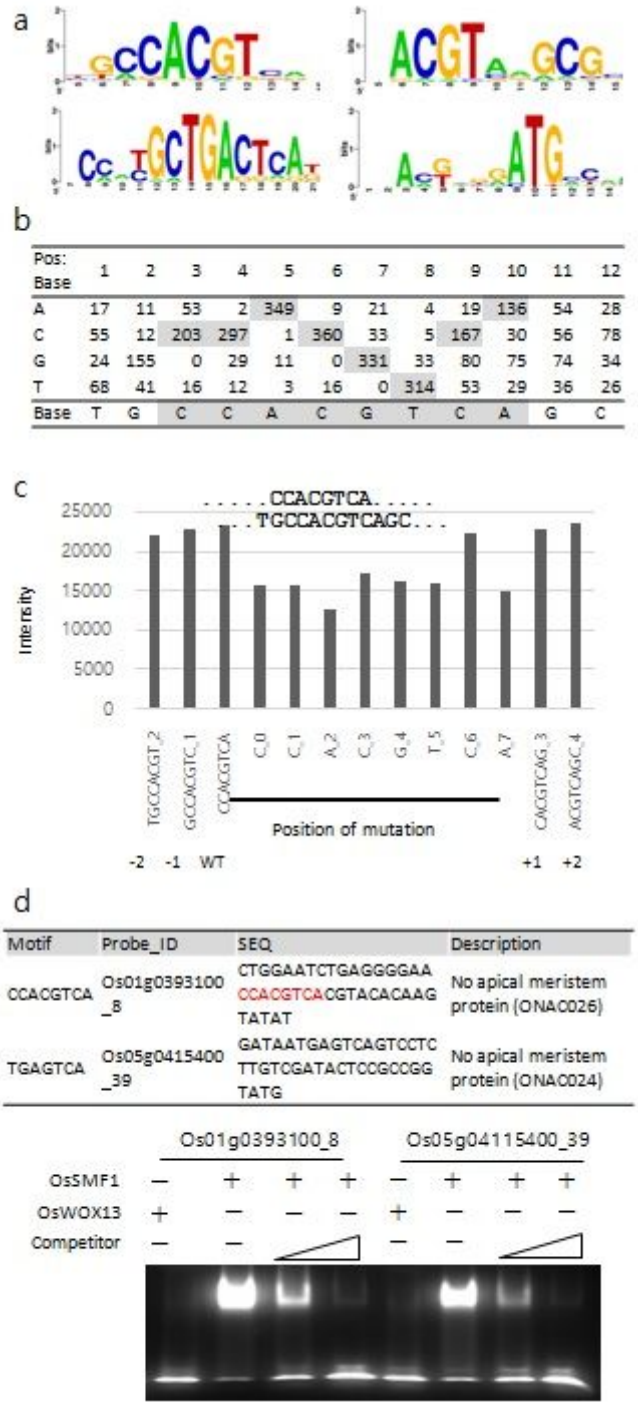

Figure 4

DNA-binding motif analysis of OsSMF1. a) DNA-binding motifs of OsSMF1 by clustering of the significant binding sequences. It gave at least 4 clusters; each cluster was analyzed, and its position weight matrix was calculated. The sequences were visualized with the Web logo program as shown in Figure 2. b) A cluster containing the GCCACGT motif and the position weight matrix. b) Comparison of the intensities of oligomers with point mutations at distinct positions in GCCACGTCA. Binding motif of OsSMF1 from the Wilcoxon-Mann-Whitney test, p-value 0. c) Mutation analysis using 9-mers. d) EMSA-based competition analysis of OsSMF1. Forty bp sequence feature probes, Os01g0393100_8 and Os05g0415400_39, representing GCCACGT and TGAGTCA clusters, respectively, were used as probes,

and competitors are depicted. EMSAs were carried out using the OsSMF1:DsRed protein and a probe 5'-labeled with FAM. Competition for the labeled sequences was tested by adding different concentrations of unlabeled probes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile4TableS3.xlsx
- Additionalfile2TableS1.xlsx
- Additionalfile7TableS6.xlsx
- Additionalfile8TableS7.xlsx
- Additionalfile6TableS5.xlsx
- Additionalfile9TableS8.xlsx
- Additionalfile3TableS2.xlsx
- Additionalfile1FigureS1S5.pptx
- Additionalfile5TableS4.xlsx
- Coverletter.docx