main gonadal (steroid) pheromone released by a 25-g female goldfish (*33*). Interference with this pheromone system offers an attractive target for selective and environmentally benign control of the sea lamprey, whose invasion of the Great Lakes represents arguably the worst ecological disaster ever to befall a large watershed (*34*).

References and Notes

1. V. Applegate, *U.S. Fish Wildl. Serv. Spec. Sci. Rep.* **55**, 237 (1950).
2. P. Manion, L. Hanson, *Can. J. Fish. Aquat. Sci.* **37**, 1635 (1980).
3. M. Fontaine, *Bull. Soc. Oceanogr. Fr.* **17**, 1681 (1939).
4. J. Teeter, *Can. J. Fish. Aquat. Sci.* **37**, 2123 (1980).
5. R. Carde, A. Minks, *Annu. Rev. Entomol.* **40**, 559 (1995).
6. M. V. Novotny, B. Jemiono, S. Harvey, D. Wiesler, A. Marchlewska-Koj, *Science* **231**, 722 (1986).
7. R. T. Mason *et al.*, *Science* **245**, 290 (1989).
8. L. E. L. Rasmussen, T. D. Lee, W. L. Roelofs, A. Zhang, G. D. Daves, *Nature* **379**, 684 (1996).
9. J. Dulka, N. E. Stacey, P. W. Sorensen, G. Van der Kraak, *Nature* **325**, 251 (1987).
10. P. W. Sorensen, T. Hara, N. E. Stacey, F. Goetz, *Biol. Reprod.* **39**, 1039 (1988).
11. T. Sveinsson, T. J. Hara, *Environ. Biol. Fishes* **42**, 253 (1995).
12. H. Yambe, M. Shindo, F. Yamazaki, *J. Fish Biol.* **55**, 158 (1999).
13. Animals were classified as spermiating males and ovulated females if milt and eggs, respectively, could be expressed by manual pressure (or otherwise as prespermiating males and preovulatory females) and used as either test subjects or odorant donors in a flow-through (0.07 m s^{-1}) maze (4.6 m by 1.2 m) with a plywood bottom and sides and a partition in the middle that extended 2.4 m from the upstream end, and with plastic meshes blocking fish movement at both the upstream and downstream ends. Odorant donors were held above the upstream mesh. Between 0700 and 1700 hours, a single test subject was acclimated for 10 min in the maze, and its behavior was video recorded for 20 min. Then, five lampreys (all of one sex and maturity) were introduced into the mesh chamber on a randomly chosen side, and the behavior of the test subject was recorded for another 20 min. When washings were used, a spermiating male was held in 10 liters of water for 4 hours, and the water was introduced into the odor chamber at 75 ml min^{-1}. Naïve observers scored videotapes for the total time spent in the experimental and control sides before odorant introduction (Be and Bc) and the time spent in both sides after odorant introduction (Ae and Ac). To measure the attraction of test subjects to the conditioned side of the maze, the scores were used to calculate an index of preference (I) = [Ae/(Ae + Be) – Ac/(Ac + Bc)]. A similar index was computed for search behaviors that involved pacing back and forth across the upstream barrier, increased swimming speed, and rapid beating of the tail by the test subject.
14. M. Siefkes, S. Yun, A. Scott, W. Li, data not shown.
15. J. R. M. Kelso, W. M. Gardner, *N. Am. J. Fish. Manage.* **20**, 132 (2000).
16. A field study was conducted in a 65-m segment of the Ocqueoc River, Presque Isle County, Michigan, a tributary to Lake Huron, with a barrier to prevent lamprey migration from the lake. The average discharge was 2.3 m^3 s^{-1}. Upstream, an island divided the streams into two channels. Cages (1 m^3) of plastic mesh (~1.5 cm mesh size) containing five male lampreys (spermiating or prespermiating) were randomly placed in the two channels. A female fitted with an external radio transmitter (*14*) was acclimated in a cage for 2 hours, released 65 m downstream, and its location was recorded every 5 min. Tests were conducted between 0700 and 1700 hours in water temperatures ranging from 12°C to 24°C.
17. A lamprey was placed in 10 liters of aerated water for 4 hours and then removed. The water was drawn through a filter paper (Whatman No. 3) and then SPE cartridges (Sep-Pak; Waters Chromatography, Millipore, Milford, MA; prewashed with 5 ml of methanol, followed by 5 ml of distilled water) at a rate of up to 20 ml min^{-1}. One liter was pumped through each cartridge, which was then washed with 5 ml of distilled water and eluted with 5 ml of methanol.
18. W. Li, thesis, University of Minnesota (1994).
19. Samples were loaded in 50 μl of ethanol on silica gel plates (Whatman type LK6DF), which were developed with chloroform/methanol (50/6, v/v) for 45 min, sprayed with 5% PBA in methanol, placed on a hot plate at 100°C for 3 to 5 min to develop the color, and photocopied.
20. A. P. Scott *et al.*, *Gen. Comp. Endocrinol.* **105**, 62 (1997).
21. W. Li, P. W. Sorensen, D. Gallaher, *J. Gen. Physiol.* **105**, 567 (1995).
22. The samples were dissolved in perdeuterated methanol or dimethyl sulfoxide and subjected to a Varian INOVA 600 spectrometer at 25°C for 2D homonuclear ^{1}H COSY and TOCSY spectra and heteronuclear ^{1}H-^{13}C HSQC, HSQC-TOCSY, and HMBC spectra. The 1D ^{13}C spectrum was acquired on a Varian VXR 500 spectrometer. Standard pulse sequences were used. Suitable window functions were applied to the time domain data for resolution or sensitivity enhancement prior to Fourier transformation. Both ^{1}H and ^{13}C chemical shifts were referenced to the solvent resonances.
23. S. Barnes, N. D. Kirk, in *The Bile Acids: Chemistry, Physiology, and Metabolism*, K. D. R. Setchell, D. Kritchevsky, P. P. Nair, Eds. (Plenum Press, New York, 1988), vol. 4, pp. 65–136.
24. G. Haslewood, L. Tokes, *Biochem. J.* **114**, 179 (1969).
25. R. Bjerselius *et al.*, *Can. J. Fish. Aquat. Sci.* **57**, 557 (2000).
26. The following mixture was shaken at 37°C for 5 hours: 10 mg of petromyzonol sulfate in 1 ml methanol, 40 mg of β-nicotinamide adenine dinucleotide (NAD) in 50 ml 0.05M 3-(Cyclohexylamino)-1-propanesulfonic acid buffer at pH 10.8, and 10 units of 3α-hydroxysteroid dehydrogenase in 100 μl of 0.1 M sodium phosphate buffer at pH 7.6. After 1 hour, 20 mg NAD and 10 units of enzyme were added. The products of the reaction were extracted with SPE cartridges (*16*) and purified by HPLC (*19*).
27. E. L. M. Vermeirssen, A. P. Scott, *Gen. Comp. Endocrinol.* **101**, 180 (1996).
28. K. Yamamoto, P. A. Sargent, M. M. Fisher, J. H. Youson, *Hepatology.* **6**, 54 (1986).
29. T. H. Maren, R. Embry, L. E. Broder, *Comp. Biochem. Physiol.* **26**, 853 (1968).
30. A. D. Pickering, *Cell Tissue Res.* **180**, 1 (1977).
31. P. W. Sorensen, N. E. Stacey, in *Advances in Chemical Signals in Vertebrates*, R. E. Johnston, D. Müller-Schwarze, P. W. Sorensen, Eds. (Kluwer Academic, New York, 1999), pp. 15–48.
32. R. Bjerselius, W. Li, P. W. Sorensen, A. P. Scott, *Proceedings of the Fifth International Symposium on the Reproductive Physiology of Fish*, P. Thomas, F. Goetz, Eds. (The University of Texas at Austin Press, Austin, 1995), p. 271.
33. A. P. Scott, P. W. Sorensen, *Gen. Comp. Endocrinol.* **96**, 309 (1994).
34. B. R. Smith, J. J. Tibbles, *Can. J. Fish. Aquat. Sci.* **37**, 1780 (1980).

35. We thank R. Bergstedt, of the U.S. Geological Survey, Lake Huron Biological Station, for space for and advice on the behavioral experiments; D. Gallaher for advice on the conversion of PS into the male pheromone; J. Kelso for advice on radio-telemetry tracking of sea lamprey; M. Twohey and R. MacDonald for supplying lampreys for this study; B. Chamberlin for assistance with the mass spectrometry analysis; and D. Trump and L. Lorenz for use of their private land as a field study site. Financed by the Great Lakes Fishery Commission.

17 December 2001; accepted 6 March 2002

Functional Annotation of a Full-Length *Arabidopsis* cDNA Collection

Motoaki Seki,[1,2] Mari Narusaka,[1] Asako Kamiya,[1] Junko Ishida,[1] Masakazu Satou,[1] Tetsuya Sakurai,[1] Maiko Nakajima,[1] Akiko Enju,[1] Kenji Akiyama,[1] Youko Oono,[2,3] Masami Muramatsu,[4,5] Yoshihide Hayashizaki,[4,5] Jun Kawai,[4,5] Piero Carninci,[4,5] Masayoshi Itoh,[4,5] Yoshiyuki Ishii,[4,5] Takahiro Arakawa,[4,5] Kazuhiro Shibata,[4,5] Akira Shinagawa,[4,5] Kazuo Shinozaki[1,2]*

Full-length complementary DNAs (cDNAs) are essential for the correct annotation of genomic sequences and for the functional analysis of genes and their products. We isolated 155,144 RIKEN *Arabidopsis* full-length (RAFL) cDNA clones. The 3′-end expressed sequence tags (ESTs) of 155,144 RAFL cDNAs were clustered into 14,668 nonredundant cDNA groups, about 60% of predicted genes. We also obtained 5′ ESTs from 14,034 nonredundant cDNA groups and constructed a promoter database. The sequence database of the RAFL cDNAs is useful for promoter analysis and correct annotation of predicted transcription units and gene products. Furthermore, the full-length cDNAs are useful resources for analyses of the expression profiles, functions, and structures of plant proteins.

Arabidopsis thaliana has been adopted as a model organism in the study of plant biology because of its small size, short generation time, and high efficiency of transformation (*1*). To sequence its small genome [125 megabases (Mb)] (*2*), scientists in Japan, Europe, and the United States collaborated in the *Arabidopsis* genome sequencing project (*3*). Two of five chromosomes (chromosomes 2 and 4, except for the nucleolar organizer regions and centromeres) were sequenced in 1999 (*4*, *5*), and the remaining three

chromosomes were sequenced in 2000 (*2*).

About 127,000 expressed sequence tags (ESTs) from *Arabidopsis* had been deposited in the EST database (dbEST) as of May 2001, including sequences from large-scale EST projects promoted by laboratory consortia in France (*6*, *7*), the United States (*8*, *9*), and Japan (*10*). These projects have produced EST data from different tissues, organs, seeds, and developmental stages (*6*–*10*). However, these EST projects are based on cDNA libraries in which most of the inserts are not full-length. ESTs are useful for making a catalog of expressed genes, but not for further study of gene function. Consequently, genome-scale collections of the full-length cDNAs of expressed genes become important for the analysis of the structure and function of genes and their products in the functional genomics era.

We previously made full-length cDNA libraries using the biotinylated CAP trapper method (*11*, *12*) from *Arabidopsis* plants (*13*). Here, we constructed *Arabidopsis* full-length cDNA libraries from plants grown under different conditions as reported previously (*11*–*15*) by the biotinylated CAP trapper method using trehalose-thermoactivated reverse transcriptase. We used λZAP (*11*, *13*) and λFLC (*16*) vectors for construction of the cDNA libraries. The λFLC vectors accommodate cDNAs in a broad range of sizes and are useful for the high-efficiency cloning of long cDNA fragments (*16*). The λFLC vectors can also be bulk-excised by a Cre-*lox*–based system free of size bias to produce the plasmid libraries. In the construction of full-length cDNA libraries [RIKEN *Arabidopsis* full-length (RAFL) 12, 13, 14, 15, 16, 17, 18, 19, and 21 (Table 1)], we used a single-strand linker ligation method (*17*), which uses DNA ligase to add a double-stranded (ds) DNA linker to single-stranded (ss) full-length cDNA. Subsequent sequencing

[1]Plant Mutation Exploration Team, Plant Functional Genomics Research Group, RIKEN Genomic Sciences Center (GSC), 3-1-1 Koyadai, Tsukuba 305-0074, Japan. [2]Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba 305-0074, Japan. [3]Master's Program in Biosystem Studies, University of Tsukuba, Tennoudai, Tsukuba, Ibaraki, 305-0074, Japan. [4]Laboratory for Genome Exploration Research Group, RIKEN GSC, Yokohama Institute 1-7-22 Suehiro-cho, Tsurumi-ku, Yokokama, Kanagawa, 230-0045, Japan. [5]Genome Science Laboratory, RIKEN, 2-1 Hirosawa, Wako, 351-0198, Japan.

*To whom correspondence should be addressed. Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba 305-0074, Japan. E-mail: sinozaki@rtc.riken.go.jp

Table 1. Summary of 3'-end single-pass sequencing of RAFL cDNA clones isolated from *A. thaliana* full-length cDNA libraries. 155,144 RAFL cDNA clones were clustered by mapping of the 3'-end single-pass-sequencing data on the genomic sequence to produce more than 14,668 cDNA groups. n.d., not determined; UV, ultraviolet; ABA, abscisic acid; JA, jasmonic acid; SA, salicylic acid; GA, gibberellin; BTH, benzo-(1,2,3)-thio-diazole-7-carbothionic acid *S*-methyl ester.

Library no.	Plant materials	Vector	Standard/ normalization/ subtraction	Number of cDNA clones subjected to clustering	Number of cDNA groups
RAFL1	Cold-treated leaves and stems	λZap	Standard*	111†	n.d.
RAFL2	Rosette plants	λZap	Standard	256	130
RAFL3	Dehydration-treated plants	λZap	Standard	223	115
RAFL4	Cold-treated plants	λZap	Standard	1,029	862
RAFL5	Dehydration-treated plants	λZap	Standard	2,030	1,672
RAFL6	Plants at various developmental stages and those treated with dehydration and cold	λZap	Standard	6,139	1,461
RAFL7	Cold-treated plants	λFLC-1-B‡	Standard	2,591	751
RAFL8	Dehydration-treated plants	λFLC-1-B	Standard	2,637	584
RAFL9	Plants at various developmental stages and those treated with dehydration and cold	λFLC-1-B	Standard	22,929	3,368
RAFL11	Plants at various developmental stages and those subjected to various stress (dry, cold, NaCl, heat, and UV) and ABA treatments. Plants grown under dark conditions. Silique tissues	λFLC-1-B	Normalization§	2,242	339
RAFL12	Cold-treated plants	λFLC-1-E‡	Subtraction§	22	2
RAFL13	Dehydration-treated plants	λFLC-1-E	Subtraction	72	5
RAFL14	Roots	λFLC-1-E	Standard	23,302	1,371
RAFL15	Siliques and flowers	λFLC-1-E	Standard	13,661	816
RAFL16	Dark-grown plants	λFLC-1-E	Standard	25,466	1,227
RAFL17	Dehydration-treated plants Rehydration (after dry 10 hours)–treated plants	λFLC-1-E	Subtraction	14,035	452
RAFL18	Cold-treated plants	λFLC-1-E	Subtraction	1,213	41
RAFL19	Siliques and flowers	λFLC-1-E	Subtraction	24,951	970
RAFL21	Plants treated with various stress (heat and UV), hormone (ABA, auxin, ethylene, JA, SA, GA, and cytokinin), and BTH treatments	λFLC-1-E	Subtraction	12,346	502

*cDNAs were neither normalized nor subtracted in the construction of standard full-length cDNA libraries. †These RAFL cDNAs were not used for clustering, because only 5'-end single-pass sequencing had been done on these clones. ‡The information on λFLC-1-B and λFLC-1-E vectors was described previously (*16*). §cDNAs were normalized or subtracted in the construction of normalized or subtracted full-length cDNA libraries as described previously (*18*).

of clones and translation of proteins from full-length cDNA are easier and more efficient because of the elimination of the GC tail. Normalization and subtraction procedures (*18*) were also introduced in the construction of full-length cDNA libraries [RAFL11, 12, 13, 17, 18, 19, and 21 (Table 1)] to reduce the representation of highly expressed mRNAs in the library and to remove cDNAs already categorized by means of one-pass sequencing, respectively. The method is based on hybridization of the first-strand full-length cDNA with several RNA drivers, including starting mRNA as the normalizing driver and run-off transcripts from rearrayed clones as subtracting drivers. This method should dramatically enhance the discovery of new cDNAs. The overall strategy for preparing cDNA libraries, including standard, normalized, and subtracted libraries, has been described previously (*19*). We constructed 19 full-length cDNA libraries from *Arabidopsis* plants grown under various stress, hormone, and light conditions from plants at various developmental stages and from various plant tissues.

We performed single-pass sequencing of the cDNA clones from the 3′ end. The 155,144 3′ ESTs were clustered and then mapped onto the *Arabidopsis* genome (Fig. 1 and supplemental text) (*15*). Finally, 14,668 nonredundant RAFL cDNA clones were identified and mapped on the *Arabidopsis* genome (Table 1 and Fig. 1). The information on the 14,668 RAFL cDNA clones (the "RAFL cDNA" genes) is available in Web tables 1 and 2 (*20*). Assuming that the total number of *Arabidopsis* genes is about 25,000, the RAFL clones should account for about 60% of all *Arabidopsis* genes. Our evaluation of 349 RAFL cDNA clones by single-pass sequencing showed that ~98% of the clones contained both start and stop codons. Thus, the cDNA libraries constructed by the biotinylated CAP trapper contained a very high proportion of full-length cDNAs.

From the 5′-end sequences of mRNAs, the promoter sequences can be obtained by comparison with the *Arabidopsis* genomic sequences. We also obtained 5′ ESTs of 14,034 RAFL cDNA clones and constructed a promoter database (*21*) using the PLACE database (*22*). The *Arabidopsis* promoter database shows genomic sequences 1000 base pairs (bp) upstream from the 5′ termini of each RAFL cDNA clone and about 300 cis-acting elements known from plants (Web table 1) (*20*).

Of the 14,668 RAFL cDNA clones mapped onto the *Arabidopsis* genome, 13,831 were matched to Munich Information Center for Protein Sequences (MIPS) protein entry codes (Fig. 2), leaving 837 RAFL cDNA clones unmatched (Fig. 2, Web fig.

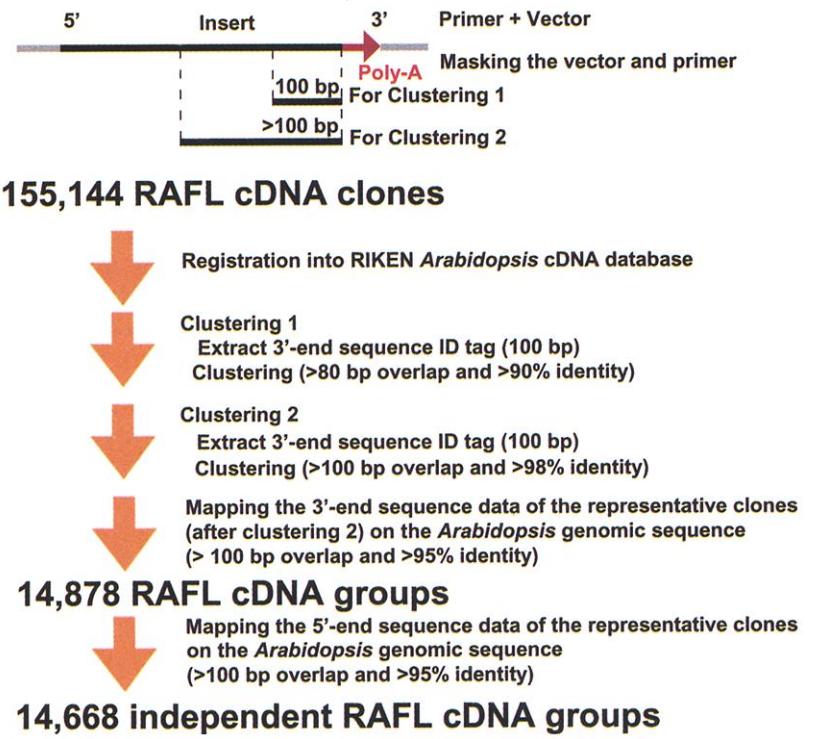

Fig. 1. Strategy for clustering of the RAFL cDNA clones. A total of 155,144 RAFL cDNA clones isolated from 19 full-length cDNA libraries were subjected to single-pass sequencing from the 3′ ends of the cDNA. The 3′-end single-pass sequencing data were used in the two steps for clustering as described in supplemental methods (*15*). After the second clustering, the best quality sequence was chosen as the representative of the group. The 3′ EST of each representative clone was then mapped onto the *Arabidopsis* genome as described in the supplemental text (*15*). As a result, 14,878 nonredundant representative 3′ ESTs were mapped on the *Arabidopsis* genome. Next, the 14,878 cDNA clones were subjected to single-pass sequencing from the 5′ end of the cDNA. The 5′ end sequencing data were then mapped onto the *Arabidopsis* genome with the BlastN program (*15*). Finally, the 14,668 nonredundant RAFL cDNA clones mapped on the *Arabidopsis* genome were identified.

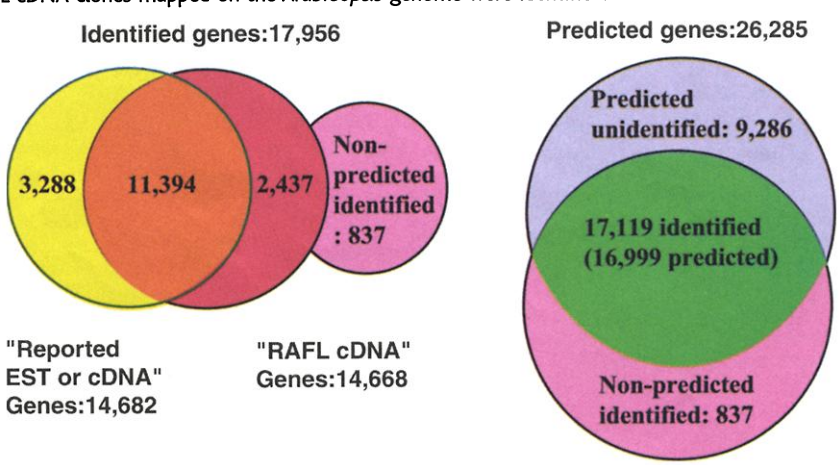

Fig. 2. Current compilation of expressed genes in *Arabidopsis*. The left-hand Venn diagram shows the two classes of the 17,956 experimentally identified genes. Of these genes, there are 14,668 RAFL cDNA genes isolated in this study (red and pink circles) and 14,682 reported EST or cDNA genes (yellow circle), including EST genes identified by EST analysis, CERES cDNAs, and *Arabidopsis* expressed genes that *Arabidopsis* researchers have cloned and sequenced by traditional cloning. Of 14,668 RAFL cDNA genes, 837 newly identified genes that were not predicted and 2437 newly identified genes that were predicted existed. The right-hand Venn diagram shows the intersection between the total number of predicted genes (26,285, blue circle) and the experimentally identified genes (17,956, pink circle). The green region of intersection shows the 17,119 experimentally identified genes that have been predicted. The blue region of nonintersection shows the 9286 predicted genes that have not been experimentally confirmed yet. The pink region of nonintersection shows the 837 identified genes that are not predicted by AGI. In addition, in some cases, pairs of seemingly separate predicted genes correspond to a single experimentally identified gene. Conversely, single predictions sometimes correspond to more than two experimentally identified genes. The last two facts explain why 17,119 genes correspond to 16,999 predicted genes.

1C, and Web table 3-3) (*20*). These 837 RAFL cDNAs have not yet been predicted by the Arabidopsis Genome Initiative (AGI) and thus represent false negatives in the genome annotation.

To analyze all known expressed *Arabidopsis* genes, we used data from: (i) 5100 complete cDNAs that *Arabidopsis* researchers have sequenced and deposited in GenBank as of 18 August 2001 (*23*), (ii) 127,031 *Arabidopsis* ESTs identified as of 22 May 2001 (*24*), and (iii) 5000 *Arabidopsis* full-length cDNAs that Ceres, Inc., released to The Institute for Genomic Research on 19 December 2000 (*25*). Altogether, these genes (the "reported EST or cDNA" genes) were subjected to homology search (*26*) against the sequence database of its corresponding MIPS protein entry code using the BlastN program. The reported EST or cDNA genes covered a total of 14,551 MIPS protein entry codes (Fig. 2). Also, 2437 of the RAFL cDNAs mapped to the MIPS protein entry codes were novel genes not identified so far (Fig. 2). ESTs or cDNA genes have been reported for 3288 MIPS protein entry codes, but no RAFL cDNA genes have been identified (Fig. 2). A total of 11,394 genes corresponded to both reported EST or cDNA and RAFL cDNA genes. These results bring the total number of *Arabidopsis* genes whose expression has been experimentally confirmed to 17,956 (Fig. 2). In comparison, AGI lists 17,119 experimentally confirmed genes, of which 16,999 were predicted (Fig. 2). The discrepancies are likely due to two predicted genes corresponding to a single experimentally identified gene (Web fig. 1A) (*20*), or single predicted genes corresponding to more than two experimentally identified genes (Web fig. 1B) (*20*). Some RAFL cDNA clones correspond to each of these circumstances (Web tables 3-1 and 3-2) (*20*).

We conclude that 9286 predicted genes need further data to be confirmed as expressed genes or unidentified genes (Fig. 2). Because these unidentified genes have not been confirmed by any ESTs, some of the predicted genes represent false positives or pseudogenes. Alternatively, these unidentified genes might have remained undetected by the EST approach because of their weak expression in specific tissues.

The biological roles and biochemical functions of RAFL cDNA clones were identified by homology search using the BLAST program (Table 2). The results show that cDNA clones of some functional categories, such as energy production, protein synthesis, and ion homeostasis are well represented in RAFL. More than 80% of cDNAs for genes involved in energy production, protein synthesis, and ionic homeostasis were found in RAFL, and ~70% of cDNAs for genes involved in metabolism, protein destination, cellular transport and transport mechanisms, and cellular organization were found in RAFL. It has been estimated that ~1500 transcription factor genes (*27*) and about 1000 protein kinase genes (*28*) exist in the *Arabidopsis* genome. The RAFL cDNA collection includes 1087 transcription factor and 506 protein kinase genes (Table 2).

Although many algorithms have been written to predict a transcription unit from genomic sequence data, the accuracy of their predictions is still limited. A more direct and efficient approach to identifying coding sequences is to sequence full-length cDNAs. Complete sequences of RAFL cDNAs will be useful for gene identification and positional cloning. The RAFL cDNA clones are publicly available from the RIKEN Bioresource Center.

Table 2. Functional classification of RAFL cDNA clones.

Functional category	No. of predicted genes	No. of RAFL cDNA clones*
Metabolism	757†	521 (68.8%)
Energy	122†	98 (80.3%)
Cell growth, cell division, and DNA synthesis	96†	54 (56.3%)
Transcription	583†	331 (56.8%)
Protein synthesis	170†	145 (85.3%)
Protein destination	236†	169 (71.6%)
Transport facilitation	252†	151 (60.0%)
Cellular transport and transport mechanisms	119†	89 (74.8%)
Cellular biogenesis	177†	107 (60.5%)
Cellular communication/signal transduction	482†	262 (54.4%)
Cell rescue, defense, death, and aging	325†	172 (52.9%)
Ionic homeostasis	4†	4 (100%)
Cellular organization	365†	255 (69.9%)
Motility	1†	0 (0%)
Development	75†	40 (53.3%)
Transposable elements and viral and plasmid proteins	132†	2 (1.5%)
Organism-specific proteins	1†	0 (0%)
Classification not yet clear-cut	691†	398 (57.6%)
Unclassified proteins	17,213†	8,745 (50.8%)
Protein kinase	1,067‡	506 (47.4%)
Transcription factor	1,533§	1,087 (70.9%)

*The number of RAFL cDNA clones corresponding to the predicted genes in each category was calculated with the BLAST program. The percentages of the RAFL cDNA clones in each category are given in parentheses. †These numbers represent the number of predicted genes in each category of the MIPS functional catalog (*29*). ‡This number represents the number of *Arabidopsis* protein kinase genes in the PlantsP database (*28*). §A recent paper (*27*) estimates 1533 genes coding for transcription factors in *Arabidopsis*.

References and Notes

1. D. W. Meinke *et al.*, *Science* **282**, 662 (1998).
2. The *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).
3. M. Bevan, *Plant Cell* **9**, 476 (1997).
4. X. Lin *et al.*, *Nature* **402**, 761 (1999).
5. The European Union Arabidopsis Genome Sequencing Consortium, The Cold Spring Harbor, Washington University in St. Louis, PE Biosystems Arabidopsis Sequencing Consortium, *Nature* **402**, 769 (1999).
6. R. Cooke *et al.*, *Plant J.* **9**, 101 (1996).
7. H. Höfte *et al.*, *Plant J.* **4**, 1051 (1993).
8. T. Newman *et al.*, *Plant Physiol.* **106**, 1241 (1994).
9. J. A. White *et al.*, *Plant Physiol.* **124**, 1582 (2001).
10. E. Asamizu, Y. Nakamura, S. Sato, S. Tabata, *DNA Res.* **7**, 175 (2000).
11. P. Carninci *et al.*, *Genomics* **37**, 327 (1996).
12. P. Carninci *et al.*, *DNA Res.* **4**, 61 (1997).
13. M. Seki *et al.*, *Plant J.* **15**, 707 (1998).
14. P. Carninci *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 520 (1998).
15. Details of experimental procedures for construction of *Arabidopsis* full-length cDNA libraries, sequencing, and clustering of the cDNA clones are available as supplemental text on *Science* Online at www.sciencemag.org/cgi/content/full/1071006/DC1.
16. P. Carninci *et al.*, *Genomics* **77**, 79 (2001).
17. Y. Shibata *et al.*, *Biotechniques* **30**, 1250 (2001).
18. P. Carninci *et al.*, *Genome Res.* **10**, 1617 (2000).
19. M. Seki *et al.*, *Plant Physiol. Biochem.* **39**, 211 (2001).
20. Web tables and figures can be viewed on *Science* Online (*15*) or at http://www.gsc.riken.go.jp/Plant/index.html.
21. The promoter database was constructed in the following ways. The 5′ ESTs of the 14,034 RAFL cDNAs were mapped onto *the Arabidopsis* genome with the BlastN program. The criterion for mapping was identity >98% within >100-bp overlap. The genomic sequences hit with the highest score were used for construction of the promoter database. The genomic sequences observed in 1000-bp upstream regions of the 5′ termini of the RAFL cDNA clones were regarded as the promoter sequence of each RAFL cDNA clone. About 300 known plant cis-acting elements were then searched in 1000-bp promoter sequences of each RAFL cDNA clone with the PLACE database (*22*).
22. K. Higo, Y. Ugawa, M. Iwamoto, T. Korenaga, *Nucleic Acids Res.* **27**, 297 (1999).
23. See http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide.
24. See http://www.tigr.org/tdb/agi/release.
25. See http://www.tigr.org/tdb/e2k1/ath1/ceres/ceres.shtml.
26. The criterion for mapping was identity >95% within >100-bp overlap.
27. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).
28. See PlantsP database (http://plantsp.sdsc.edu/).
29. See http://mips.gsf.de/proj/thal/db/tables/tables_func_frame.html.

30. We thank the following for technical assistance and discussion: N. Hayatsu, Y. Shibata, T. Shiraki, C. Sakai, T. Tanaka, K. Nishi, S. Akahira, A. Yasunishi, K. Imotani, T. Hiraoka, T. Akimura, A. Arai, R. Numasaki, F. Takahashi, D. Sasaki, Y. Sogabe, A.

Tagawa, T. Tomoya, K. Nomura, T. Hanagaki, and T. Matsuyama. We thank T. Takase and M. Matsui for sampling of the dark-grown plants. Supported by a grant for Genome Research from RIKEN; the Program for Promotion of Basic Research Activities for Innovative Biosciences; the Special Coordination Fund of the Science and Technology Agency; and a Grant-in-Aid from the Ministry of Education, Science and Culture of Japan to K.S. Also supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Science" from the Ministry of Education, Science, Sports and Culture of Japan to M.S. and by a Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H.

19 February 2002; accepted 12 March 2002
Published online 21 March 2002;
10.1126/science.1071006
Include this information when citing this paper.

Conserved Structure for Single-Stranded Telomeric DNA Recognition

Rachel M. Mitton-Fry,[1] Emily M. Anderson,[1] Timothy R. Hughes,[2]* Victoria Lundblad,[2,3] Deborah S. Wuttke[1]†

The essential Cdc13 protein in the yeast *Saccharomyces cerevisiae* is a single-stranded telomeric DNA binding protein required for chromosome end protection and telomere replication. Here we report the solution structure of the Cdc13 DNA binding domain in complex with telomeric DNA. The structure reveals the use of a single OB (oligonucleotide/oligosaccharide binding) fold augmented by an unusually large loop for DNA recognition. This OB fold is structurally similar to OB folds found in the ciliated protozoan telomere end-binding protein, although no sequence similarity is apparent between them. The common usage of an OB fold for telomeric DNA interaction demonstrates conservation of end-protection mechanisms among eukaryotes.

Telomeres are the specialized nucleoprotein complexes that cap eukaryotic chromosomes, protecting chromosome ends from unregulated degradation and end-to-end fusion. Telomeric DNA is typically composed of repetitive, noncoding sequence terminating in a single-stranded TG-rich overhang. Several mechanisms have been identified for capping this overhang, ranging from sequestration through protein binding in ciliates and yeasts to t-loop formation in mammals (*1–3*). Proteins that specifically bind to this single–stranded overhang, such as the *Oxytricha nova* telomere end-binding protein (TEBP) (*4*, *5*), the *Schizosaccharomyces pombe* protection of telomeres 1 (Pot1) and human Pot1 (*6*), and the *Saccharomyces cerevisiae* Cdc13 (*7*, *8*), are involved in telomeric end protection. For example, depletion of Cdc13 activity causes extensive resection of the 5′ strand of the yeast telomere and DNA damage–dependent cell cycle arrest (*9–12*), whereas deletion of the *pot1* gene leads to complete telomere loss and cell death (*6*). Cdc13 is also required for telomere elongation as a positive regulator of telomerase (*7*, *13*). Cdc13 is believed to fulfill both of these important, yet disparate, roles through localization to the 3′ single-stranded telomeric end, followed by recruitment of relevant complexes to the telomere through protein-protein interactions (*14–16*).

Evidence for conservation of telomeric end-protection proteins among distantly related eukaryotes has been elusive. Although the Pot proteins were originally identified on the basis of weak sequence similarity to the NH_2-terminal portion of the α subunit of the heterodimeric *O. nova* TEBP (*6*), no similarity was apparent between any of these proteins and Cdc13. To investigate the requirements for telomeric end protection and sequence-specific interaction with single-stranded DNA (ssDNA), we determined the solution structure of the Cdc13 DNA binding domain (DBD) in complex with telomeric ssDNA. This 23.5-kD domain retains DNA binding activity and specificity (*17–19*), and fusions of the DBD with other components of the end-protection or telomerase machinery eliminate the need for full-length protein in vivo (*14*, *15*). The ssDNA 11-nucleotide (nt) oligomer dGTGTGGGTGTG in the complex is the minimal Cdc13 binding site (*17*) and the complement to the center of the coding region of the telomerase RNA template (*20*).

The high-resolution Cdc13 DBD structure in complex with ssDNA (Fig. 1) was calculated from a total of 2865 nuclear magnetic

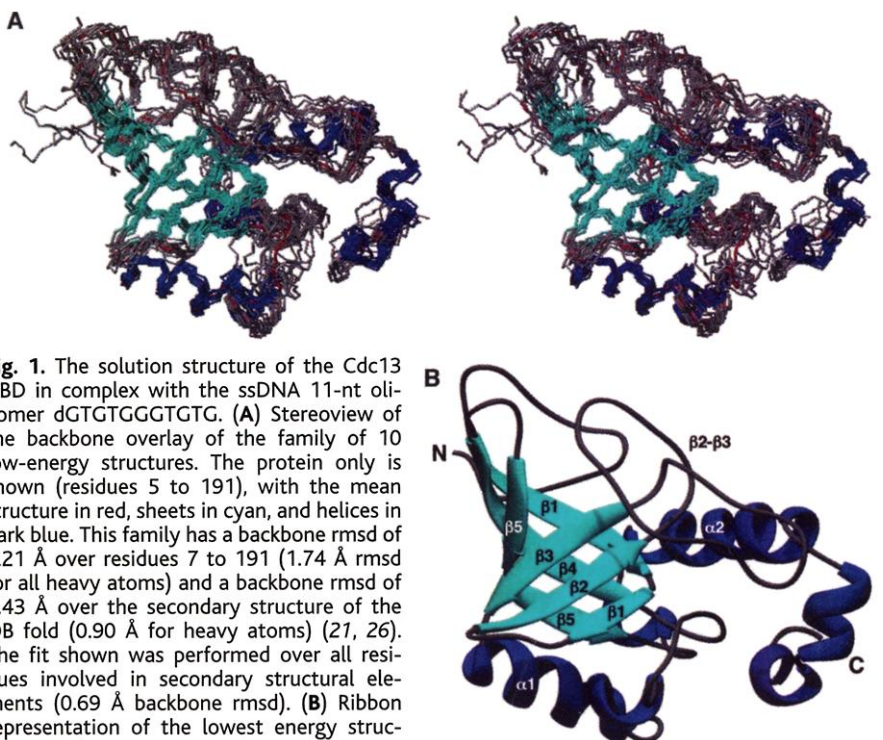

Fig. 1. The solution structure of the Cdc13 DBD in complex with the ssDNA 11-nt oligomer dGTGTGGGTGTG. (**A**) Stereoview of the backbone overlay of the family of 10 low-energy structures. The protein only is shown (residues 5 to 191), with the mean structure in red, sheets in cyan, and helices in dark blue. This family has a backbone rmsd of 1.21 Å over residues 7 to 191 (1.74 Å rmsd for all heavy atoms) and a backbone rmsd of 0.43 Å over the secondary structure of the OB fold (0.90 Å for heavy atoms) (*21*, *26*). The fit shown was performed over all residues involved in secondary structural elements (0.69 Å backbone rmsd). (**B**) Ribbon representation of the lowest energy structure, residues 7 to 191. Figures were prepared with MOLMOL (*33*) and RIBBONS (*34*).

[1]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309, USA. [2]Interdepartmental Program in Cell and Molecular Biology, [3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA.

*Present address: University of Toronto, Banting and Best Department of Medical Research, Toronto, Ontario M5G 1L6, Canada.

†To whom correspondence should be addressed. E-mail: deborah.wuttke@colorado.edu