PCP
PLANT & CELL PHYSIOLOGY

Special Issue – Regular Paper

In Silico Identification of Short Nucleotide Sequences Associated with Gene Expression of Pollen Development in Rice

Motohiro Mihara[1], Takeshi Itoh[2] and Takeshi Izawa[1,*]

[1]Plant Genomics Research Unit, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, 305-8602 Japan
[2]Bioinformatics Research Unit, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, 305-8602 Japan

Microarray analysis of tiny amounts of RNA extracted from plant section samples prepared by laser microdissection (LM) can provide high-quality information on gene expression in specified plant cells at various stages of development. Having joined the LM-microarray analysis project, we utilized such genome-wide gene expression data from developing rice pollen cells to identify candidates for *cis*-regulatory elements for specific gene expression in these cells. We first found a few clusters of gene expression patterns based on the data from LM-microarrays. On one gene cluster in which the members were specifically expressed at the bicellular and mature pollen mitotic stages, we identified gene cluster fingerprints (GCFs), each of which consists of a short nucleotide representing the gene cluster. We expected that these GCFs would contain *cis*-regulatory elements for stage- and tissue-specific gene expression, and we further identified groups of GCFs with common core sequences. Some criteria, such as frequency of occurrence in the gene cluster in contrast to the total tested gene set, flanking sequence preference and distribution of combined GCF sets in the gene regions, allowed us to limit candidates for *cis*-regulatory sequences for specific gene expression in rice pollen cells to at least 20 sets of combined GCFs. This approach should provide a general purpose algorithm for identifying short nucleotides associated with specific gene expression.

Keywords: *Cis*-regulatory elements • Microarray • Pollen development • Rice • Transcription factor-binding site (TFBS) • Transcription start site (TSS).

Abbreviations: GCF, gene cluster fingerprint; EST, expressed sequence tag; GA20ox, gibberellin 20-oxidase; GA3βox gibberellin 3β-oxidase; GRSF, germline-restrictive silencing factor; LDSS, local distribution of short sequences; LM, laser microdissection; RAP, Rice Annotation Project; TFBS, transcription factor-binding site; TSS, transcription start site.

Introduction

In the course of efforts to elucidate the molecular mechanisms of biological events, extensive research has been conducted to increase our understanding of how specific gene expression is regulated. In addition to experimental analyses using several reporter genes in transgenic organisms and transiently expressed cells, many methods have been used to find *cis*-regulatory elements in a given gene group *in silico*. Multiple EM for Motif Elicitation (MEME; Bailey and Elkan 1994, Bailey et al. 2006), AlignACE (Hughes et al. 2000) and Motif-Sampler (Thijs et al. 2001, Thijs et al. 2002) are popular standards used in this field. These software packages utilize expectation maximization (EM) or the Gibbs sampling algorithm and can find consensus sequences occurring more frequently than at background levels from gene sequence information in a given gene group. However, these programs are basically statistical processing software that does not consider the genome composition related to *cis*-regulatory elements in the target organism or the local distribution of *cis*-regulatory elements. In contrast, positional information on candidates from transcription start sites (TSSs) is considered by some programs, such as the new version of A-GLAM (Defrance and Touzet 2006, Kim et al. 2006, Kim et al. 2008), because transcription-factor binding sites (TFBSs) are usually distributed significantly frequently in a range of upstream regions of TSSs (Hughes et al. 2000, Wray et al. 2003). However,

*Corresponding author: E-mail, tizawa@nias.affrc.go.jp

Plant Cell Physiol. 49(10): 1451–1464 (2008) doi:10.1093/pcp/pcn129, available online at www.pcp.oxfordjournals.org

it is not yet easy to find consistent rules for the positional preferences of TFBSs (Wray et al. 2003). Genome composition is also considered in a new algorithm called 'gene enrichment motif searching' (GEMS) for finding *cis*-regulatory elements in a non-model organism, *Plasmodium falciparum* (Young et al. 2008). In plants, a position-sensitive extraction method called 'local distribution of short sequences' (LDSS) has been applied to rice (*Oryza sativa*) and *Arabidopsis thaliana* genomes and has identified promoter constituents categorized into major three groups, REG, TATA box and Y Patch (Yamamoto et al. 2007). These data were recently summarized in a database, termed ppdb (Yamamoto and Obokata 2008). Some databases, such as PLACE, compile information on *cis*-regulatory elements found in the literature (Higo et al. 1999).

The extraction of microarray data from RNA samples prepared by laser microdissection (LM) (Nakazono et al. 2003) can provide ideal noise-free gene expression data in which a group of genes—a gene-cluster—is expressed in specified cells at a certain developmental stage, and not expressed in other type of cells. In joining the project on the LM-microarray analysis of pollen development in rice, as described in the accompanying papers (Hirano et al. 2008, Hobo et al. 2008, Suwabe et al. 2008), we decided to use the LM-microarray data on rice pollen development to find candidates for *cis*-regulatory elements by an ad hoc method in order to develop an algorithm for a future systematic *in silico* search for *cis*-regulatory elements. Here, we developed a new method of finding candidates for *cis*-regulatory elements, termed the combined GCF (gene cluster fingerprint) method. The method consists of two steps: (i) identifying GCFs from promoter sequence information, which are a catalog of representative short nucleotide sequences that may determine the specificity of a given gene cluster; and (ii) searching common core sequences in the GCFs and picking up combined GCF sets, each of which may represent a TFBS. These candidates for *cis*-regulatory elements were evaluated in paralogous gene groups clustered in an annotation-similarity-based genome comparison database, termed SALAD (Surveyed Alignment and Associating Dendrogram; http://salad.dna.affrc.go.jp/salad/), which we recently started. All the microarray data described here are freely available to the public through the SALAD web site as presented in genome-wide annotation similarity clusters (http://salad.dna.affrc.go.jp/CGViewer/MicroArrayPollen/).

Results

Identification of gene clusters from LM-microarray data

From a total of 31 samples of microarray data obtained by using rice 44K oligo microarrays (Agilent Technologies, Palo Alto, CA, USA), we removed data with outliers (see Materials and Methods) and selected expression data with 5,110 probes for clustering on the basis of specific gene expression. With the Z-score calculated from the selected raw data, pairwise correlation coefficients between all possible combinations of probes were obtained and were considered as distances between the gene probes. This pairwise distance was utilized to find clusters in a two-dimensional way by the average distance method using Cluster software (Eisen et al. 1998). The results showed three major gene clusters of specific gene expression: (A) a gene cluster for the late bicellular (BC in **Fig. 1**) and mature pollen (tricellular; TC in **Fig. 1**) stages; (B) a gene cluster for tapetum-specific gene expression; and (C) a gene cluster for the microspore (UN in **Fig. 1**) and early bicellular pollen stages (BC in **Fig. 1**) (Singh and Bhalla 2007). These gene clusters consisted of 186 probes, 312 probes and 325 probes, respectively (**Fig. 1**). We arranged the probe data on the basis of identifier, and obtained 128 genes, 199 genes and 233 genes in each respective gene cluster. Note that the 5,110 probes corresponded to 3,795 genes of rice. In light of the specificity of gene expression shown in **Fig. 1**, in this work, we focused on the use of the first gene cluster, the '128-gene cluster', which consists of 128 genes specifically expressed at the late bicellular and mature pollen stages (Supplementary Table S1).

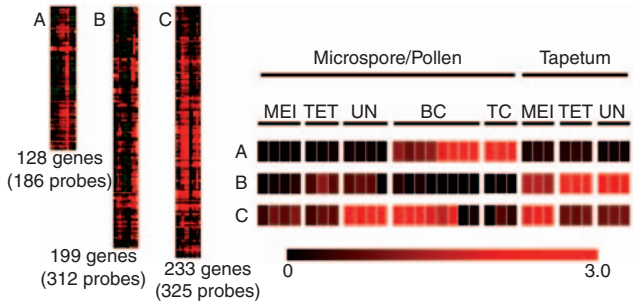

Fig. 1 Clustering of genes using 31 profiles of microarray expression in the pollen development stage of rice. Data are presented in a matrix format (left): each row represents an individual gene, and each column corresponds to 31 microarray samples at various pollen developmental stages. Normalized gene expression is depicted by a pseudocolor scale, with red indicating high gene expression and black indicating low expression (Z-score). Average expression pattern of each cluster (right). The results showed three major gene clusters (A, B and C) of specific gene expression. These gene clusters consisted of 128 genes, 199 genes and 233 genes, respectively

Identification of fingerprints of gene clusters at the late bicellular and mature pollen stages

We introduced the GCF as a new concept (**Fig. 2**). A GCF is any short nucleotide which can contribute to the specificity of a given gene cluster and is found at significantly higher rates in the promoter regions of members of the gene cluster than in

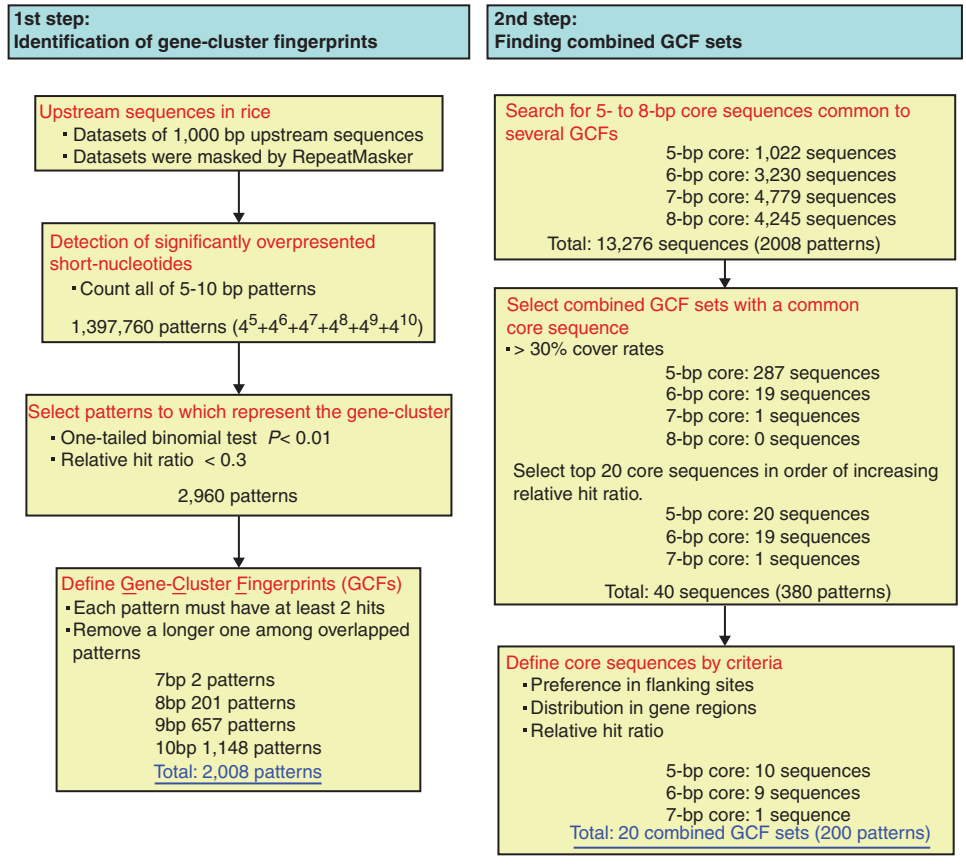

Fig. 2 Flowchart of the two-step analysis of the combined GCF method: how to define GCFs and how to find core sequences of combined GCFs. In this study we analyzed the 128-gene cluster specifically expressed at the late bicellular and mature pollen stages (see Fig. 1).

those of all the genes tested. Thus, GCFs can be used to represent the entire gene cluster with reduced complexity. Some of these GCFs can be expected to be the *cis*-regulatory elements that confer stage- and tissue- or other specific gene expression.

First, we collected information on the 1,000 bp promoter sequences upstream of the TSSs registered in the Rice Annotation Project Database (RAP-DB; http://rapdb.dna.affrc.go.jp) for all 28,237 rice genes (with some cDNA evidence); these sequences were expected to contain major *cis*-regulatory elements for the expression of specific genes at the developmental stage in addition to promoter constituents. After removal of the repeated DNA sequences from these 1,000 bp promoter regions by RepeatMasker with TIGR's Oryza Repeat data (v. 3.1) to reduce the sequence complexity, we then counted all possible 5, 6, 7, 8, 9 and 10 bp nucleotide sequences (in total, 1,397,760 sequence patterns; **Fig. 2**) in the 1,000 bp promoter regions of the 128-gene cluster and of all of the rice genes without outlier data (3,795 genes) (**Fig. 3**). If short nucleotide sequences were distributed randomly throughout the upstream regions, then the frequency of occurrence of the sequence should obey a binomial distribution. Therefore, we examined the statistical significance of the frequency of occurrence of each nucleotide sequence in the 128 genes on the basis of the binomial distribution, considering the frequency of occurrence of the same pattern in the promoter regions of all the 3,795 genes as the probability of binomial distribution. We further calculated the relative hit ratios (hit rate for the 3,795 genes/hit rate for the 128-gene cluster; see Materials and Methods) of each nucleotide sequence between the 128-gene cluster and the tested 3,795 genes. Here, a smaller relative hit ratio means a higher specificity of each sequence in the genes expressed at the late bicellular and mature pollen stages. Both the calculated significance of frequency and the relative hit ratio were utilized to determine the GCFs of the 128 genes. In this work, we defined a GCF as occurring with a frequency significance of $P < 0.01$ and a relative hit ratio of <0.3 (**Fig. 4**). At this point, 2,960 short sequence patterns met the criteria (**Fig. 2**). To simplify the GCFs as much as possible, short sequence patterns that were included only once in the selected short sequences were removed. In addition, if a short nucleotide

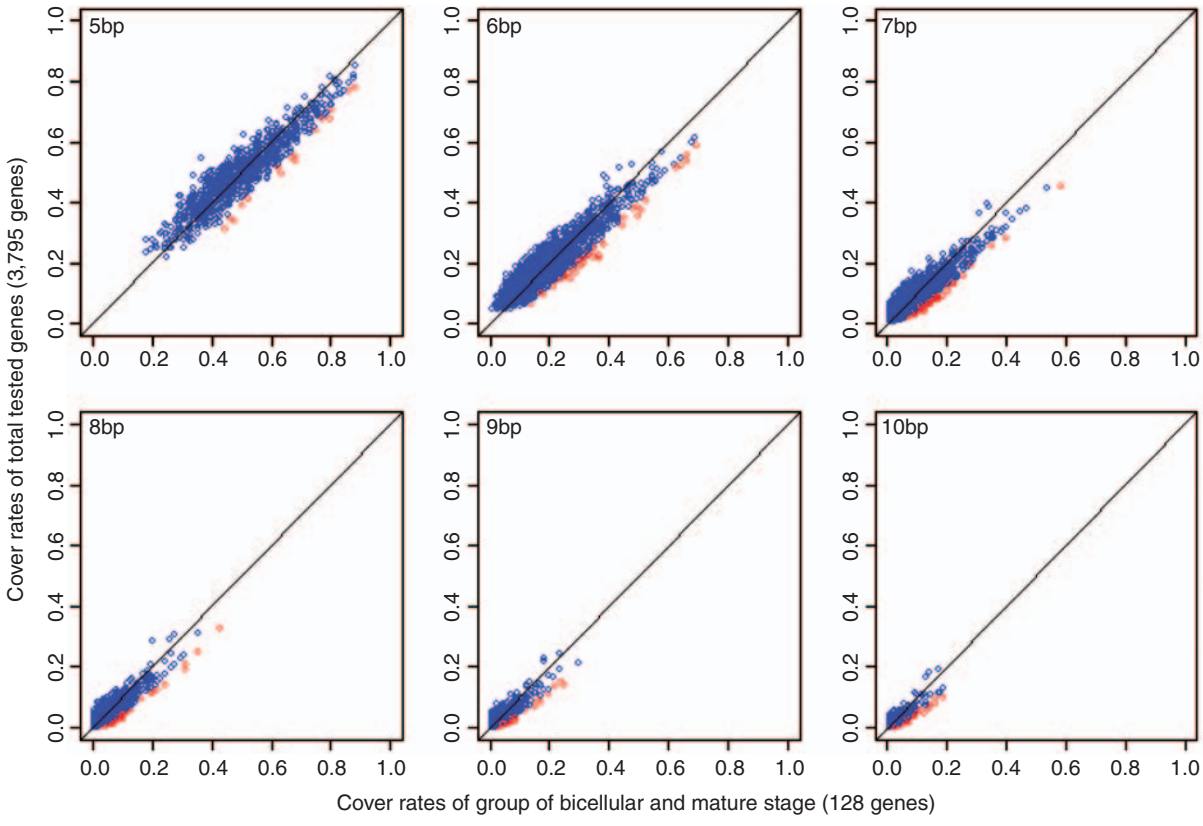

Fig. 3 Comparison of cover rates of short nucleotides in 1,000 bp promoter regions of the cluster of 128 genes expressed at the bicellular and mature pollen stages (*x*-axis) and of all the 3,975 genes without outliers (*y*-axis). Red circles indicate $P < 0.01$ by one-tailed binomial test.

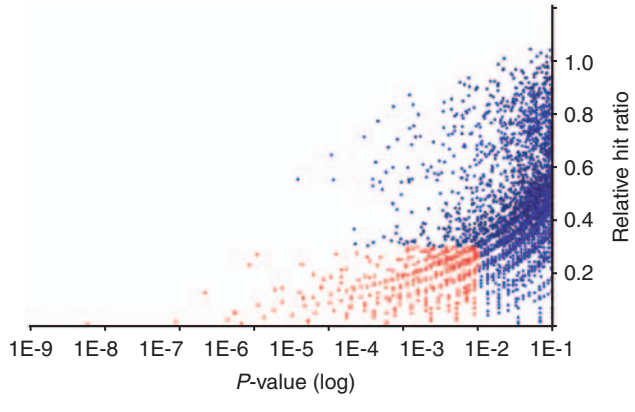

Fig. 4 Scatter plot with *P*-values and relative hit ratios of short nucleotides. Red dots indicate short nucleotides that have *P*-values <0.01 and relative hit ratios of <0.3. The relative hit ratio of a short nucleotide is equal to (hit rate for the total genes of 3975/hit rate for the gene cluster). The hit rate is equal to (numbers of occurrence of a short nucleotide/numbers of subjected genes).

pattern overlapping with a shorter one was included in the GCFs, we kept only the shorter one as a GCF to reduce the complexity of the GCFs. Finally, 2008 short nucleotide patterns remained (**Fig. 2** and Supplementary Table S2). We then defined these 2,008 nucleotide sequence patterns as GCFs for the 128-gene cluster, in which the members were expressed specifically at the late bicellular and mature pollen stages.

Search for GCFs with core sequences

Because of the low cover rates of most GCFs in the promoter regions of the 128 genes, most of the GCFs alone could not be considered as *cis*-regulatory elements conferring specific gene expression. Therefore, we next tried to find core sequences common to some groups of GCFs. We examined all possible 5, 6, 7 and 8 bp nucleotide sequences in the 2,008 GCFs and found 1,022 common 5 bp sequences (among a total of 1,024 sequences of 5 bp), 3,230 common 6 bp sequences (among 4,096 sequences), 4,779 common 7 bp sequences (among 16,384 sequences) and 4,245 common 8 bp sequences (among 65,536 sequences) (**Fig. 2**). We considered

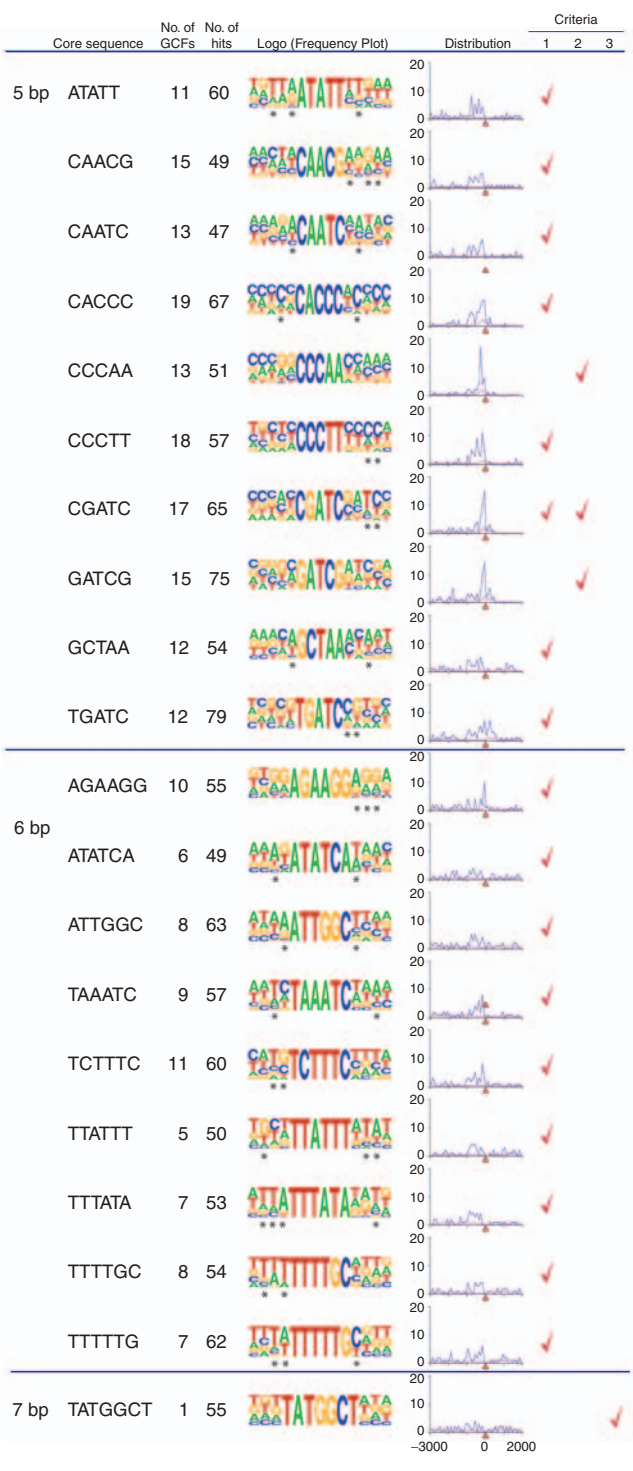

Fig. 5 Information of the combined GCFs. After selection, 20 combined GCF sets were listed. The third column shows the number of GCFs in a combined GCF set. The fourth column shows the number of GCFs belonging to a combined GCF set found in the 128 genes; all those GCFs were used to create the logos in the fifth column. Sites in the flanking sites with significant preferences (see Materials and Methods) are shown (*). The sixth column shows the positional distribution pattern of each of the 20 combined GCFs in the 5,000 bp gene region. The *x*-axis indicates the position from the TSS, and the *y*-axis indicates the occurrence rate (/128 genes) of each pattern. Blue lines show the patterns in the 128 genes, while red lines indicate the corresponding patterns for the total 3,795 tested genes. Three criteria were used to pick up the candidate combined GCFs. (i) Preference in the flanking regions. More than 0.5 appearance of any single nucleotide at least two flanking sites (asterisks). (ii) Positional distribution. Combined GCFs with acute positional peaks were selected. Note that all GCFs were over-represented in the promoter regions. (iii) Relative hit ratio of the core sequence of combined GCF <0.3. Although TTTTGC and TTTTTG 6 bp core sequences may be related, we missed it as a 7 bp core sequence due to the <0.3 cover rate. Note that the second best 7 bp core sequence was TTTTTGC, which was not selected on the GCF selection by the criteria in this work (see Fig. 2).

these sequences to be core sequences and categorized the 2,008 GCFs into subsets with common core sequences. We then examined the cover ratios (numbers of genes containing subset GCF members/128 genes) of those subsets of GCFs in the promoter region (1,000 bp from TSS) of the 128-gene cluster. The subset of GCFs could be combined because of their common possession of a particular core sequence and represented by the core sequences. Many TFBSs for a certain transcription factor often included a rigid core sequence and flexible flanking sequences. Therefore, if we could find a subset of GCFs that had a certain core sequence and exhibited a high cover ratio among the 128 genes, then we could consider these GCFs as candidates for *cis*-regulatory elements conferring specific gene expression at the late bicellular and mature pollen stages. First, we empirically selected cover ratios of >30%; this limited the numbers of candidate 5, 6, 7 and 8 bp core sequences to 287, 19, 1 and 0 sequences, respectively (**Fig. 2**, Supplementary Table S3). GCFs with the top 20 of the 5 bp core sequences in the order of cover ratios for the 128-gene cluster, GCFs with all the nineteen 6 bp core sequences and GCFs with the one 7 bp core sequence were further examined to pick up more likely candidates for *cis*-regulatory elements. We used three more criteria here: (i) the preferential sequences flanking the core sequences; (ii) local distribution in a 5,000 bp gene region (3,000 bp upstream and 2,000 bp downstream of the TSS), because TFBSs are usually distributed significantly frequently in a range of upstream regions of TSS (Hughes et al. 2000, Wray et al. 2003); and (iii) the cover ratio of core sequences. Finally, 20 core sequences (ten 5 bp, nine 6 bp and one 7 bp) found in 200 GCFs remained (**Figs. 5, 6**). We therefore concluded that these 200 GCFs (belonging to 20 combined GCF sets) were candidates for *cis*-regulatory elements conferring specific gene expression at the late bicellular and mature pollen stages (**Fig. 5**). Each combined GCF set contained 11–19 GCFs (for the 5 bp core), 5–11 GCFs (for the 6 bp core) and one GCF (for the 7 bp core) (**Fig. 6**).

			Group of bicellular and mature stage (128 genes)				Total tested genes (3,795 genes)				Combined GCFs	Combined GCFs	Core sequence
			5,000 bp				5,000 bp				5,000 bp	1,000 bp	1,000 bp
Length	Core pattern	No. of GCF	No. of genes	Cover rate	No. of hits	Hit rate	No. of genes	Cover rate	No. of hits	Hit rate	Relative hit ratio	Relative hit ratio	Relative hit ratio
5 bp	ATATT	11	46	0.36	60	0.47	566	0.15	631	0.17	0.35	0.14	0.95
	CAACG	15	41	0.32	49	0.38	458	0.12	504	0.13	0.35	0.12	0.90
	CAATC	13	41	0.32	47	0.37	457	0.12	502	0.13	0.36	0.12	1.05
	CACCC	19	47	0.37	67	0.52	619	0.16	695	0.18	0.35	0.18	0.79
	CCCAA	13	40	0.31	51	0.40	459	0.12	527	0.14	0.35	0.19	0.97
	CCCTT	18	43	0.34	57	0.45	502	0.13	543	0.14	0.32	0.12	0.85
	CGATC	17	46	0.36	65	0.51	590	0.16	689	0.18	0.36	0.19	0.72
	GATCG	15	51	0.40	75	0.59	679	0.18	804	0.21	0.36	0.19	0.78
	GCTAA	12	44	0.34	54	0.42	528	0.14	586	0.15	0.37	0.14	0.96
	TGATC	12	59	0.46	79	0.62	619	0.16	707	0.19	0.30	0.12	0.90
6 bp	AGAAGG	10	45	0.35	55	0.43	780	0.21	956	0.25	0.59	0.23	0.74
	ATATCA	6	41	0.32	49	0.38	702	0.18	804	0.21	0.55	0.23	0.84
	ATTGGC	8	53	0.41	63	0.49	897	0.24	1033	0.27	0.55	0.23	0.57
	TAAATC	9	47	0.37	57	0.45	849	0.22	999	0.26	0.59	0.25	0.66
	TCTTTC	11	47	0.37	60	0.47	665	0.18	771	0.20	0.43	0.16	0.67
	TTATTT	5	43	0.34	50	0.39	621	0.16	694	0.18	0.47	0.20	0.89
	TTTATA	7	48	0.38	53	0.41	665	0.18	752	0.20	0.48	0.20	0.73
	TTTTGC	8	46	0.36	54	0.42	759	0.20	866	0.23	0.54	0.20	0.88
	TTTTTG	7	49	0.38	62	0.48	889	0.23	1036	0.27	0.56	0.23	0.98
7 bp	TATGGCT	1	50	0.39	55	0.43	860	0.23	978	0.26	0.60	0.30	0.30

Fig. 6 Comparison of occurrence of combined GCFs for each core between the 128-gene cluster and total tested genes (3,795 genes). We calculated the relative hit ratio (hit rate for the 3,975 total genes/hit rate for the 128-gene cluster) of both combined GCFs and the core sequence. The 7 bp pattern TATGGCT contains only one GCF, which means that the same 7 bp pattern is the core. Therefore, the relative hit ratio is equal between combined GCFs and core sequence in this case. 1,000 bp and 5,000 bp indicate the ranges of searched sequences. 1,000 bp, 1,000 bp upstream sequences of TSS; 5,000 bp, 3,000 bp upstream of TSS and 2,000 bp downstream of TSS.

MEME motif	Representative motif sequence	No. of sites	E-value	This analysis
1	CTCCTCCTCC	90	1.70E−72	-
2	AAAAAAAAAA	93	1.70E−58	-
3	CGGCGGCGGC	66	2.40E−38	-
4	TTTTTTTTTT	102	5.50E−34	TTATTT
5	GAGAGAGAGA	50	4.40E−30	
6	TCCCCTCCCC	84	1.60E−26	CCCTT
7	GGAAGAGGAG	49	5.90E−06	-
8	CTCTCTCTCT	44	2.80E−05	-
9	GGGCCCCACC	27	7.90E−05	-
10	CGCCGCCGCG	34	1.20E−03	-
11	GCCCGGCCCA	14	8.60E−02	-
12	GCGGCGGCG	27	1.30E+00	-
13	CCGCGCGCGC	27	1.90E−01	-
14	CCCACCCAC	30	4.10E+00	CACCC
15	AAAAGAAAA	30	5.70E+00	-
16	CGATCGATCC	18	1.50E+03	CGATC, GATCG
17	CGGCCCAA	31	2.10E+04	CCCAA
18	CCACCACACC	22	2.30E+04	-
19	CGCCGGCGTG	4	1.70E+05	-
20	GTGGTGGGGG	16	2.40E+04	-

Fig. 7 Top 20 results of MEME analysis for the 128-gene cluster. The order is based on E-values. Six core sequences of combined GCF sets were detected in the MEME results.

In silico evaluation of GCFs with core sequences as candidates for cis-regulatory elements

We compared the 20 combined GCF sets with the results of an MEME analysis (**Figs. 7, 8**) and a ppdb search (**Fig. 9**), indicating higher specificity of candidate selection by our analysis compared with the MEME analysis (compare **Figs. 6** and **8**) and greater sensitivity of detection compared with the ppdb search. The top 10 OsREGs sorted by occurrence rates by the ppdb search were often found with a similar core ppdb motif, GCCCA (**Fig. 9**). Paralogous gene groups, including members of the gene cluster expressed in the late bicellular and mature pollen stages, were further examined. In our SALAD database, clusterings and dendrograms obtained from pairwise scores based on similarity of conserved amino acid sequence motifs are provided for all genome annotation data for rice, *Arabidopsis thaliana* and a red sea alga. We integrated the LM-microarray data into our dendrogram in the SALAD database. This allowed us to compare stage-specific gene expression in paralogous gene groups (http://salad.dna.affrc.go.jp/CGViewer/MicroArrayPollen/) (**Fig. 10**). Because gene functions in paralogous groups are likely to be related, this site should provide users with useful information from the LM-microarrays. From this site, we selected some sets of microarray data in nine paralogous groups containing members of the 128-gene cluster (**Fig. 10** and Supplementary Figs. S1–S7). We then examined the gene regions of selected paralogous groups to look for the 200 GCFs (**Fig. 11**). The results showed that the 200 GCFs were present at greater frequencies in the 128-gene cluster members than in other members of the paralogous gene groups. These data strongly suggested that our algorithms can be used efficiently to

MEME Motif	Representative motif sequence	Group of bicellular and mature stage (128 genes) 1,000 bp No. of genes	Cover rate	No. of hits	Hit rate	Total tested genes (3,795 genes) 1,000 bp No. of genes	Cover rate	No. of hits	Hit rate	3,795 genes/128 genes 1,000 bp Relative hit ratio
1	CTCCTCCTCC	92	0.72	207	1.62	2421	0.64	6301	1.66	1.03
2	AAAAAAAAAA	83	0.65	140	1.09	1890	0.50	3106	0.82	0.75
3	CGGCGGCGGC	74	0.58	153	1.20	2254	0.59	5455	1.44	1.20
4	TTTTTTTTTT	90	0.70	152	1.19	2190	0.58	3678	0.97	0.82
5	GAGAGAGAGA	57	0.45	99	0.77	1340	0.35	2195	0.58	0.75
6	TCCCCTCCCC	100	0.78	208	1.63	2398	0.63	5392	1.42	0.87
7	GGAAGAGGAG	63	0.49	99	0.77	1482	0.39	2433	0.64	0.83
8	CTCTCTCTCT	70	0.55	113	0.88	1704	0.45	2646	0.70	0.79
9	GGGCCCCACC	41	0.32	61	0.48	868	0.23	1143	0.30	0.63
10	CGCCGCCGCG	67	0.52	120	0.94	2010	0.53	4332	1.14	1.22
11	GCCCGGCCCA	37	0.29	51	0.40	986	0.26	1290	0.34	0.85
12	GCGGCGGCG	58	0.45	102	0.80	1418	0.37	2815	0.74	0.93
13	CCGCGCGCGC	47	0.37	64	0.50	1125	0.30	1632	0.43	0.86
14	CCCACCCAC	60	0.47	97	0.76	1458	0.38	2276	0.60	0.79
15	AAAAGAAAA	75	0.59	127	0.99	1822	0.48	2940	0.77	0.78
16	CGATCGATCC	30	0.23	37	0.29	685	0.18	809	0.21	0.74
17	CGGCCCAA	51	0.40	74	0.58	970	0.26	1364	0.36	0.62
18	CCACCACACC	71	0.55	127	0.99	1949	0.51	3617	0.95	0.96
19	CGCCGGCGTG	39	0.30	57	0.45	1328	0.35	2204	0.58	1.30
20	GTGGTGGGGG	27	0.21	33	0.26	768	0.20	1004	0.26	1.03

Fig. 8 Comparison of occurrence of the MEME motifs between the 128-gene cluster and total tested genes. Data are shown as in Fig. 6. High relative hit ratios (the right column) with motif sequences selected by MEME indicate high hit rates (numbers of occurrence/numbers of tested genes) of the motifs in all the tested genes.

select candidate sequences for *cis*-regulatory elements. Finally, we counted the numbers of hits with the selected 200 GCFs for the 128-gene cluster and the background genes tested [3,667 (3,795 – 128) genes] (**Fig. 12**). As expected, rates of genes with hit numbers >2 were higher for the members among the 128-gene cluster than those observed in the background genes (3,667 genes). When we searched for genes with >5 combined GCFs in the promoter regions (1,000 bp) among the entire genes of rice (28,237 genes), we found 145 genes and 58 genes in the 28,237 genes and in the 128-gene cluster, respectively. Regarding non-members of the 128-gene cluster [87 (145 – 58) genes], 11 genes were specifically expressed at late bicellular and mature pollen stages in the LM-microarray data set (data not shown). In contrast, among 87 randomly selected genes, only three genes were expressed at late bicellular and mature pollen stages in the LM-microarray data set. Therefore, it is likely that we could pick up genes which are expressed at late bicellular and mature pollen stages efficiently only from this combined GCF search. On the other hand, we also noticed that there were some genes without any 20 combined GCF sets (which consists of 200 GCFs) in their promoter regions, although they were specifically expressed at the given stage.

ID	Sequence	ppdb motif	Group of bicellular and mature stage (128 genes) No. of genes
OsREG631	GGCCCCAC		16
OsREG542	CCGGCCCA	GCCCA	14
OsREG626	GGCCCAAC	GCCCA , CCAACGG	13
OsREG471	AGGCCCAA	GCCCA	13
OsREG445	ACGGCCCA	GCCCA	12
OsREG527	CCCACCAC	GCCCA	12
OsREG563	CGGCCCAA	GCCCA	10
OsREG587	GAGGCCCA	GCCCA	10
OsREG421	AAAGCCCA	GCCCA	10
OsREG612	GCCCCACC		10

Fig. 9 Top 10 list of REG octamers found in the 128-gene cluster. The order is based on the number of occurrences of OsREGs in the 128 gene regions. Two ppdb motifs were found in these OsREGs. GCCCA* (ppdb motif) is found in eight OsREGs (out of 10 OsREGs) and is known as 'Element II of Arabidopsis PCNA-2, cell cycle/meristematic expression'.

Discussion

Gene expression in pollen development

The impacts of LM-microarray analysis reported in the accompanying papers (Hirano et al. 2008, Hobo et al. 2008,

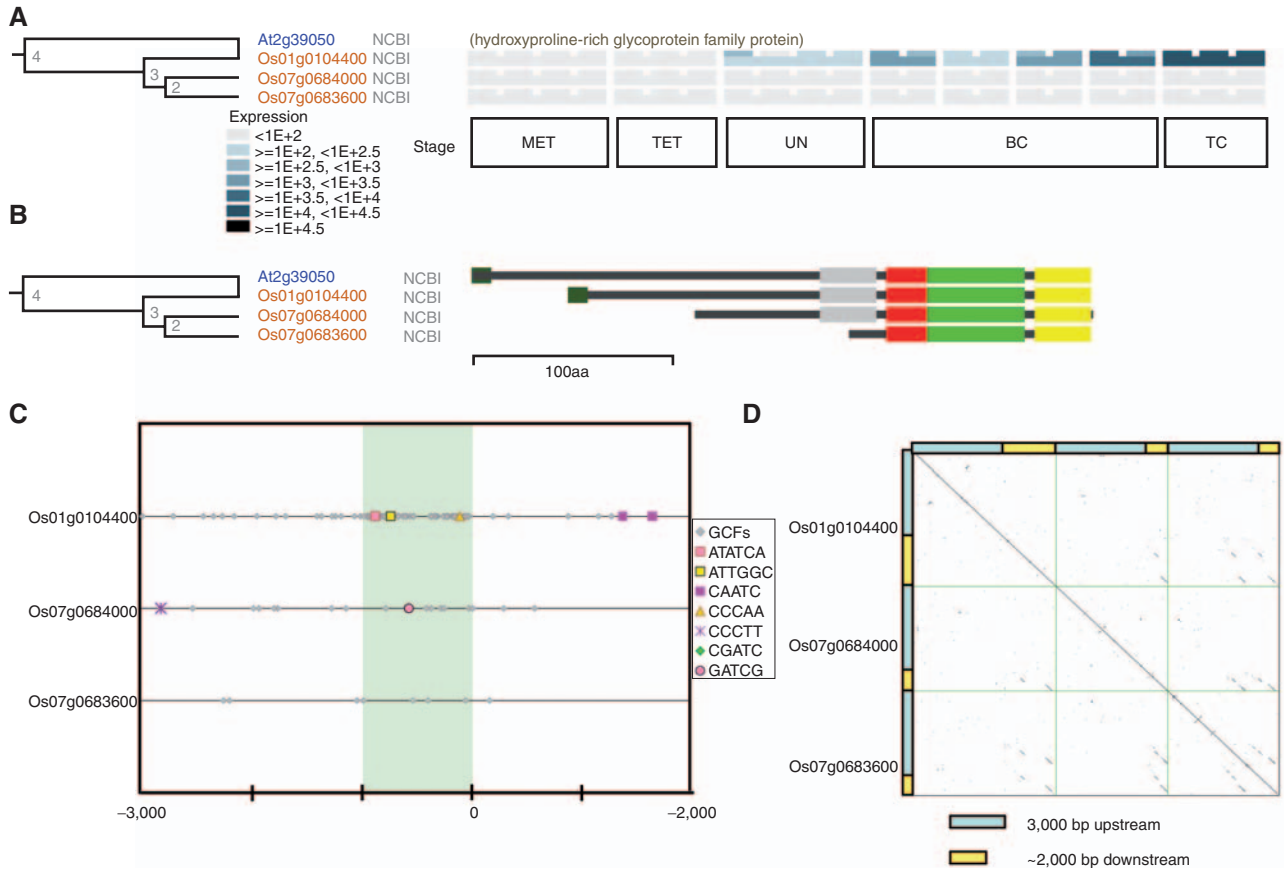

Fig. 10 Expression profiles in paralogous gene groups. (A) Gene expression profiles of a paralogous gene group clustered in the SALAD database (http://salad.dna.affrc.go.jp/salad/). (B) SALAD clustering with the diagram based on MEME-extracted motifs. (C) Distribution of combined GCFs in the paralogous genes. (D) Dot matrix between paralogous genes. In this case, no clear conservation was observed in the promoter region. Junctions of light blue and yellow indicate the TSS.

Suwabe et al. 2008) strongly motivated us to try *in silico* identification of *cis*-regulatory elements for specific gene expression in pollen development of rice. There have been several genome-wide analyses of pollen development in *A. thaliana*, maize, lily and rice (Engel et al. 2003, Endo et al., 2004, Pina et al. 2005, Okada et al. 2006). Although microarray analysis has been performed in *A. thaliana*, the developmental stages of pollen cells—such as microspore, early bicellular stage, late bicellular stage and mature pollen stage—were not considered in this work (Pina et al. 2005). In maize, >5,000 expressed sequence tags (ESTs) have been obtained from sperm cells isolated by fluorescence-activated cell sorting (Engel et al. 2003). In lily, 886 ESTs have been obtained from generative cells (Okada et al. 2006). However, the sequence information of promoter regions of the genes corresponding to these ESTs has not yet been obtained in maize or lily. Therefore, these data were not available to search for *cis*-regulatory elements. Some genes specific to pollen cells, such as genes involved in nucleotide excision repair, cell cycle progression and ubiquitin-related pathways, are expressed in pollen cells, but the developmental stages at which this expression occurs have not been identified (Singh and Bhalla 2007). In *Oryza sativa*, a 4K cDNA array was utilized to analyze gene expression in anthers, resulting in identification of more than a hundred genes specifically expressed in rice anthers, although there are many rice genes remaining to be analyzed (Endo et al. 2004). Thus, clearly the LM-microarray data in this project have provided a tremendous amount of information on gene expression in pollen development in plants (Suwabe et al. 2008). Our clustering analysis of the LM-microarray data revealed several gene clusters (**Fig. 1**). The 128 members of the gene cluster that is the main subject of this work—one of the identified gene clusters—were expressed only in the late bicellular stage and mature pollen stage of rice pollen development. Recently, a repressor termed GRSF (germline-restrictive silencing factor) was found to repress the activation of germline-specific genes in vegetative cells but not in sperm cells during pollen development

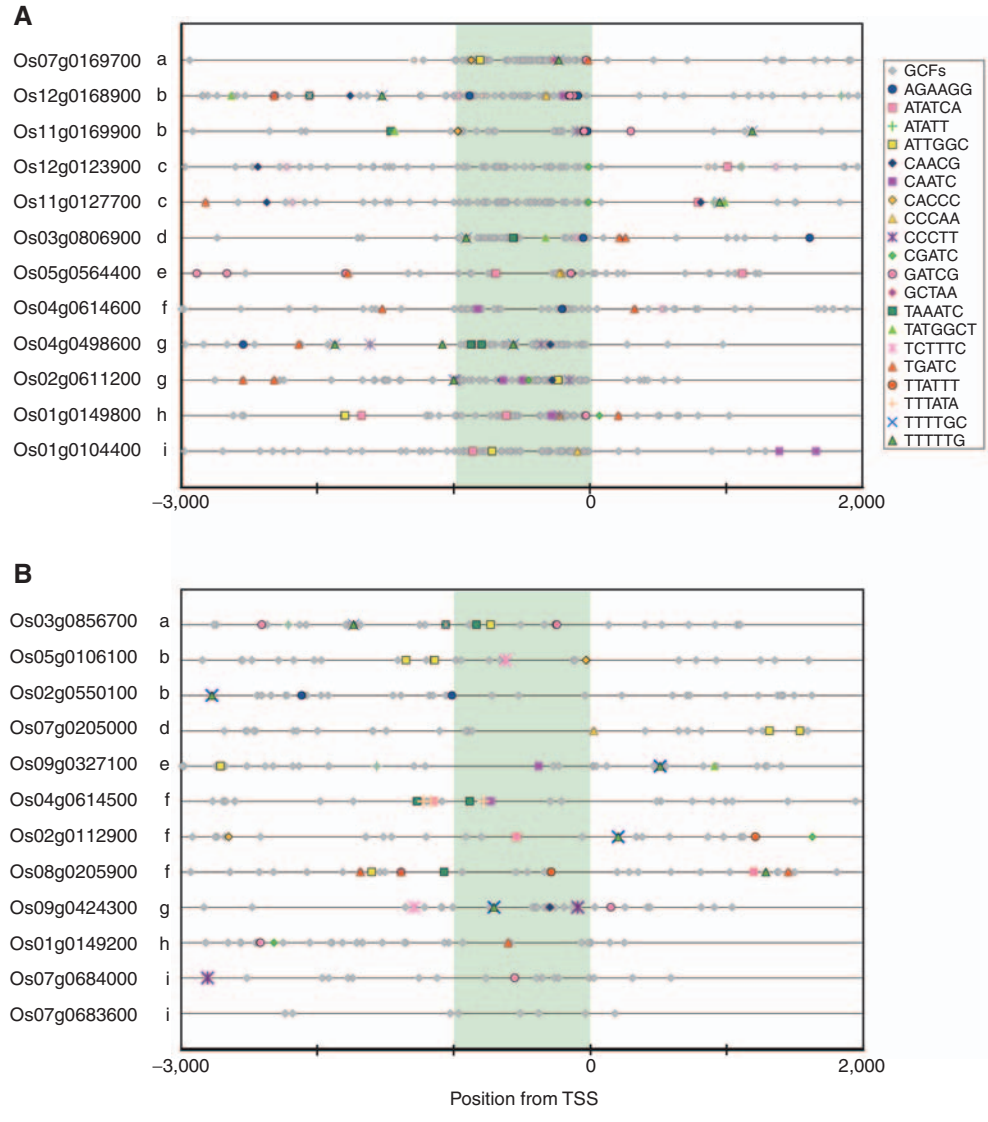

Fig. 11 Comparison of positional distribution of combined GCFs in paralogous genes. (A) Members in the 128-gene cluster. (B) Paralogous genes of the members not in the 128-gene cluster. The letters (a–i) in the second column indicate the paralogous relationship between (A) and (B). Note that color labels, which represent the positions of GCFs in the combined GCF sets, were observed in (A) more than in (B). GCFs (gray labels) were often observed in the 1,000 bp promoter regions of members.

(Haerizadeh et al. 2006). A simple search for the 8 bp GRSF-binding site in the promoter regions of the 128-gene cluster revealed that more than half of the genes contained a perfect match with the binding site (data not shown). This strongly suggested that the gene cluster reflected gene expression in sperm cells. Therefore, in this work, we tried to identify the *cis*-regulatory elements that may be involved in the activation of gamete-specific genes in rice by using the 128-gene cluster as a data set in order to provide a general purpose algorithm for identifying short nucleotides associated with specific gene expression.

Comparison with other *in silico* search methods for cis-regulatory elements

There are many methods available for finding *cis*-regulatory elements in given gene groups such as MEME (Bailey and Elkan 1994, Bailey et al. 2006), AlignACE (Thijs et al. 2001, Thijs et al. 2002) and Motif-Sampler (Hughes et al. 2000). Unfortunately, these software programs often do not provide enough information for the detection of genuine *cis*-regulatory elements for specific gene expression. In addition to the fact that they can record hits on common sequences of basic promoter constituents, an apparent lack of consideration

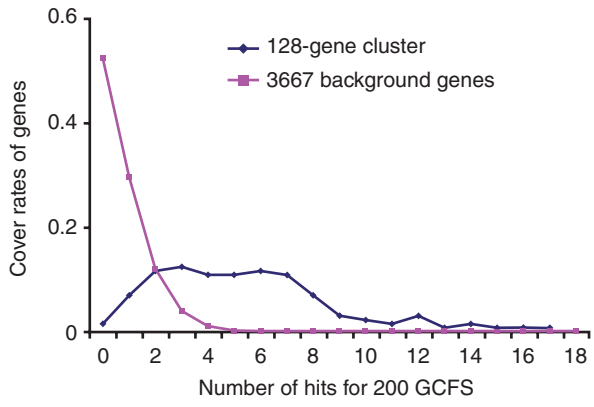

Fig. 12 Comparison of numbers of hits for the 200 GCFs between the 128-gene cluster and the 3,667 background genes. Hit numbers of the 200 GCFs in the 1,000 bp promoter regions were counted between the gene cluster of 128 genes and the 3,667 (3,795 − 128) background genes without outlier data. Gene ratios for hit numbers are presented.

for genome composition related to *cis*-regulatory elements and the local distributions of *cis*-regulatory elements gives these software packages severe limitations. Recently, with increasing genome information, LDSS was applied to the rice and *Arabidopsis* genomes and identified many promoter constituents categorized into major three groups, REG, TATA box and Y Patch (Yamamoto et al. 2007). This approach has provided a large amount of information on plant promoters, but it may miss the less frequently occurring *cis*-regulatory elements that are involved in specific gene expression in specified cells at particular developmental stages.

In this work, we searched for all possible short sequences in the target gene cluster (128 genes) and the total tested genes of rice (3,795 genes), and then compared their frequency of appearance on the basis of a simple binomial distribution to make a data set of GCFs. The approach to searching for short nucleotides in the target sequences is similar to the LDSS approach, but the purpose is completely different. In LDSS, promoter constituents were searched for by using the localization patterns of short nucleotides in the promoter regions; in contrast, to reduce complexity in the target sequences and remove common short nucleotides, we searched for short nucleotides that were over-represented in the target gene cluster in comparison with their occurrences in the background genes. By this selection using 1,000 bp promoter regions upstream of TSSs, we were able to highlight the short nucleotides that could be involved in specific gene expression, such as gene expression in rice pollen cells at the late bicellular and mature pollen stages. The criteria for making this GCF catalog were determined ad hoc in this work, and we therefore do not consider them optimal for determining GCFs for a given gene group yet. These criteria will need to be reconsidered empirically in the future. Note that the GCFs excluded the common promoter constituents such as the TATA box, since all the GCFs appeared with significantly higher frequencies in the gene cluster than in the background gene group. We tried several more stringent selection criteria in our preliminary experiments; they were still good enough to determine GCFs, but they were not sufficient to extract candidates for *cis*-regulatory elements efficiently since the use of these stringent criteria for GCFs did reduce the numbers of candidates for *cis*-regulatory elements. Therefore, in this work, we decided to make a catalog of GCFs that was as large as possible in order just to remove redundant sequences such as basic promoter constituents by using information from the tested rice genes. These trials allowed us to set our GCF selection criteria as <0.3 for the relative hit ratio (hit rates in the tested genes/hit rates in the target gene cluster) and $P < 0.01$ for frequency.

Twenty combined GCF sets containing 200 of the 2,008 GCFs were found (**Fig. 5**). Although there may be more GCFs containing the same core sequences, we did not consider them as candidates in this work, since we selected GCFs only when >30% of the 128 genes had GCFs containing a common core sequence. This cover rate of 30% for the selection of GCFs with a particular common core sequence also has not yet been proven to be optimal, and the value chosen should be reconsidered empirically. The point here is that GCFs combined into a set through the possession of a common core sequence can be considered as a TFBS. For some DNA-binding proteins, in vitro DNA binding site preferences have been examined in detail. Usually there is a common core sequence, such as ACGT, recognized by many bZIP proteins (Izawa et al. 1993, Foster et al. 1994). Preferences for sequences flanking the core sequences are more flexible. For instance, many bZIP proteins can bind GCCACGTGGC, but not CACACGTGTG (Izawa et al. 1993, Foster et al. 1994). In this case, one can consider CACGTG, termed the G-box, as a core sequence. Therefore, to find TFBSs *in silico*, we considered this characteristic of DNA-binding transcription factors: many TFBSs consist of a rigid core and flexible flanking sequences. This is why we tried to find any common sequences in GCFs as core sequences: the combined GCF sets can represent TFBSs containing both the rigid core and flexible flanking sequences. Therefore, after finding a candidate subset of GCFs with a common core sequence, we examined the preferences for sequences flanking the core sequence. Not all GCFs combined with a common core sequence and exhibiting a >30% cover ratio exhibited clear flanking sequence preferences, suggesting that these GCF subsets do not act as a TFBS to determine the specificity of gene expression. In contrast, some combined GCF sets showed clear preferences beyond the core sequences and may have a TFBS-like nature, with a rigid core and flexible

flanking sequences [logos and criterion (i) in **Fig. 5**]. If nucleotides in the flanking sites were distributed randomly, the probability for picking up the combined GCF set was less than one in a thousand. However, it was easy to find some preferences in the flanking regions for the selected GCFs among the 2,008 GCFs. Therefore, it is likely that the selected 2,008 GCFs often imply some common structure on the nucleotide sequence inside.

In addition, we also considered localization patterns as typical of TFBSs. It is well known that many *cis*-regulatory elements that may serve as TFBSs are localized in the promoter regions (Hughes et al. 2000, Wray et al. 2003). The LDSS method has already been used to demonstrate the importance of local distributions of *cis*-regulatory elements (Yamamoto et al. 2007). Because our 2,008 GCFs were selected from the 1,000 bp promoter regions in the 128-gene cluster, we expected the preferential occurrence of the combined GCFs in this 1,000 bp region. Therefore, we looked for local distribution patterns with a peak or more other than flat patterns in the 1,000 bp regions (**Fig. 5**); such localization patterns also imply that the combined GCFs can act as a TFBS.

Furthermore, GCFs with a core sequence of higher frequency in the gene cluster than in the tested genes were also considered to be candidates for *cis*-regulatory elements. Only one sequence, TATGGCT—a core sequence—exhibited a lower hit rate in the 1,000 bp promoter region of the background genes (the relative hit ratio 30%) (**Fig. 6**). Note that the TATGGCT as a GCF set exhibited the same 30% relative hit ratio and covered 39% of the 128-gene cluster, indicating that this 7 bp short sequence itself was enough to be considered a TFBS (**Figs. 5, 6**).

We therefore succeeded in proposing the 20 combined GCF sets (in total 200 GCFs) as candidates for *cis*-regulatory elements for pollen cell-specific gene expression at the late bicellular and mature pollen stages of rice pollen development. This two-step selection process—(i) selection for GCFs and (ii) selection for combined GCFs with a common core sequence—should provide us with a refined way to find *cis*-regulatory elements *in silico* and should give useful hints for elucidating transcriptional regulation in the specific expression of target genes.

In silico evaluation of these candidates readily exhibited both the effectiveness and limitation of our approach. First, six of the top 20 of the simple MEME analysis hit our core sequences, and the motifs in the MEME results were not significantly higher for the 128-gene cluster than the background since the relative hit ratios of these 20 MEME motifs (hit rates for the 3,795 genes/hit rates for the 128-gene cluster) were >62% (**Fig. 8**), indicating that our method, unlike MEME (Bailey and Elkan 1994, Bailey et al. 2006), could pick up novel candidates which are characteristics of the target gene cluster. It is likely that we can remove common promoter constituents efficiently using our algorithm.

In addition, eight OsREGs among the top 10 list (**Fig. 9**) of OsREGs found in the 128-gene cluster contained a ppdb motif, GCCCA, indicating that registered OsREGs are not enough to extract candidates of *cis*-regulatory elements for specific gene expression in rice pollen cells. Here, the ppdb motifs are a set of core sequences found in the OsREG sequences found by the LDSS method for all the rice genes (Yamamoto and Obokata 2008) (**Fig. 9**). This motif, GCCCA, was not selected as a core sequences for any combined GCFs. This may be because that LDSS often picked up the popular *cis*-regulatory elements in the genomes.

The expression profiles within the SALAD dendrogram in some paralogous groups support the possibility that these 20 combined GCF candidates contribute to specific gene expression, although there were notable exceptions (**Fig. 11**; Supplementary figures). In addition, examination of the profiles of genes aligned with the numbers of hits for the 200 GCFs (**Fig. 12**) showed that the majority of the 128 genes in the gene cluster had high numbers of hits for the 200 GCFs in the promoter regions, indicating that the 200 GCFs can represent a majority of the 128-gene cluster. Although it is possible that from our data we could find more GCFs combined through the possession of other common core sequences, we are not however sure whether >20 independent TFBSs (or combined GCFs) are involved in pollen cell-specific gene expression yet. Therefore, in the future, positional information between the combined GCFs and the effects of combination among these combined GCFs (possible TFBSs) would be the next priorities for the evaluation of candidates, as is experimental validation. Note that we selected GCFs from the 1,000 bp promoter regions of 128 genes in comparison with those of 3,795 genes without outlier data. Therefore, it is possible to evaluate the combined GCF sets to find genes which are expressed at the late bicellular and microspore stage from microarray data other than the data set of the 3,795 genes as an *in silico* validation experiment. We found 11 genes specifically expressed at those stages among 87 genes with more than five combined GCF hits; meanwhile, only three such genes were found among 87 genes randomly selected from the entire rice genome (data not shown). This result indicated that our method is in some way useful to predict genes with specific gene expression in pollen cells of rice.

Regarding the GRSF-binding sites, although there are four 8 bp core GRSF-binding site reported, none of them exhibited a significant occurrence rate by the binominal test. Therefore, we could not select such GRSF core sequences as GCFs in this work. This may be due to the fact that any repressor-binding sites may not be over-represented in promoter regions of genes exhibiting specific gene expression.

Finally, our combined GCF method would be useful to extract efficiently those short nucleotide sequences associated with stage- and tissue-specific gene expression and to

select candidates for *cis*-regulatory elements from microarray data, although experimental validation for these candidates has not been done as yet.

Facilitation of microarray data with the SALAD dendrogram

In general, microarray data are often clustered with normalized expression data and/or sample identifiers, that provide information on gene sets (for instance, see **Fig. 1**). This kind of clustering facilitated our search for *cis*-regulatory elements and the functional relationship among annotations in the same cluster, and, as a result, protein similarity between gene annotations has been paid less attention for the purpose of clustering microarray analysis data.

Here, we used our similarity dendrogram from the SALAD database to facilitate the analysis of microarray data (see **Fig. 10**) and provided a web site (URL: http://salad.dna.affrc.go.jp/CGViewer/MicroArrayPollen/) in which the microarray data are arranged by similarity of gene annotation. This type of presentation of microarray data provides a novel view of the regulation of gene expression in paralogous gene groups. For example, as published previously and in this issue (Chhun et al. 2007, Hirano et al. 2008), some genes related to gibberellin metabolism are expressed only in the late bicellular and mature pollen stages. On the web site we have provided it is apparent that one of four gibberellin 20-oxidase (GA20ox) genes exhibits specific gene expression at this stage, as does one of four gibberellin 3β-oxidase (GA3βox)-like genes (URL: http://salad.dna.affrc.go.jp/CGViewer/MicroArrayPollen/cgv_array_view.jsp?pfamid=At4g21690). This is surely useful information on gene function and on the redundancy of gene regulation. For instance, the GA20ox and GA3βox genes expressed specifically only at the mitotic pollen stage might be good targets for evaluation of the roles of gibberellin in pollen development by molecular genetics techniques since other related genes were not expressed at that stage.

Materials and Methods

Clustering of arrays and genes

To filter out noise from the 31 microarray samples of 43,624 genes, we selected genes in all arrays with the following criteria: glsSaturated = 0, glsFeatNonUnifOL = 0, glsBGNonUnifOL = 0, glsFeatPopnOL = 0, glsBGPopnOL = 0, glsBGPopnOL = 0, glsPosAndSignif = 1, and glsWellAboveBG = 1. Obtained expression data on a total of 5,110 genes were normalized by Z-score. Hierarchical clustering was performed with Cluster software (http://rana.lbl.gov/EisenSoftware.htm) (Eisen et al. 1998) with a single-linkage hierarchical clustering algorithm and visualized by TreeView (http://rana.lbl.gov/EisenSoftware.htm). Pearson's correlation was chosen as the similarity measure.

Upstream sequences in rice

We downloaded the GFF file, which is a format for describing genes, from the RAP-DB download site (http://rapdb.dna.affrc.go.jp/rapdownload/) (Rice Annotation Project 2008). Using the rep.gff file, we collected TSS information from a total of 28,237 representative genes for which we have information on full-length cDNA or ESTs in *O. sativa* genomes. Using this information and the International Rice Genome Sequencing Project genome sequence (build 4 assembly), we created two types of data sets (International Rice Genome Sequencing Project 2005). One type contained 1,000 bp upstream sequences, and the other contained 5,000 bp sequences (3,000 bp upstream and 2,000 bp downstream). If a gene was shorter than 2,000 bp, we used the shorter sequence as the downstream sequence. To remove repeat sequences, the data set was masked by RepeatMasker with TIGR's Oryza Repeat data (v. 3.1).

Detection of significant short nucleotides in upstream sequences of co-expressed genes

We used a total of 1,397,760 patterns (**Fig. 2**), which included all possible patterns of 5–10 bp sequences, for the analysis. We set two criteria to define GCFs in a gene cluster (i.e. a group of co-expressed genes). One was the *P*-value in a one-tailed binomial test. *P*-values were calculated using binom.test in R software (http://www.r-project.org/). To determine the probability of successfully selecting relevant sequences, we performed a statistical evaluation of the number of genes in the cluster that shared a common pattern in the 1,000 bp region upstream as a ratio of the number of genes in the 3,975 tested genes (derived from 5,110 probe data without outliers) that had the same common pattern. We used a *P*-value of <0.01.

The second criterion used to select GCFs was the pattern frequency. To choose patterns observed more frequently in the gene cluster than in the 3,975 genes, we calculated the relative hit ratio of GCFs, defined as:

Relative hit ratio of GCFs = (No. of patterns in the 3,975 tested genes × No. of genes in the gene cluster)/(No. of patterns in the gene cluster×No. of genes in the 3,975 tested genes).

We selected patterns with a relative hit ratio of <0.3.

Furthermore, we chose only those patterns with more than one hit in the cluster. We used the shorter one if an inclusion existed and then removed the inclusion of the patterns by removal of the longer ones as far as possible. The patterns remaining were defined as GCFs. In this works, 2,000 GCFs were selected.

Finding meaningful core sequences in the GCFs

To find meaningful common core sequences, we classified the GCFs by common 5, 6, 7 and 8 bp patterns. We counted

the occurrence of GCFs that had each core sequence in common in both the gene cluster and the 3,975 subject genes of the 5,000 bp data set, and we calculated the ratio of genes containing each core sequence between the gene cluster and the 3,975 genes. We then selected sets of GCFs possessing a certain core sequence that was present in >30% of genes in the cluster. We calculated the relative hit ratio of combined GCFs for each core sequence, defined as:

Relative hit ratio of combined GCFs = (No. of genes containing GCFs belonging to a combined GCF in the gene cluster×No. of genes in the 3,975 genes)/(No. of genes containing GCFs belonging to a combined GCF in the 3,975 genes×No. of genes in the gene cluster)

We then examined combined GCF candidates in the order of the increasing relative hit ratio.

Selection of combined GCFs with common core sequences

We further examined: (i) frequency-based sequence logos using WebLogo ver. 2.8.2 (Crooks et al. 2004) to search for preferences beside the core sequence (we considered that preference existed if the rate of occurrence of two specific bases was >0.5); (ii) the positional distribution pattern of each core sequence, to determine whether core sets were distributed in specific positions; and (iii) the relative hit ratios of GCFs as core sequences.

MEME analysis

We examined the upstream 1,000 bp of the 128 genes by using MEME (http://meme.sdsc.edu/meme/intro.html) to identify conserved nucleotide motifs present in the promoter regions. The program was set to output the top 20 motifs and a motif length ranging between 5 and 10 bp. The other option values used were the defaults. We compared the MEME motifs with the defined core sequences (**Fig. 8**).

ppdb search

A total of 242 LDSS-positive REG octamers in *O. sativa* (OsREG) were searched in the 1,000 bp promoter regions of the 128-gene cluster and aligned with the decreasing numbers of hits (**Fig. 9**).

Display of microarray data on the SALAD database

We created data for each RAP locus in the SALAD database. However, because the rice 44K oligo microarray was designed on the basis of full-length cDNAs or ESTs, a single RAP locus may sometimes correspond to several items of microarray data. We therefore calculated an average gene expression value for each RAP locus. This value was then depicted by a color gradient, with deeper colors indicating higher levels of expression in the SALAD database (http://salad.dna.affrc.go.jp/CGViewer/MicroArrayPollen/) (**Fig. 10**).

Supplementary Material

Supplementary Material are available at PCP Online.

Funding

The Genomics for Agricultural Innovation Project (grant Nos. GIR1001 to T. Itoh, GIR1002, RTR0004 to T. Izawa.); the Ministry of Agriculture, Forestry, and Fisheries of Japan; the Ministry of Education Grants-in-Aid for Scientific Research in Priority Areas (to T. Izawa.).

Acknowledgments

We thank the members of the LM-microarray of rice pollen project for providing microarray data.

References

Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2: 28–36.

Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34: W369–W373.

Chhun, T., Aya, K., Asano, K., Yamamoto, E., Morinaka, Y., Watanabe, M., et al. (2007) Gibberellin regulates pollen viability and pollen tube growth in rice. *Plant Cell* 19: 3876–3888.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res* 14: 1188–1190.

Defrance, M. and Touzet, H. (2006) Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC Bioinformatics* 7: 396.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* 95: 14863–14868.

Endo, M., Tsuchiya, T., Saito, H., Matsubara, H., Hakozaki, H., et al. (2004) Identification and molecular characterization of novel anther-specific genes in japonica rice, *Oryza sativa* L. by using cDNA microarray. *Genes Genet. Syst.* 79: 213–226.

Engel, M.L., Chaboud, A., Dumas, C. and McCormick, S. (2003) Sperm cells of Zea mays have a complex complement of mRNAs. *Plant J.* 34: 697–707.

Foster, R., Izawa, T. and Chua, N.H. (1994) Plant bZIP proteins gather at ACGT elements. *FASEB J.* 8: 192–200.

Haerizadeh, F., Singh, M.B. and Bhalla, P.L. (2006) Transcriptional repression distinguishes somatic from germ cell lineages in a plant. *Science* 313: 496–499.

Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27: 297–300.

Hirano, K., Aya, K., Hobo, T., Sakakibara, H., Kojima, M., Shim, R.A., et al. (2008) Comprehensive transcriptome analysis of phytohormone biosynthesis and signaling genes in microspore/pollen and tapetum of rice. *Plant Cell Physiol.* 49: 1429–1450.

Hobo, T., Suwabe, K., Aya, K., Suzuki, G., Yano, K., et al. (2008) Various spatiotemporal expression profiles of anther-expressed genes in rice. *Plant Cell Physiol.* 49: 1417–1428.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296: 1205–1214.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800.

Izawa, T., Foster, R. and Chua, N.H. (1993) Plant bZIP protein DNA binding specificity. *J. Mol. Biol.* 230: 1131–1144.

Kim, N.K., Tharakaraman, K., Marino-Ramirez, L. and Spouge, J.L. (2008) Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics* 9: 262.

Kim, N.K., Tharakaraman, K. and Spouge, J.L. (2006) Adding sequence context to a Markov background model improves the identification of regulatory elements. *Bioinformatics* 22: 2870–2875.

Nakazono, M., Qiu, F., Borsuk, L.A. and Schnable, P.S. (2003) Laser-capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: identification of genes expressed differentially in epidermal cells or vascular tissues of maize. *Plant Cell* 15: 583–596.

Okada, T., Bhalla, P.L. and Singh, M.B. (2006) Expressed sequence tag analysis of *Lilium longiflorum* generative cells. *Plant Cell Physiol.* 47: 698–705.

Pina, C., Pinto, F., Feijo, J.A. and Becker, J.D. (2005) Gene family analysis of the Arabidopsis pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation. *Plant Physiol.* 138: 744–756.

Rice Annotation Project (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* 36: D1028–D1033.

Singh, M.B. and Bhalla, P.L. (2007) Control of male germ-cell development in flowering plants. *Bioessays* 29: 1124–1132.

Suwabe, K., Suzuki, G., Takahashi, H., Shiono, K., Endo, M., et al. (2008) Separated transcriptomes of male gametophyte and tapetum in rice: validity of a laser microdisection (LM) microarray. *Plant Cell Physiol.* 49: 1407–1416.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., et al. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17: 1113–1122.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., et al. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.* 9: 447–464.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., et al. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20: 1377–1419.

Yamamoto, Y.Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., et al. (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* 8: 67.

Yamamoto, Y.Y. and Obokata, J. (2008) ppdb: a plant promoter database. *Nucleic Acids Res.* 36: D977–D981.

Young, J.A., Johnson, J.R., Benner, C., Yan, S.F., Chen, K., Le Roch, K.G., et al. (2008) In silico discovery of transcription regulatory elements in *Plasmodium falciparum*. *BMC Genomics* 9: 70.

(Recieved July 31, 2008; Accepted August 29, 2008)