MASSIVE COLLECTION OF FULL-LENGTH COMPLEMENTARY DNA CLONES AND MICROARRAY ANALYSES: KEYS TO RICE TRANSCRIPTOME ANALYSIS*

SHOSHI KIKUCHI†
Plant Genome Research Unit
Division of Genome and Biodiversity Research
National Institute of Agrobiological Sciences (NIAS)
Kannondai 2-1-2
Tsukuba, Ibaraki 305-8602, Japan

Completion of the high-precision genome sequence analysis of rice led to the collection of about 35,000 full-length cDNA clones and the determination of their complete sequences. Mapping of these full-length cDNA sequences has given us information on (1) the number of genes expressed in the rice genome; (2) the start and end positions and exon–intron structures of rice genes; (3) alternative transcripts; (4) possible encoded proteins; (5) non-protein-coding (np) RNAs; (6) the density of gene localization on the chromosome; (7) setting the parameters of gene prediction programs; and (8) the construction of a microarray system that monitors global gene expression. Manual curation for rice gene annotation by using mapping information on full-length cDNA and EST assemblies has revealed about 32,000 expressed genes in the rice genome. Analysis of major gene families, such as those encoding membrane transport proteins (pumps, ion channels, and secondary transporters), along with the evolution from bacteria to higher animals and plants, reveals how gene numbers have increased through adaptation to circumstances. Family-based gene annotation also gives us a new way of comparing organisms. Massive amounts of data on gene expression under many kinds of physiological conditions are being accumulated in rice oligoarrays (22K and 44K) based on full-length cDNA sequences. Cluster analyses of genes that have the same promoter *cis*-elements, that have similar expression profiles, or that encode enzymes in the same metabolic pathways or signal transduction cascades give us clues to understanding the networks of gene expression in rice. As a tool for that purpose, we recently developed "RiCES", a tool for searching for *cis*-elements in the promoter regions of clustered genes.

*This work was partly supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Integrated Research Project on Plants, Insects, and Animals Using Genome Technology, no. QT1003) and by a grant from the budget of the Generation Challenge Program SP4 project.

1. Introduction

1.1. *Massive collection of full-length cDNA clones and their sequence information: starting materials for transcriptome analysis*

Highly accurate genome sequences of rice are available (1–4). Genome sequences of rice (*O. sativa* ssp. *japonica* 'Nipponbare') have been assembled independently by The Institute for Genome Research (TIGR), the International Rice Genome Sequencing Project (IRGSP), and the Beijing Genomics Institute (BGI). In addition, full-length complementary DNA (FL-cDNA) sequences (5, 6) and expressed sequence tags (ESTs) (7–9) from rice have served as valuable resources for genomic and genetic studies.

FL-cDNA collections have been established for several organisms (10–12). The cDNA information has been used extensively to determine gene annotation, structures, and the start and end sites of transcription. The cDNA information is also indispensable for validation of gene function by reverse genetics. The rice FL-cDNA collection has also been used to complement the information obtained by genome sequencing. The first report of a rice FL-cDNA collection, published in 2003 (5), described the characteristics and annotation of 28K FL-cDNA sequences. The collection has now expanded to 578K FL-cDNA clones, among which 35K cDNA clones were completely sequenced and annotated (available at Knowledge-based Oryza Molecular biological Encyclopedia (KOME): http://cdna01.dna.affrc.go.jp/cDNA/).

Annotation of the complete genome of Nipponbare was performed by the Rice Annotation Project (RAP). All functional annotations for proteins and non-protein-coding RNA (npRNA) candidates were manually curated. Functions were identified or inferred for 19,969 (70%) of the proteins, and 131 possible npRNAs (including 58 antisense transcripts) were found. Almost 5000 annotated protein-coding genes were found to be disrupted in insertional mutant lines. The rice loci were determined by using cDNA sequences obtained from rice and other representative cereals. Our conservative estimate based on these loci and an extrapolation suggested that there are about 32,000 rice genes—fewer than previous estimates (13).

Here we show the comparison of the transcription units (TUs) mapped by the full-length cDNA sequences to the rice genome sequence and the loci predicted by the gene annotation computer program (TIGR-CDS). About 5400 TUs are generated only by the full-length cDNA mapping and are not predicted by the gene annotation program. Those are so-called non-annotated expressed (NAE) genes. Detailed structural, gene expression, and homology analyses have

revealed that many of the rice NAE genes are similar to the npRNA genes in mouse.

1.2. *Gene-family–based functional annotation*

Family-based gene annotation was launched after global gene annotation. We will first focus here on the genes encoding membrane-associated proteins. Cells maintain their biological activities by importing and exporting various substances. Provision of energy and nutrients and efflux of salts, biochemicals, and ions are necessary to maintain biological activity in prokaryotic and eukaryotic cells. Environmental situations within cells differ among organisms: unicellular organisms cannot control the ion concentrations outside cells, but multicellular eukaryotes (especially animals) can precisely regulate the ion concentrations of their cellular environments within micromolar ranges. Therefore, we can expect organisms to differ in gene number, structure, and function according to their biological abilities and environmental situations.

We searched for orthologs of known membrane transport genes among the 35,180 full-length rice cDNA sequences (5, 6) and genomic sequence data from Arabidopsis (14) and *japonica* rice (1,4), and among global functional gene annotations in Arabidopsis and rice (Munich Information Center for Protein Sequences [MIPS] data service, http://mips.gsf.de/proj/plant/jsf/ (15, 16); Rice Annotation Project Data Base [RAP-DB], http://rapdb.lab.nig.ac.jp/ (13, 17); TIGR Rice Genome Annotation, http://www.tigr.org/tdb/e2k1/osa1/index.shtml (18)). Transmembrane proteins have a hydrophobic structure, a pore-forming sequence, and molecule-binding sites. Because of these specific structural features, the identification of membrane transport orthologs is clear from computer calculations. Previous reports have characterized individual transporter protein families but have not extended to whole transport systems in general (19–22). In a more general analysis of various organisms, the features of prokaryotes were contrasted with those of eukaryotes (23). However, differences among eukaryotes—especially animals and plants—were not a focus of that analysis. We also searched for orthologs of membrane transport genes in various organism databases (Human Gene Nomenclature Database Search Engine, http://www.genenames.org/cgi-bin/hgnc_search.pl (24); Genomic Comparison of Membrane Transport Systems [TransportDB], http://www.membranetransport.org/index.html (25); Functional Genomics of Plant Transporters [PlantsT], http://plantst.genomics.purdue.edu/ (26); ARAMEMNON, http://aramemnon.botanik.uni-koeln.de/ (27)). We compare total membrane transport systems from diverse organisms and conclude that

membrane transport genes exemplify the evolutionary diversity of homeostatic systems. The evolutionary changes in gene families indicate the dynamics of alterations in biological systems and gene networks. Therefore, analysis of large categories of gene families may reveal many basic concepts of biological systems.

1.3. *Establishment of rice microarray systems and development of a tool for searching for* cis*-elements in the promoter regions of clustered genes*

On the basis of the results of our large-scale FL-cDNA analysis (5), we have constructed a monitoring system that uses an oligonucleotide array to monitor gene transcriptional levels and to develop genome-wide functional analysis of rice. The array (22K rice oligoarray) was composed of 21,938 probes with 60-mer oligonucleotides synthesized at gene-specific regions (28–30) from 32,127 FL-cDNAs. Mapping of these cDNA clones to genomic DNA revealed that there were about 20,500 TUs, and clustering of the clones revealed a unique clone set. Two 22K array platforms are registered in NCBI-GEO (National Center for Biotechnology Information – Gene Expression Omnibus). Platform GPL477 is a prototype version of the 22K array. One study, which was the comparison of gene expression profiles of rice callus by the Gibberellic acid and Abscisic acid treatments has used this array system (31). Platform GPL892 is the commercial version of the 22K array (Agilent Technologies catalog number G4138A) and has been used to accumulate many environmental stress data. At the Rice Genome Resource Center at our institute we have used these arrays and a new 4 × 44K format from Agilent Technologies to develop a new rice oligoarray (GPL6864) covering all the genes expressed in the rice genome.

If good-quality RNA samples are prepared, the oligomicroarray system provides us with highly reproducible gene expression data. It takes just a few days to obtain the gene expression data, but the subsequent data mining process can take many months. For systematic data analyses, functional annotation data on each probed gene must be well facilitated and good tools must be available. One such useful tool would search for *cis*-elements in the promoter regions of clustered gene sets after microarray analysis. The existence of common *cis*-elements in the promoter regions of clustered gene sets may suggest that those genes are controlled by the same transcription factor.

Cis-elements in the promoter regions of genes and *trans*-acting transcription factors are major biological features to be characterized if we are to achieve an understanding of the systems that regulate gene expression. Identification of candidate *cis*-elements corresponding to genes is now practicable through the

use of available sequence and genome mapping information, combined with information about the responses of genes to specific experimental conditions; such responses have been elucidated by using the gene expression profiles now publicly available.

Exhaustive sequence analysis using available public databases can identify *cis*-element candidate motifs for further examination, but such approaches are not efficient. One confounding factor is that public databases are independently constructed and not generally optimized to facilitate the integration of information from many sources with local experimental data. A more perplexing issue for experimental researchers who are not familiar with bioinformatics techniques is the challenge of finding unknown but biologically notable relationships among genes, *cis*-elements, and experimental conditions from the huge number of possible combinations generated by large experimental datasets.

To resolve some of these issues, we developed a novel data mining tool to identify *cis*-elements in the rice genome. It performs the complex bioinformatics analysis mentioned above, and then lists *cis*-element candidates for genes. The genes can be grouped by similarity of expression profiles and other criteria for assessment by researchers, and the tool then annotates them with related public database information.

Similar tools have been developed previously. Helden released RSAT, which includes a program that can detect motifs over-represented in the upstream regions of co-regulated genes (32). Holt et al. established CoReg, which links the hierarchical clustering of co-expressed gene sets with frequency tables of promoter elements (33). Zhao et al. established TRED, which integrates a database and a system for predicting *cis*- and *trans*-elements in mammals (34).

Our novel tool searches for *cis*-element candidates in the upstream, downstream, or coding regions of differentially regulated genes. The tool first lists *cis*-element candidates by motif searching based on the supposition that if there are *cis*-elements playing important roles in the regulation of a given set of genes then they will be statistically over-represented and will be conserved. Then it evaluates the likelihood scores of the listed candidate motifs by association rule analysis. This strategy depends on the idea that motifs over-represented in the promoter region could play specific roles in the regulation of expression of these genes. The tool is designed so that any biological researchers can use it easily at the publicly accessible Internet site http://hpc.irri.cgiar.org/tool/nias/ces. We evaluated the accuracy and utility of the tool by using a dataset of auxin-inducible genes that have well-studied *cis*-

elements. The test showed the effectiveness of the tool in identifying significant relationships between *cis*-element candidates and related sets of genes.

2. Results and Discussion

2.1 *Comparison of gene models predicted by gene prediction programs and transcription units identified by the mapping of full-length cDNA sequences*

We mapped 578,000 rice (ssp. *japonica*, 'Nipponbare') FL-cDNA clones (DDBJ accession numbers: completely sequenced FL-cDNA: AK058203–068528, AK068530–068912, AK068914–70720, AK070722–074028, AK098843–112119, AK119160–122186, AK240633–243692; one-pass sequences of FL-cDNA, 5′ end: CI285358–311811 and CI563340–778739, 3′ end: CI000001–285357 and CI311812–563339) to five genome assemblies: TIGR4, IRGSP3, IRGSP4, the Nipponbare genome determined by Syngenta, and the 93-11 genome by BGI (Table 1). The mapping criteria were 95% identity and 90% coverage. For proper comparison between the assemblies, the results of mapped cDNAs in the respective assemblies were compared with those of TIGR4. The numbers of mapped FL-cDNA clones differed among assemblies, with the highest in TIGR4. The orientation in which some cDNA clones were mapped onto a chromosome and the chromosome on which some clones were mapped were not consistent among the assemblies. Of the 32,775 completely sequenced FL-cDNA clones mapped in TIGR4, 29,925 were also mapped in all of the other assemblies; however, the number of clones commonly mapped in both TIGR4 and a given assembly differed (Table 1). The maximum and minimum numbers of common clones were 32,730 in IRGSP4 and 30,162 in 93-11, respectively. The number of mapped clones was greater in the *japonica* rice genomes than in the *indica* genomes; this might reflect differences in the genome sequences between subspecies. The number of clones common to TIGR4 and IRGSP4 was close to that common to TIGR4 and IRGSP3, and both numbers were greater than the number of clones common to TIGR4 and the Syngenta sequence. This suggests that the differences in numbers of common clones may have resulted from differences in the sequencing methods adopted in the assemblies (TIGR4 and IRGSP by the map-based method; Syngenta sequence by the whole-genome shotgun method). Mapping of 578K FL-cDNA clones identified about 28,500 loci in the *japonica* genome and 27,800 loci in the *indica* genome. A total of 29,925 completely sequenced FL-cDNAs were mapped in all the genome assemblies, and more than 90% of the FL-cDNAs

were mapped in all five assemblies (Table 1). We therefore decided to use only the mapping results of TIGR4 for further analyses, not those from the other assemblies. The number of predicted loci was about 56K, which was sufficient for our data analysis but probably not sufficient to reach complete accuracy of gene prediction and annotation of TIGR4.

Table 1. FL-cDNA clones mapped to five rice genome assemblies.
(Source: Satoh et al. PLoS One 2, e1235(6))

	origin sequencing All	*japonica* genome				*indica* genome
		Map-base cloning			whole shotgun	
		TIGR	IRGSP4	IRGSP3	Syngenta	93-11
FL-cDNA	35,187	32,775	32,745	32,640	31,928	30,354
5endFLEST	241,854	212,598	212,539	211,564	208,606	199,001
3endFLEST	536,885	483,657	484,358	482,909	482,665	465,775
FL-cDNA locus	Chr1	4,026	4,021	4,039	4,050	3,940
	Chr2	3,196	3,198	3,215	3,186	3,153
	Chr3	3,569	3,567	3,566	3,597	3,607
	Chr4	2,531	2,530	2,534	2,477	2,493
	Chr5	2,313	2,305	2,310	2,338	2,329
	Chr6	2,292	2,293	2,290	2,262	2,266
	Chr7	2,183	2,185	2,193	2,165	2,021
	Chr8	1,933	1,934	1,939	1,912	1,827
	Chr9	1,605	1,605	1,574	1,545	1,515
	Chr10	1,538	1,528	1,536	1,502	1,416
	Chr11	1,685	1,683	1,675	1,486	1,333
	Chr12	1,693	1,692	1,705	1,523	1,435
	Chr0 [a]				434	497
	Total	28,564	28,541	28,576	28,477	27,832
Comparison of FL-cDNA mapping with TIGR4		Both mapped	32730	32623	31741	30162
		Same Chr-Same Strand	32646	32611	30422	28760
		Same Chr-Reverse Strand	80	10	317	335
		Differential Chr.	4	2	1002	1067
		Mapped on only TIGR	45	152	1034	2613
		Unmapped on only TIGR	15	17	187	192
		Both unmapped	2397	2395	2225	2220
		mapped on all assemblies			29925	
		unmapped on all assemblies			2186	

[a]: sequence-assembled contigs that were not localized to one of the 12 chromosomes.
doi:10.1371/journal.pone.0001235.t001

A total of 55,890 gene loci were predicted in the rice genome according to TIGR OSA1 release 4. Mapping of FL-cDNA clones on TIGR4 revealed that 533,667 FL-cDNA clones were derived from 28,564 FL-cDNA loci (Table 2). FL-cDNA loci were cross-referenced with TIGR4 loci to examine the overlaps between the two groups. According to the sources of mapped loci and the occurrence of overlap, the loci were classified as follows: (1) when a FL-cDNA locus overlapped with a TIGR4 locus, the FL-cDNA locus was defined as FL-AE (annotated expressed) and the TIGR4 locus was defined as coding-sequence–AE (CDS-AE); (2) a FL-cDNA locus that did not overlap with any TIGR4 locus was defined as FL-NAE (non-annotated expressed); and (3) a

TIGR4 locus that did not overlap with any FL-cDNA locus was defined as CDS-ANE (annotated non-expressed). On the basis of these definitions, the loci were classified into 23,117 FL-AE, 23,193 CDS-AE, 5447 FL-NAE, and 32,697 CDS-ANE (Table 2).

Table 2. Comparisons of FL-cDNA loci and TIGR4 loci
(Source: Satoh et al. PLoS One 2, e1235)

		Class		
		AE	NAE	ANE
TIGR CDS		23193	0	32697
FL-locus		23117	5447	0
mapping information	FL-cDNA	29808	2967	0
	5endFLEST	201343	11255	0
	3endFLEST	465816	17481	0
	FL-clones	511817	21850	0

doi:10.1371/journal.pone.0001235.t002

The classification of loci as defined above raised questions about whether any characteristic distinctions existed between FL-AE and FL-NAE, and why FL-NAE loci were not predicted in TIGR4. To answer these questions, we analyzed the structures of genes belonging to the respective groups.

Open reading frames of FL-cDNA clones mapped on the genome

We mapped 32,775 FL-cDNAs at 22,943 FL-cDNA loci (FL-AE, 20,324; FL-NAE, 2619). The numbers of FL-cDNAs mapped to FL-AE and FL-NAE were 29,808 and 2967, respectively (Table 2). The median lengths of the FL-cDNA mapped to FL-AE and FL-NAE were 1540 and 1173 bp, respectively (Figure 1a, Table 3).

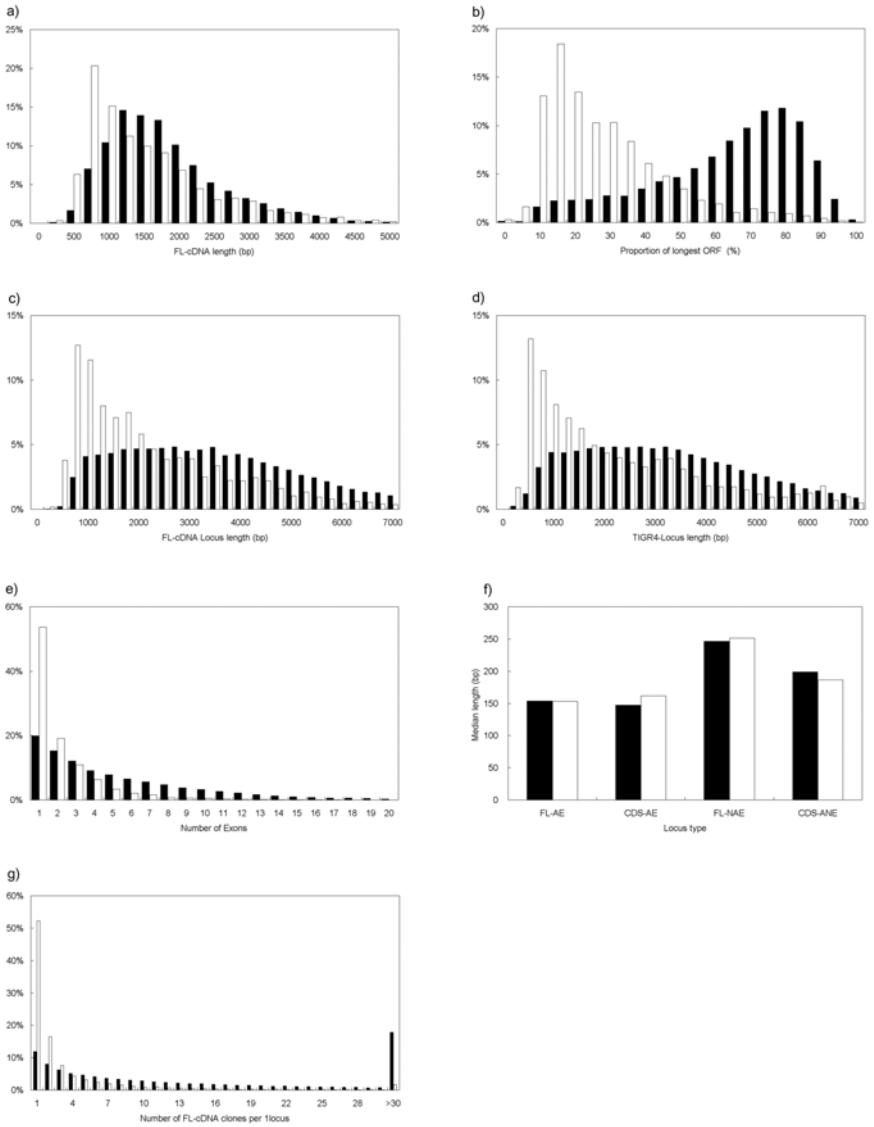

Figure 1 Gene structure analysis in rice. (Source: Satoh et al. PLoS One 2, e1235)
(a) Length distribution of FL-cDNA for FL-AE (black) and FL-NAE (white). (b) Distribution of open reading frame (ORF) proportions for FL-AE (black) and FL-NAE (white). (c) Distribution of FL-cDNA locus lengths for FL-AE (black) and FL-NAE (white). (d) Distribution of locus lengths for CDS-AE (black) and CDS-ANE (white) in TIGR4. (e) Distribution of number of exons for FL-AE (black) and FL-NAE (white). (f) Distribution of exon (black) and intron (white) lengths for the respective locus types. (g) Distribution of number of FL-cDNA clones mapped per single FL-AE (black) and FL-NAE (white) locus.

The proportions of the length of the longest open reading frames (ORFs) to that of FL-cDNA were also considerably different between FL-cDNAs mapped to FL-AE and those mapped to FL-NAE. The median proportion of the longest ORF in FL-AE was 66%, versus only 21% in FL-NAE (Figure 1b, Table 3). The results indicate that FL-NAE clones generally encode shorter peptides, and that the ORF lengths differ considerably between FL-AE and FL-NAE (based on the clone length × ORF proportion).

Table 3 Structural characteristics of locus types (Source: Satoh et al. PLoS One 2, e1235)

	FL-cDNA length (median)	ORF ratio (median)	Locus length (median) [a]	Variation of locus length	Number of exons (average) [b]	Exon length (median) [b]	Intron length (median) [c]	Ave.nunber of mapped FL-cDNA clones
FL-AE	1540	66%	3354	Rich	5.3	154	153	22.3
FL-NAE	1173	21%	1727	Poor (short)	2.4	247	251	4.1
CDS-AE	-	-	3173	Rich	5.8	147	162	-
CDS-ANE	-	-	1643	middle	3.9	199	186	-

[a]: For the calculation of locus lengths, we used the maximum lengths of individual loci.
[b]: Exons shorter than 10 bp were excluded from the analysis. Thus, the definition of an exon in FL-cDNA loci differs from that in TIGR OSA1.
[c]: Introns shorter than 10 bp were excluded from the analysis. Thus, the definition of an intron in FL-cDNA loci differs from that in TIGR OSA1.
doi:10.1371/journal.pone.0001235.t003

Locus length

The start and end sites for transcription were determined in 24,164 loci (FL-AE, 21,263; FL-NAE, 2901) out of 28,564 FL-cDNA loci. The distance between the start and the end sites (i.e., locus length) was calculated by using TIGR4. The median locus lengths differed considerably between FL-AE and FL-NAE, with lengths of 3354 and 1727 bp, respectively (Figure 1c, Table 3). The average ratio of locus length to FL-cDNA clone length was greater than 2 for FL-AE, but less than 1.5 for FL-NAE. The median locus lengths in CDS-AE and CDS-ANE were 3173 and 1643 bp, respectively (Figure 1d). The locus lengths in CDS-ANE appeared to be more variable than those in CDS-AE. The patterns of locus length variation in CDS-AE and CDS-ANE were similar to those in FL-AE and FL-NAE, respectively (Figure 1c, d, Table 3). We could not compare the lengths of FL-cDNA loci with those of the TIGR4 loci, because FL-cDNAs were constructed from coding and 5′- and 3′-end untranslated regions, whereas many gene structures in TIGR4 were predicted only from coding regions. However, from the results above, we expect that FL-AE and FL-NAE would differ from each other in this characteristic, as would CDS-AE and CDS-ANE.

Exon-intron structure

The average numbers of exons per FL-AE and FL-NAE were 5.3 and 2.3, respectively (Figure 1e, Table 3). The frequency of loci with a single exon was highest in both FL-AE and FL-NAE when the loci were distributed according to the numbers of exons. However, the proportions of loci with a single exon were significantly different between FL-AE and FL-NAE, with more than 50% in FL-NAE and about 20% in FL-AE (Figure 1e). When FL-cDNA loci with one exon were excluded, the ratio between the lengths of exons and introns in individual loci was approximately 1 irrespective of the locus type, but exon and intron lengths were significantly different ($P < 0.01$; Student's *t*-test) between FL-AE and FL-NAE, with median lengths of about 150 and 250 bp, respectively (Figure 1f, Table 3). We also analyzed the exon–intron structures of CDS-AE and CDS-ANE from the information at TIGR OSA1. The number of exons was higher in CDS-AE (5.8) than in CDS-ANE (3.9) (Table 3). The exon and intron lengths also differed significantly ($P < 0.01$, as calculated by Student's *t*-test) between CDS-AE and CDS-ANE, with median exon lengths of 147 and 199 bp and median intron lengths of 162 and 186 bp, respectively (Figure 1f, Table 3).

Number of mapped FL-cDNA clones at a locus

FL-NAE accounted for about 19% of the entire FL-cDNA loci, whereas the proportion of FL-cDNA clones mapped as FL-NAE was only 5% (21,850) of all mapped FL-cDNA clones (533,667) (Table 2). The average numbers of FL-cDNA clones mapped per locus (collection efficiency) differed significantly between FL-AE (22.3 clones) and FL-NAE (4.1) (Figure 1g, Table 3). The collection efficiency was 1 for more than half of the FL-NAE loci (Figure 1g, Table 3), suggesting that FL-cDNA clones derived from FL-NAE are more difficult to collect than those from FL-AE.

Gene annotation

We analyzed the homology of FL-cDNA mapped on TIGR4 with Arabidopsis CDSs in The Arabidopsis Information Resource (TAIR6, http://www.arabidopsis.org/) (35) using BlastX software. On the basis of the significance of similarity, FL-cDNA clones were classified into highly homologous (E-value $< 10^{-50}$), weakly homologous ($10^{-50} <$ E-value $< 10^{-10}$), and non-homologous (E-value $> 10^{-10}$). Under these criteria, the numbers of FL-cDNA clones classified as highly homologous, weakly homologous, and non-homologous were 17,759, 7103, and 7913, respectively. Of these clones, 99.5% of highly homologous FL-cDNAs were mapped to 59% of FL-AE, and 92% of

FL-NAE coded non-homologous genes (Table 4). Thus, the results indicate that nearly all highly homologous genes were derived from FL-AE, and that most FL-NAE loci encoded genes likely to be specific to rice or other monocots. This is consistent with the findings of a previous report that some rice genes with no homologs in Arabidopsis are similar to genes in the sorghum genome (3).

Causes of inconsistency between gene prediction and FL-cDNA mapping

Table 4. Frequency of occurrence of Arabidopsis homologous genes at each FL-locus (Source: Satoh et al. PLoS One 2, e1235(6))

	FL-AE		FL-NAE		Total	
homology [a]	Locus	FLcDNA	Locus	FLcDNA	Locus	FL-cDNA
HH	11898	17669	75	90	11973	17759
LH	4763	6941	140	162	4903	7103
NH	3663	5198	2404	2715	6067	7913
Total	20324	29808	2619	2967	22943	32775

[a]: HH, LH, NH: highly-, low- or non-homologous FL-cDNA with Arabidopsis CDS
doi:10.1371/journal.pone.0001235.t004

Cross-examination between FL-cDNA and TIGR4 loci revealed the existence of FL-NAE clones. The results of our analyses of the locus structures of FL-AE and FL-NAE suggest explanations of why some expressed genes were not annotated. One possible reason is the characteristic lengths of ORFs in the different classes of loci. The proportion of the longest ORF in FL-NAE (median ratio 21%) is significantly lower than that in FL-AE (median ratio 66%) indicating that the transcripts from FL-NAE are more likely to encode either small peptides or no peptide. In the TIGR OSA1, 687 CDSs encoding fewer than 50 amino acids were excluded from the predicted gene model. Thus, even though the number of excluded CDSs was less than the number of FL-NAE sequences, FL-cDNA sequences overlapping with the excluded CDSs in TIGR4 might have been mapped as FL-AE. Another possible reason may be the difference in exon–intron structure between FL-NAE and FL-AE. The locus lengths of FL-NAE were generally shorter than those of FL-AE, and more than half of the FL-NAE loci contained only one exon (Figure 1e). Meanwhile, the lengths of both exon and intron in FL-NAE were generally greater than those in FL-AE, CDS-AE, and CDS-ANE (Figure 1f). If we consider the structures of FL-AE, CDS-AEs, and CDS-ANE as standard for rice genes, then the structure

of many FL-NAE loci may be recognized as an irregular form. In light of the unique features of ORFs and the exon–intron structures in FL-NAE, it may have been difficult to assign proper annotations to the genes in FL-NAE through the use of gene prediction software.

The structural difference between FL-AE and NAE is also similar to that between protein-coding mRNA and mRNA-like non-coding RNA (npRNA) in the mouse (36). In the mouse, the total length of npRNA is shorter and the exon is longer than in mRNA. Moreover, more than 70% of npRNA is constructed from one exon. In our classification, FL-AE overlaps with predicted CDSs that encode >50 amino acid sequences, so cDNAs mapped on FL-AE originate from protein-encoding mRNAs. In addition, the diversity of FL-cDNA lengths between FL-AE and FL-NAE is not large and the proportions of ORFs between FL-AE and FL-NAE are reversed (Figure 1a, b, Table 3). So it seems that the proportion of protein-coding FL-NAE loci is not large, and many FL-NAE loci encode mRNA-like npRNA. Therefore, the structural diversity between FL-AE and FL-NAE may correspond to the difference between protein-coding mRNA and mRNA-like npRNA in rice. In addition, these results suggest that the structural differences between protein-coding mRNA and mRNA-like npRNA are also conserved between plants and mammals.

We categorized the gene loci in TIGR4 into three types by cross-examination between FL-cDNA loci and TIGR4 loci. Collection efficiency varied considerably depending on locus type (FL-AE = 22.3; FL-NAE = 4.1; CDS-ANE = 0). These differences may be associated with the levels of mRNA or the transcription activity of each locus type. Moreover, the general features of locus structure and the average levels of homology with Arabidopsis genes were distinctively different among the locus types. Thus, these findings may indicate an interrelationship among locus structure, transcription activity, and the assignability of gene annotation. An association of transcription activity with locus structure has been reported in plants (37): highly expressed genes have longer primary transcripts, ORFs, and exon and intron sequences and have more exons than weakly expressed genes. The results from a previous report (37) are consistent with our hypothesis that locus structure affects transcription activity, except that the results for intron length differed from our results. The locus length of FL-AE was greater than that of FL-NAE, and the collection efficiency of FL-AE was also greater than that of FL-NAE. Moreover, the cloning efficiency of mRNA-like npRNA in the mouse is lower than that of protein-coding mRNA, and half of the npRNA in the mouse has an efficiency of 1 (36). Cloning efficiency features in the mouse are also similar to those of FL-NAE; this may imply that many FL-NAE loci encode mRNA-like npRNA.

We analyzed the diversity between FL-AE and FL-NAE identified from FL-cDNA mapping and found some differences between FL-AE and FL-NAE. The difference between FL-AE and FL-NAE is similar to that between protein-coding mRNA and mRNA-like npRNA. In our classification, FL-AE implies a protein-coding locus and FL-NAE is an npRNA-coding locus, which might explain why FL-NAE loci are not predicted in TIGR4. Although gene prediction software can identify loci that encode proteins, it does not detect loci that are transcribed into ncRNA. Therefore, prediction software cannot find an FL-NAE locus that encodes npRNA.

2.2 *Gene-family–based annotation of rice genes, such as membrane-transport-protein–coding genes, contributes to study of comparative biology from bacteria to higher plants*

Membrane transport proteins carry various materials for homeostasis. The transport proteins have many clear domain features (e.g., transmembrane, pore-forming, ATP-binding, molecular capture) and functional features. In accordance with their structures and functional systems, membrane transport proteins have been divided into three categories: pump, channel, and secondary transporter. The pump system is the slowest system (1–103 molecules/s) but is environmentally independent and consumes energy (mainly ATP) for transport. The channel system is the most rapid system (107–108 molecules/s) and is non-energy-consuming, but it needs concentration gradients previously (transport directions are only according to the gradients). The secondary transport system adapts the movement energy of co-transport molecules to carry molecules. Therefore, it needs co-transport molecules, and the transport direction depends on the environmental conditions; the speed of this system (102–104 molecules/s) is midway between those of the pump and channel systems. We summarized all three categories (pump, channel, and secondary transporter) of genes and compared the total numbers of membrane transport genes in *Escherichia coli*, *Arabidopsis thaliana*, *Oryza sativa*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Neurospora crassa*, and *Saccharomyces cerevisiae*. The genome sizes among these organisms were diverse (4.6–3150 megabases), and the numbers of transmembrane genes ranged from 300 to 350 in *E. coli*, fungi, and yeast to about 1000–1200 in Arabidopsis and rice. This suggests that a minimum number of about 300 gene species is required to retain cell homeostasis. The greater numbers of transmembrane transport genes in Arabidopsis and rice suggest additional redundancy as well as the modification of genes for new roles (e.g., addition of new substances, adaptation of systems for regulating transport, divergence of stage- and tissue-

specific material transport), specialization for the various tissues and cells of multicellular organisms, and the increased complexity of cells, which in eukaryotes have many additional organelles. The greater relative increase (plant versus bacterium, fungus, and yeast) in the numbers of membrane transport genes was less than has been reported for other gene categories, such as transcription factor genes and metabolic enzyme genes, in higher eukaryotes (38). This suggests that adaptations in membrane transport are critical for the survival of organisms during evolution. The total numbers of membrane transport genes in higher plants (Arabidopsis, about 1000; rice, 1200) are 1.2–2.0 times those in animals (fly, 600; nematode, 650; human, 750). These differences in numbers of transporter genes may be related to differences in the need for efflux and influx systems in restricted habitation environments. Because of their immobility and the simplicity of their uptake systems, plant cells have more opportunity than animals to absorb inappropriate substances; they can also absorb greater amounts of substrates and synthesize larger amounts of secondary products.

We compared the composition ratios of the three classes of protein (pump, channel, and secondary transporter; Fig. 2). The numbers of pump genes in animals (72–82) were almost the same as in bacteria (70). The numbers of secondary transporter (animals, 350; bacteria, 230) and channel (animals, 160–320; bacteria, 15) genes were increased in animals. In particular, vertebrates (humans) had more (322) channel gene species than plants (130–180). We considered that this gene diversity in the development of channel systems was caused by the acquisition of a nervous system. The electrical transmission systems in the nervous systems supplying organs (e.g. muscles, kidneys) need precisely controlled ion concentrations and the ability to make immediate changes in gradients. The development of active transport systems in animals allowed the regulation of rapid movements of the body and organs. Therefore, animals presumably acquired genes for the fastest transport-system channels. Plants also had more channel gene species than bacteria, although fewer than animals. Because plants do not transmit signals for quick movement of their organs, they do not need to regulate membrane voltages as precisely as animals. Additionally, signal-transmitting systems with ligand molecules (e.g. neurotransmitters) are not specific, unlike in animals. Therefore, the numbers of voltage-gated ion channels (VICs) and ligand-dependent channels were smaller in plants than in higher animals (Table 5). On the other hand, higher plants had increased numbers of genes for pumps (170–250) and secondary active transporters (660–760). Plant cells have chloroplasts, which synthesize carbohydrates for many biological activities, including protein synthesis and

functioning of ATP-dependent pumps. Plants presumably use ATP-consuming systems more easily than animals, and the pumps transport the molecules that act as the driving forces of the secondary active transporters. Additionally, plant-specific organelles and vacuoles provide pools of ions and catabolite molecules. Co-transport molecules for secondary transport are also safely and stably stored in the vacuoles. Therefore, plants are presumably able to constantly supply co-transport molecules for secondary active transporters, independently of environmental conditions. The existence of vacuoles gives plant cells more self-sufficiency than animal cells and explains the evolution of membrane transport genes for individual cell homeostasis in plants. Therefore, pump and secondary transporter systems in plants are more divergent than in animals.

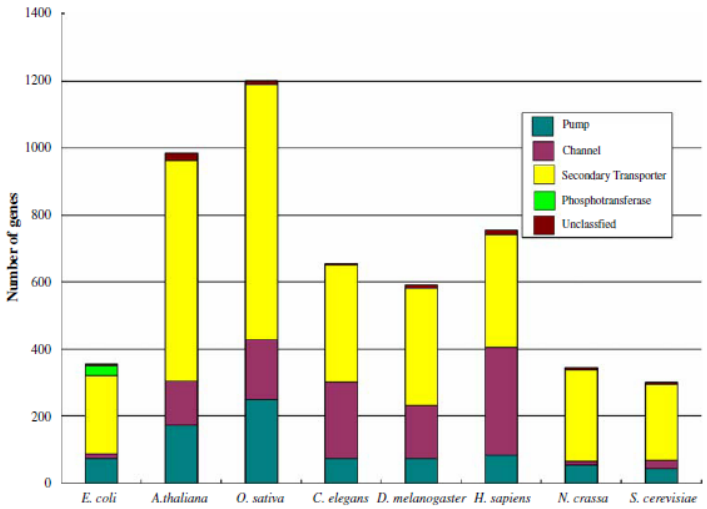

Fig. 2 **Numbers of membrane transporter proteins of each class**. Membrane transporter proteins were categorized into three classes (ATP-dependent [pump], channel, and secondary transporter) and compared among *Escherichia coli* K12-MG1655, *Arabidopsis thaliana*, *Oryza sativa*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens* NCBI, *Neurospora crassa* 74-OR23-IVA, and *Saccharomyces cerevisiae* S228C (Source: Nagata et al. Plant Mol Biol. 2008 Apr;66(6):565–85. Epub 2008 Feb 22.(58))

Table 5 Comparison of genome size and total and membrane transport gene numbers in various organisms

	E. coli K12	*A. thaliana*	*O. sativa*	*C. elegans*	*D. melanogast*	*H. sapiens*	*N. crassa 74*	*S. cerevisiae*
Genome Size (Mb)	4.6	125	430	97	120	3150	40	13
Total gene number	4,290	26,000	32,000	20,621	13,489	30,000	10,082	5,804
Total Transporter Proteins	354	984	1200	654	590	754	344	300
Transporters per Mb genome	76.96	7.87	2.79	6.74	4.92	0.24	8.60	23.08
Transporters per whole gene	0.08	0.04	0.04	0.03	0.04	0.03	0.03	0.05
ATP-dependent	72	173	245	73	91	99	63	70
	20.3%	17.6%	20.4%	11.2%	15.4%	13.1%	18.3%	23.3%
Ion Channels	15	160	144	230	180	353	14	22
	4.2%	16.3%	12.0%	35.2%	30.5%	46.8%	4.1%	7.3%
Phosphotransferase	29	0	0	0	0	0	0	0
	8.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Secondary Transporters	233	635	721	349	359	337	271	223
	65.8%	64.5%	60.1%	53.4%	60.8%	44.7%	78.8%	74.3%
Unclassified	4	20	11	4	9	14	6	6
	1.1%	2.0%	0.9%	0.6%	1.5%	1.9%	1.7%	2.0%

Comparative analysis of membrane transporters among these eight diverse organisms reveals differences in the types of cell homeostasis, as evidenced by the patterns of gene conservation and diversification. Evolutionary changes in gene families, in general, indicate the dynamics of alterations in biological systems and gene networks. Therefore, analysis of large categories of gene families may reveal many basic concepts of biological systems. In practice, analyses of the membrane transporter mechanism are useful in revealing changes in the absorption of molecules by, or their efflux from, cells and tissues. This information also is useful for examining changes in soil adaptability, nutritional demand, and stress tolerance in plants. It may also help to improve the harvest of crop cultivars or extend the areas habitable by plant species. Gene networks are intricately related, and analysis of the whole genetic structure is needed if we are to gain a full understanding of biological phenomena and systems of gene regulation. We are continuing to analyze whole categories of genes in an effort to develop an overview of total gene networks.

2.3 *RiCES: a tool for* cis*-element searches in the promoter regions of clustered genes after microarray analysis*

The tool, called Rice *Cis*-Element Searcher (RiCES), consists of a *cis*-element–searching pipeline controlled via a Web-based user interface. Figure 3 summarizes the procedure. The pipeline first reads a list of gene identifiers from the user and retrieves the promoter sequences corresponding to the listed genes. Then a preliminary list of *cis*-element candidates is built by aligning information

from the built-in list of plausible motifs, or by *ab initio* motif searching of the sequence data. Association rule analysis is carried out and reported to support the candidacy of the resulting *cis*-element list.

Gene list

RiCES assumes that a user has already identified genes of interest from experimental analysis (e.g., clusters of coordinately regulated genes). The list of identifiers is input into a Web-based data entry form. RiCES recognizes GenBank accession numbers, identifiers of TUs as defined in the TIGR pseudomolecular assemblies (18), and several other major gene identification systems. Using the list, it retrieves the set of associated upstream, downstream, or coding region sequences flanking the specified genes from available genomic sequence data.

Preliminary *cis*-element candidate list

The second step of the analysis is the compilation of a list of motifs as candidate *cis*-elements. RiCES supports two methods of achieving this. The first method depends on *ab initio* motif searching based on the supposition that if there are *cis*-elements playing important roles in the regulation of a given set of genes, they will be statistically over-represented in the associated promoter sequences as conserved motifs that can be identified by using a suitable motif search program. There are several programs implementing several algorithms. We have chosen to use MEME, which is a publicly available motif discovery program (39) supporting an expectation maximization algorithm. In our analysis algorithm, MEME is invoked to identify motifs 6 to 8 bp long that look highly conserved among promoter sequences of the selected genes. Users can modify some of the search parameters of the MEME program via the Web form. The second method relies on the hypothesis that common, known *cis*-elements play important roles under the experimental conditions that gave rise to the list of genes specified by the user. Therefore, RiCES searches for matches to a pre-compiled list of known *cis*-elements. Several databases of plant *cis*-elements are publicly available. PLACE (40) is one of the most popular databases of known *cis*-elements in plant genomes. AtcisDB, a part of AGRIS (41), includes information on *cis*-elements involved in gene regulation in *Arabidopsis thaliana*. Although these databases are extremely useful resources, it is not straightforward to cross-link information from them directly to the researcher's own data. Current databases are not exhaustive enough to distinguish 'core'

motifs, which decide the function of *cis*-elements, from co-existing sequences in neighboring regions. As a result, many *cis*-element sequence data in these databases include superficial core motifs for which no evidence of functionality has been obtained. The use of such data prohibits effective information analysis. The *cis*-elements are collected from reports of experiments such as gel shift assays and footprint analyses, categorized by transcription factor, and documented with respect to known activity in the plant genome. Some *cis*-elements known only in organisms other than plants are also listed, in consideration of their possible, albeit unknown, roles in plants. The database includes four types of *cis*-elements: (1) G-box and E-box, which bind to common sequences such as bHLH or bZIP in many organisms; (2) A-box, T-box, and GGTTTAG repeats, which bind to common sequences in many organisms, such as homeodomain and Myb; (3) CArG boxes and GCC-box, which bind to plant MADS, zinc finger, and AP2/EREBP elements; and (4) other *cis*-elements binding only in animals, such as HSF, PcG, and HMG.

Association rule analysis

The third step of the analysis is the likelihood evaluation of the *cis*-element candidates by association rule analysis, which is a data mining method designed to discover significant relationships between pairs of characteristics observed in datasets. Candidates showing the highest likelihood (specificity) are retained in the final *cis*-element candidate list. Association rule analysis has been applied to mechanisms that regulate gene expression (e.g. 42, 43). We used it to find relationships between identified *cis*-elements and gene expression profiles. The strategy depends on the idea that motifs over-represented in the promoter region of the genes of interest could play specific roles in regulation of the expression of those genes. Implied cause-and-effect relationships documented as ‘rules’ are evaluated by using several well-known indices of likelihood, including support, confidence, and lift (42). On the basis of sample datasets, the lift index appeared to best discriminate significant relationships between experimental conditions and *cis*-element candidates. If the presence of motif X in a gene implies that the gene is a member of group Y, then lift is the ratio of the posterior probability (the probability that the gene is in group Y if it possess motif X) to the prior probability (the probability of X possession, irrespective of the membership of Y). When lift > 1.0, coexistence of X and Y is not a random occurrence but suggests some causal relationship between them. If lift < 1.0, it is not considered probabilistically significant. Consequently, we set the default threshold of lift to 1.0, and the *cis*-element candidates are included in the final candidate list only if their lift values are higher than this threshold. RiCES also evaluates pairwise

combinations of motifs in the preliminary candidate list (upper right-hand box in Fig. 3), in consideration of possible protein–protein interactions of multiple transcription elements binding *cis*-elements, as illustrated by experimental evidence (44, 45).

Output

The final *cis*-element candidate list is presented as an association table with the identifier of the submitted genes (TU identifiers based on TIGR gene model annotation are used in the current version) annotated with any available corresponding information from RiceCyc (http://www.gramene.org/pathway/) and Gene Ontology (46). RiCES also provides information on candidate motifs, including the positions of the element in the promoter regions of corresponding TUs, the sequence, and related information from AtcisDB (41). The position of the *cis*-element candidates is also presented in both text and graphics.

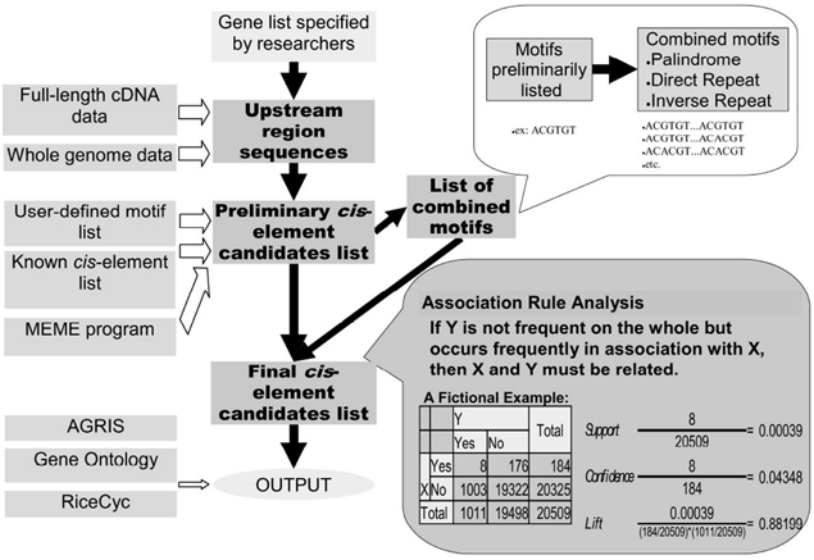

Fig. 3 Features of RiCES. (Source: Doi et al. BMC Plant Biology 2008, 8:20(59))

Validation

To test whether or not the output of RiCES was meaningful, we validated it with a list of auxin-inducible genes with known characteristics, compiled from RiceTFDB 2.0 (http://ricetfdb.bio.uni-potsdam.de/). First, Aux/IAA genes stored in RiceTFDB were applied as queries in a BLASTN search (47) of

GenBank, returning a list containing 28 rice TUs. These genes were fed into the pipeline. When the MEME program was called, the length of target motifs was set to 6, 7, or 8 bases, the number of occurrences of each motif was set to 7, 14, or 21, and the search algorithm was set to 'zoops' to check zero or one occurrence per sequence. The outputs of each option setting were merged but not otherwise filtered.

Many Aux/IAA genes are auxin-inducible (48) and contain the TGTCTC element (49). This element is commonly found in the upstream regions of auxin-responsive genes. Thus, detection of all instances of the motif by the pipeline could serve as a validation of the pipeline algorithm. The auxin-responsive element (AuxRE) containing the TGTCTC motif in some cases requires another proximal AuxRE for biological activity (44, 50). In other contexts, AuxRE functions only when it occurs with its palindromic components separated by 7 or 8 nucleotides (51). A search of AtcisDB for these motifs returned 4 showing a partial match to the record of 'PRHA binding sites', which is derived from the report of Plesch et al. (52), describing auxin-induced expression of the Arabidopsis *prha* homeobox gene. Another 4 motifs contained the TGTCTC element. The result was consistent with results of previous work, as TGTCTC was listed as a candidate in the single-motif search of Aux/IAA genes. The analysis returned 22 *cis*-element candidates with lift > 1.0. Some of these candidates were suggested by previous studies to have some kind of relationship to auxin response. For example, RAV1 was found in the promoter region of ABP, which encodes an auxin-binding protein (53). Expression of LEAFY (LFY) is affected by the auxin gradient in Arabidopsis (54). ETT is another auxin response factor (55), and LFY and ETT expression are closely correlated (45, 56). The position of a *cis*-element is important information to consider in relation to the function of the *cis*-element. For biological activity to occur, the distance of some *cis*-elements from the coding region or other collaborating elements is constrained. To this end, RiCES highlights the distribution of *cis*-element candidates. It provides tables of identified *cis*-element motifs and graphical motif maps to help researchers grasp positional relationships among the candidate elements.

The positions of the listed elements, some of which include TGTCTC, varied among upstream regions of genes, and it was hard to detect any skewed distribution of motifs. Goda et al. (57) studied the distribution of TGTCTC motifs in the genome of *A. thaliana*, and they pointed out that 25% of genes investigated had TGTCTC motifs in the upstream region within 1000 bp of the start codon, and 14% within 500 bps. Our results do not seem to conflict with theirs. TGTCTC motifs are scattered over wide regions of many plant species. It

is possible that the variety of the roles of genes reflects the variety of mechanisms regulating gene expression and positions of *cis*-elements, even if the genes in question can be classified as 'auxin-responsive genes' in a larger sense. A major research concern is how to pick up *cis*-element candidates worthy of further experimentation. Computational and manual selection of *cis*-element candidates should play complementary roles to resolve this issue. It should be emphasized that *cis*-element candidates listed by RiCES are rated according to the likelihood provided by association rule analysis. On the other hand, researchers can check the significance of candidates in detail by using related information derived from several databases. The supported databases include AGRIS, Gene Ontology, and RiceCyc, as well as the map information described above. The outputs are not only easily accessible in a Web browser, but are also usable in further statistical or bioinformatics analyses, as they are also provided in XML format, which is a tagged plain-text format compatible with various computer programs. In some cases, the results of the analysis from the pre-compiled list of elements will be easily comparable with prior knowledge. In other cases involving solely *ab initio* evidence from MEME, the results of motif searches should be interpreted carefully, because the result will change considerably in accordance with the options selected. An appropriate set of motif search options should be determined each time, by trial and error. However, as described above, a motif search can find *cis*-element candidates whose sequences do not exactly match those of known *cis*-elements. Although RiCES is focused on the role of *cis*-elements in *Oryza sativa* ssp. *japonica*, the methodology can be applied easily to studies of other plant species, or of other genome sequence motifs involving gene expression regulation, such as motifs in coding regions of genes or downstream of the gene sequence. Such work can be made possible by replacing the reference dataset containing whole genes of rice with other datasets.

We have presented here a newly developed tool for searching for *cis*-element candidates in lists of genes. A case study showed the applicability of the tool. The tool is easy to use and publicly available. We expect that its use will deepen our understanding of the mechanisms that regulate gene expression in plants.

Contributions

This presentation is based on three recently published reports from our laboratory (6, 58, 59). I sincerely thank all the authors of the reports.

References

1. S.A. Goff, D. Ricke, T.H. Lan, G. Presting, R. Wang, M. Dunn, et al., *Science* **296**, 92 (2002).
2. J.Yu, S. Hu, J. Wang, G.K. Wong, S. Li, et al., *Science* **296**, 79 (2002).
3. J.Yu, J. Wang, W. Lin, S. Li, H. Li, et al. *PLoS Biol*. Feb, 3e38 (2005).
4. International Rice Genome Sequencing Project, *Nature* **436**, 793 (2005).
5. Rice Full-Length cDNA Consortium, *Science* **301**, 376 (2003).
6. K. Satoh, K. Doi, T. Nagata, N. Kishimoto, K. Suzuki, et al., *PLoS ONE* **2**, e1235 (2007).
7. J. Wu, T. Maehara, T. Shimokawa, S. Yamamoto, C. Harada, et al., *Plant Cell* **14**, 525 (2002).
8. Y. Zhou, J. Tang, M.G. Walker, X. Zhang, J. Wang, et al., Genomics *Proteomics Bioinformatics* **1**, 26 (2003).
9. J. Zhang, Q. Feng, C. Jin, D. Qiu, L. Zhang, et al., *Plant J* **42**, 772 (2005)
10. M. Seki, M. Narusaka, A. Kamiya, J. Ishida, M. Satou, et al., *Science* **296**, 141 (2002).
11. T. Imanishi, T Itoh, Y Suzuki, C O'Donovan, S Fukuchi, et al., *PLoS Biol* **2** (2004).
12. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, *Nature* **409**, 685 (2001).
13. T. Itoh, T. Tanaka, R. A. Barrero, C. Yamasaki et al., *Genome Research* **17**, 175 (2007).
14. Arabidopsis Genome Initiative, *Nature* **408**,796 (2000).
15. H. Schoof H, R. Ernst, V. Nazarov, L. Pfeifer, H.W. Mewes, K. and F. Mayer, *Nucleic Acids Res.* **32** (database issue), D373 (2004).
16. W. M. Karlowski, H. Schoof, V. Janakiraman, V. Stuempflen and K. F. Mayer, *Nucleic Acids Res.* **31**, 190 (2003).
17. T. Tanaka, B. A. Antonio, S. Kikuchi, T. Matsumoto et al., *Nucleic Acids Res.* **36** (database issue), D1028 (2008).
18. S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee et al., *Nucleic Acids Res*. **35**, D883 (2007)
19. B. H. Eng, M. L. Guerinot, D. Eide and M. H. Jr. Saier, *J. Membr. Biol*. **166**, 1 (1998)
20. S. S. Pao, I. T. Paulsen and M. H. Jr. Saier, *Microbiol. Mol. Biol. Rev.* **62**, 1 (1999).
21. P. Mäser, S. Thomine, J. I. Schroeder, J. M. Ward, K. Hirschi, H. Sze, I. N. Talke, A. Amtmann, F. J. Maathuis, D. Sanders, J. F. Harper, J. Tchieu, M. Gribskov, M. W. Persans, D. E. Salt, S. A. Kim and M. L. Guerinot, *Plant Physiol*. **126**, 1646 (2001).
22. R. Sánchez-Fernández, T. G. Davies and J. O. Coleman, *J. Biol. Chem.* **276**, 30231 (2001).
23. Q. Ren and I. T. Paulsen, *PLoS Comput. Biol.* **3**, e27 (2005).

24. H. M. Wain, M. J. Lush, F. Ducluzeau, V. K. Khodiyar and S. Povey, *Nucleic Acids Res.* **32** (database issue), D255 (2004).
25. Q. Ren, K. H. Kang and I. T. Paulsen, *Nucleic Acids Res* **32** (database issue), D284 (2004).
26. J. H. Tchieu, F. Fana, J. L. Fink, J. Harper, T. M. Nair, R. H. Niedner, D. W. Smith, K. Steube, T. M. Tam, S. Veretnik, D. Wang, M. Gribskov, *Nucleic Acids Res.* **31**, 342 (2003).
27. R. Schwacke, A. Schneider, E. van der Graaff, K. Fischer, E. Catoni, M. Desimone, W. B. Frommer, U. I. Flügge and R. Kunze, *Plant Physiol.* **131**, 16 (2003).
28. M. G. Carter, T. Hamatani, A. A. Sharov, C. E. Carmack, Y. Qian, K. Aiba, N. T. Ko, D. B. Dudekula, P. M. Brzoska, S. S. Hwang and M. S. H. Ko, *Genome Res*. **13**, 1011 (2003).
29. T. R. Hughes, M. Mao, A. R. Jones, J. Burchard, et al. *Nat. Biotechnol*. **19**, 342 (2001).
30. D. D. Shoemaker, E. E. Schadt, C. D. Armour, Y. D. He, P. Garrett-Engele, P. D. McDonagh, P. M. Loerch, and A. Leonardson. *Nature* **409**, 922 (2001).
31. J. Yazaki, Z. Shimatani, A. Hashimoto, Y. Nagata, F. Fujii et al. *Physiol. Genomics*. **17**, 87 (2004).
32. J. van Helden, *Nucleic Acids Res*. **31**, 3593 (2003).
33. K. E. Holt, A. H. Millar and J. Whelan, *Plant Methods* 2, 8 (2006).
34 F. Zhao, Z. Xuan, L. Liu, M. Q. Zhang. *Nucleic Acids Res.* **33**, D103 (2005).
35. S.Y. Rhee, W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, et al., *Nucleic Acids Res.* **31**, 224 (2003).
36. T. Ravasi, H. Suzuki, K. C. Pang, S. Katayama, M. Furuno, et al., *Genome Res*. **16**, 11 (2006).
37. X. Y. Ren, O. Vorst, M. W. Fiers, W. J. Stiekema, J. P. Nap, *Trends Genet* **22**, 528. (2006).
38. G. A. Wray , M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, L. A. Romano. *Mol. Biol. Evol.* **20**, 1377 (2003).
39. T. L. Bailey and C. Elkan *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28 (1994).
40. K. Higo, Y. Ugawa, M. Iwamoto and T. Korenaga *Nucleic Acids Res*. **27**, 297 (1999).
41. R. V. Davuluri, H. Sun, S. K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz and E. Grotewold, *BMC Bioinformatics* **4**, 4:25 (2003).
42. P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. M. Carazo and A. Pascual-Montano, *BMC Bioinformatics* **7**, 54 (2006).
43. D. Conklin, I. Jonassen, R. Aasland and W. R. Taylor, *Bioinformatics* **18**, 182 (2002).
44. T. Ulmasov, Z. B. Liu, G. Hagen and T. J. Guilfoyle, *Plant Cell* **7**, 1611 (1995).
45. T. Ulmasov, G. Hagen and T. J. Guilfoyle *Plant J.* **19**, 309 (1999).

46. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al. *Nat. Genet.* **25**, 25 (2000).
47. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
48. J. W. Reed *Trends Plant Sci.* **6**, 420 (2001).
49. S. B. Tiwari, X. J. Wang, G. Hagen and T. J. Guilfoyle, *Plant Cell* **13**, 2809 (2001).
50. Z. B. Liu, G. Hagen and T. J. Guilfoyle, *Plant Physiol.* **115**, 397 (1997).
51. T. Ulmasov, G. Hagen, T. J. Guilfoyle, *Science* 276, 1865 (1997).
52. G. Plesch, K. Stoermann J. T. Torres, R. Walden, I E. Somssich, *Plant J.* **12**, 635 (1997).
53. Y. Kagaya, K. Ohmiya and T. Hattori, *Nucleic Acids Res*. **27**, 470 (1999).
54. T. A. Ezhova, O. P. Soldatova, A. Iu. Kalinina, S. S. Medvedev, *Genetika* **36**, 1682 (2000).
55. A. Sessions, J. L. Nemhauser, A. McColl, J. L. Roe, K. A. Feldmann and P. C. Zambryski, *Development* **124**, 4481 (1997).
56. D. L. Remington, T. J. Vision, T. J. Guilfoyle and J. W. Reed, *Plant Physiol.* 135, 1738. (2004).
57. H. Goda, S. Sawa, T. Asami, S. Fujioka, Y. Shimada and S. Yoshida, *Plant Physiol*. **134**, 1555 (2004).
58. T. Nagata, S. Iizumi, K. Satoh and S. Kikuchi, *Plant Mol. Biol.* **66**, 565 (2008).
59. K. Doi, A. Hosaka, T. Nagata, K. Satoh, K. Suzuki, R. Mauleon, M. J. Mendoza, R. Bruskiewich, S. Kikuchi. *BMC Plant Biol.* Feb 27 **8**, 20 (2008).