DNA RESEARCH 7, 175–180 (2000)

A Large Scale Analysis of cDNA in *Arabidopsis thaliana*: Generation of 12,028 Non-redundant Expressed Sequence Tags from Normalized and Size-selected cDNA Libraries

Erika ASAMIZU, Yasukazu NAKAMURA, Shusei SATO, and Satoshi TABATA*

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan

(Received 19 May 2000)

Abstract

For comprehensive analysis of genes expressed in the model dicotyledonous plant, *Arabidopsis thaliana*, expressed sequence tags (ESTs) were accumulated. Normalized and size-selected cDNA libraries were constructed from aboveground organs, flower buds, roots, green siliques and liquid-cultured seedlings, respectively, and a total of 14,026 5′-end ESTs and 39,207 3′-end ESTs were obtained. The 3′-end ESTs could be clustered into 12,028 non-redundant groups. Similarity search of the non-redundant ESTs against the public non-redundant protein database indicated that 4816 groups show similarity to genes of known function, 1864 to hypothetical genes, and the remaining 5348 are novel sequences. Gene coverage by the non-redundant ESTs was analyzed using the annotated genomic sequences of approximately 10 Mb on chromosomes 3 and 5. A total of 923 regions were hit by at least one EST, among which only 499 regions were hit by the ESTs deposited in the public database. The result indicates that the EST source generated in this project complements the EST data in the public database and facilitates new gene discovery. The EST sequence data of individual cDNA clones are available at the web site: http://www.kazusa.or.jp/en/plant/arabi/EST/.
Key words: *Arabidopsis thaliana*; cDNA; EST

1. Introduction

Arabidopsis thaliana has been adopted as a model organism in the study of plant biology since it has the advantages of small size, short generation time, and ease of transformation.[1] Because the *A. thaliana* genome is the smallest genome among known higher plant species (130–140 Mb),[2,3] the genome sequencing project of this plant is underway as a joint project of Japan, Europe, and the United States.[4] To date, two of five chromosomes (chromosomes 2 and 4) have been sequenced except for the nucleolar organizer regions and centromeres,[2,3] and sequencing of the remaining three chromosomes is near completion.

Under these circumstances, the accurate assignment of protein coding regions on the genomic sequence gains importance as the logical next step. In this respect, information on cDNA structure is essential. Also, comprehensive analysis of cDNA sequences is an effective way to catalogue genes expressed in an organism with a large genome. A large number of EST (expressed sequence tag) sequences of several crop plants have been deposited in the public database, dbEST (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). In addition, EST accumulation of several model plants have been initiated.[5–7]

In *A. thaliana*, more than 45,000 EST sequences have been deposited in dbEST, including sequences from large-scale EST projects promoted by two consortia of laboratories, one in France and the other in the United States.[8–10] The French program generated sequence data from ten kinds of cDNA libraries prepared from different tissues, organs and developmental stages. They deposited the 5′- and 3′-end sequences of approximately 6,000 non-redundant clones in dbEST.[8,9] The U.S. group produced 31,000 ESTs mainly from a single library made from a mixture of mRNAs from four different tissues.[10] These EST clones altogether cover approximately 34% of the predicted genes on chromosome 4.[3]

To complement the EST data currently available in the public database and facilitate new gene discovery, we constructed normalized and size-selected cDNA libraries from five different tissues of *A. thaliana*, and accumulated 5′-end and 3′-end ESTs.

Communicated by Mituru Takanami
* To whom correspondence should be addressed. Tel. +81-438-52-3933, Fax. +81-438-52-3934, E-mail: tabata@kazusa.or.jp

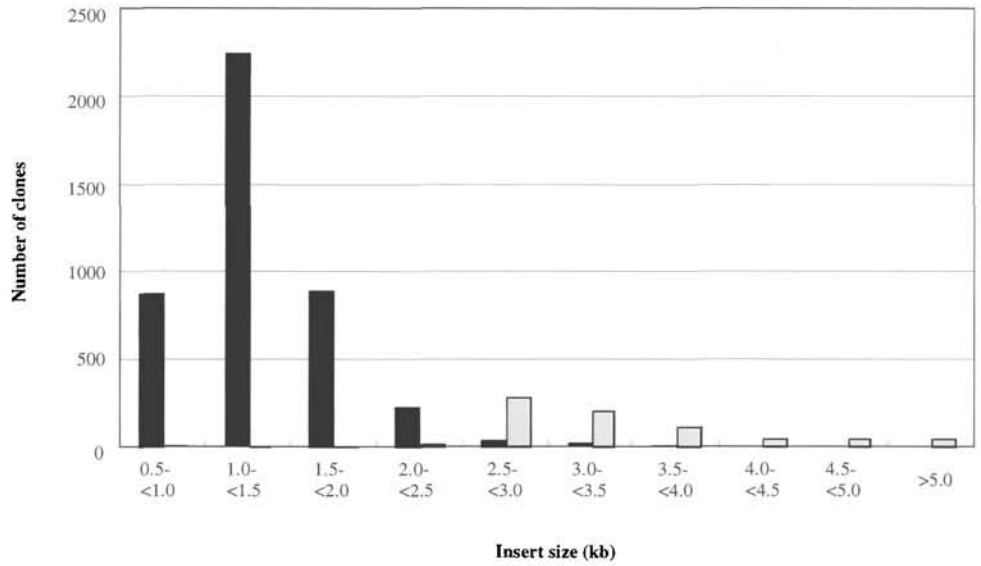

Figure 1. Size distribution of the inserts in analyzed cDNA clones from the normalized (solid bars) and size-selected libraries (gray bars) of aboveground organs.

2. Materials and Methods

2.1. Preparation of tissues

Arabidopsis thaliana Columbia accession was used for analysis and was grown in soil under a 16 hr photoperiod at 22°C. Aboveground organs were harvested from 2- to 6-week-old plants. Flower buds and green siliques were also collected from the soil-grown plants. For liquid culture, sterile seeds were sown in medium [1/2 B5 medium,[11] 1/1000 HYPONeX (Hyponex Japan), 1% sucrose, pH adjusted to 5.7] and grown under continuous light at 22°C with rotation for 2 weeks. Seedlings and roots were collected from the liquid-cultured plants.

2.2. Isolation of poly(A)$^+$ RNA and construction of cDNA libraries

Total RNA was extracted from aboveground organs, flower buds, roots, and liquid-cultured seedlings by the guanidium thiocyanate/CsCl ultracentrifugation method, and from green siliques by the SDS/phenol method as described previously.[5,7] Purification of poly(A)$^+$ RNA, conversion to cDNA, and size-selection of cDNA was performed as described.[5] Normalization was performed for the library containing 0.5- to 3-kb fragments as described.[5,12] The names of cDNA libraries refer to the tissue used for construction: AP, aboveground organs; FB, flower buds; RZ, roots; SQ, green siliques; pAZNII, liquid-cultured seedlings.

2.3. Template preparation and sequencing

For generation of all the 5′-end sequences as well as some of the 3′-end sequences from the AP and RZ libraries, PCR amplified fragments were used as a template. Vector-derived sequences were used as primers (5′-TGTGCTGCAAGGCGATTAAGTTGGG-3′, and 5′-TCATTAGGCACCCCAGGCTTTACAC-3′), and PCR was performed by Taq DNA polymerase (TaKaRa, Japan) using a Perkin-Elmer 9600 Thermal Cycler: 30 cycles of 10 sec at 98°C, 6 min at 68°C, and a final extension for 10 min at 72°C. The amplified products were precipitated by adding 1/3 to the final volume of 20% PEG6000 in 2.5 M NaCl. Plasmid DNA was used as a template for generation of the rest of 5′- and 3′-end sequences. Plasmid DNA preparation and insert size determination of each clone was performed as described.[5] Sequence reaction was performed by Dye Terminator, dRhodamine Terminator, and BigDye Terminator Cycle Sequencing Ready Reaction Kit (PE Applied Biosystems, USA) and electrophoresed on the automated DNA sequencers (ABI PRISM 373 and 377XL, PE Applied Biosystems, USA).

2.4. Sequence data analysis

Only the 3′-end sequences were subjected to the data analysis process. The vector-derived sequence and ambiguous sequences were removed from the collected EST sequences prior to the computer-aided analyses. Each sequence was translated into its amino acid sequences in six frames and subjected to similarity search against the non-redundant protein database provided by NCBI using the BLAST algorithm.[13] Similarity between a deduced amino acid sequence and a known sequence was judged to be significant when the P-value was less than 1.0^{-14}. To identify the number of independent EST species, clustering of the EST sequences was performed. The 3′-end sequences were compared with a dataset of itself using

Table 1. Number of 5′-end and 3′-end ESTs generated from cDNA libraries of five different tissues.

Tissue	Library type	Number of 5′ end ESTs	Number of 3′ end ESTs
Aboveground organs	Normalized	4996	6863
	Size-selected	2172	1753
Flower buds	Normalized	ND	5827
Roots	Normalized	5798	8505
	Size-selected	245	3161
Green siliques	Normalized	ND	11843
	Size-selected	ND	909
Liquid-cultured seedlings	Normalized	815	346
Total		14026	39207

ND: Not determined

the BLASTN program and clones that showed over 95% identity for more than 50 bp were included in the same group.

3. Results and Discussion

3.1. Quality of cDNA libraries

The size distribution of the inserts in cDNA was analyzed for the clones from the cDNA libraries of aboveground organs. As shown in Fig. 1, 72.3% of the clones from the normalized library contained inserts of 0.5 to 1.5 kb, while 96.0% of the clones from the size-selected library had inserts longer than 2.5 kb. The average insert-length of the clones in the normalized library was 1.28 kb, whereas that of the size-selected library was 3.17 kb. It is therefore evident that the size selection procedure is effective for generation of long cDNA species.

The quality of the libraries with respect to the intactness of cDNA was assessed by comparison of the 5′-end sequences to known protein sequences. Among 116 clones randomly chosen from the normalized library and 122 clones from the size-selected library, 74 (63.8%) and 85 (69.7%) were found to contain a translation initiation codon, respectively, indicating that roughly two-thirds of the cDNA clones are full-length in both libraries. This result shows that the two libraries contain an abundance of intact cDNA species with shorter and longer sizes. However, we only assessed the quality of libraries using those prepared from aboveground organs. The quality may be different among libraries from different tissues.

3.2. Generation of ESTs

cDNA clones were randomly chosen from the cDNA libraries constructed, and a total of 14,026 clones were sequenced from the 5′-ends and 39,207 clones were sequenced from the 3′-ends. The number of ESTs generated from the respective libraries are summarized in Table 1. The GC content of the randomly selected 659 ESTs (279,604 bases) was estimated to be 43.4%.

Table 2. Classification of 3′-end ESTs by similarity search against the non-redundant protein database.

Similarity	Number of clones	Number of non-redundant ESTs
Genes of known function[a)]	24892	4816
Hypothetical genes[b)]	5071	1864
No similarity[c)]	9244	5348
Total	39207	12028

a) showed similarity to genes of known function, b) showed similarity to hypothetical genes that have no definition of function, c) showed no similarity

To identify the number of independent EST species, clustering of the 3′-EST sequences was performed. As a result, the 39,207 3′-EST sequences were clustered into 12,028 independent groups. This number is supposed to be close to the actual number of gene species represented by ESTs. However, a more accurate number of independent gene species should be obtained by allocating the EST sequences on the genome, because the stringency used for clustering was not strict (95% identity for 50 bp).

3.3. Sequence similarity of ESTs

When the non-redundant EST groups deduced from the 3′-end ESTs were searched for similarity using the non-redundant protein database, 6680 groups had significant similarity to the registered sequences and the remaining 5348 groups were novel. Among the 6680 EST groups with significant similarity, 4816 showed similarity to genes with known function and the remaining 1864 to hypothetical genes, with no functional definition largely predicted from the *A. thaliana* genome sequences (Table 2). Genes whose functions could be predicted from a similarity search were classified according to the biological roles or biochemical functions[2] as shown in Table 3. The search results of the individual clone are available at the web site, http://www.kazusa.or.jp/en/plant/arabi/EST/.

3.4. Estimation of gene coverage by ESTs

Gene coverage of the non-redundant EST groups was investigated using the annotated genomic sequences, 10,009,832 bp in length, on chromosomes 3[14] and 5[15] (http://www.kazusa.or.jp/kaos/). The sequences taken were of 1 P1 clone on chromosome 3, and 106 P1 and 30 TAC clones on chromosome 5. Along the genomic sequences, 2324 regions have been assigned as potential protein-coding genes. Analysis indicated that 788 were hit by at least one EST group. In addition, 135 EST

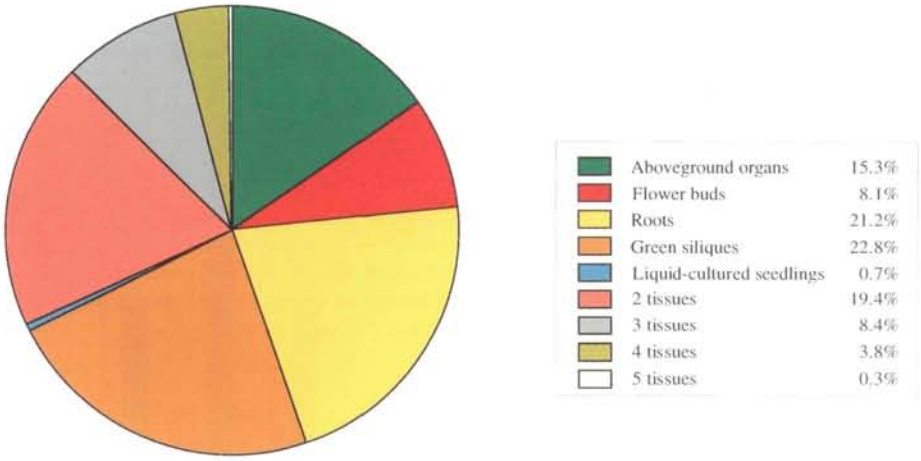

Figure 2. The proportions of EST groups identified only in one of the five tissues and those identified in two to five tissues. The proportion of each category are given as percentages and indicated by color codes.

Table 3. Classification of the non-redundant EST groups with similarity to known protein genes by their functional categories.

Functional categories	Number of non-redundant groups
Energy metabolism	538
Regulatory functions	483
Cellular structure, organization and biogenesis	416
Protein fate	369
Signal transduction	360
Protein synthesis	328
Transport and binding proteins	319
Cellular processes	218
Secondary metabolism	209
Growth and development	184
Fatty acid and phospholipid metabolism	132
Amino-acid biosynthesis	125
Environmental response	113
Pathogen responses	99
DNA metabolism	89
General transcription	76
Purines, pyrimidines, nucleosides, and nucleotides	68
Central intermediary metabolism	53
Biosynthesis of cofactors, prosthetic groups, and carriers	52
Other categories	23
Unclassified	562
Total	4816

groups could be located at regions where no gene assignment has been done. Gene coverage by ESTs deposited in the database was examined, and only 499 out of 923 regions hit by our EST groups were found to hit.

We also analyzed gene coverage by mapping of the EST groups on the completed sequence of *A. thaliana* chromosome 2, on which 4037 genes have been assigned.[2] As a result, 1775 EST groups were allocated on the genomic sequence, of which 626 groups were found to have similar sequences in the registered ESTs. Although the gene coverage data observed for different chromosomal sequences can not be compared directly, the data obviously indicate that the non-redundant ESTs generated in this project contain many new cDNA species.

3.5. Classification of ESTs with respect to tissue-specific expression

To gain information on the expression specificity of genes identified by EST analysis, the occurrence of non-redundant ESTs in the population of 3′-end ESTs generated from each tissue was counted. The 12,028 non-redundant EST groups were classified into nine categories: The groups identified only in one of the five tissues, and those identified in two to five tissues. The percentages of the EST groups classified in each category to the total non-redundant EST groups are shown in the pie chart in Fig. 2. Although the population of the 3′-end ESTs in each tissue is not large enough to speculate, the proportion of EST groups identified only in one of the five tissues was surprisingly high (68.1%) compared with those identified in multiple tissues. The result implies that the classified EST groups are good sources for finding genes with tissue-specific expression. In Table 4, the identity of genes which are abundantly represented by ESTs in four different tissues (aboveground organs, flower buds, roots and green siliques) is listed. Some genes on this list have been reported to show tissue-specific expression.[16–18] However, five kinds of ribosomal protein genes are seen in the flower bud groups, indicating the necessity of further analysis with more large EST populations.

The EST sequences reported in this paper appear in the GenBank/EMBL/DDBJ databanks with accession numbers AB038710-AB038726, AV439465-AV442830 and AV517879-AV567728.

Acknowledments: We thank A. Watanabe, T. Wada, N. Nakazaki, K. Naruo, M. Ishikawa, and M. Yamada for excellent technical assistance. This work was supported by the Kazusa DNA Research Institute Foundation.

Table 4. The identity of genes abundantly represented by 3′-end ESTs in aboveground organs, flower buds, roots and green siliques.

Number of clones	Definition of the most similar sequence
	Aboveground organs
8	COL2 [Arabidopsis thaliana]
8	strong similarity to Arabidopsis 2A6 (gb|X83096). [Arabidopsis thaliana]
6	ethylene-forming enzyme [Arabidopsis thaliana]
5	unknown
5	involved in starch metabalism [Solanum tuberosum]
5	unknown
5	nitrate reductase NR1 (393 AA) [Arabidopsis thaliana]
5	unknown
5	unknown
4	beta-1,3-glucanase 2 [Arabidopsis thaliana]
	Flower buds
19	PsCL18 ribosomal preprotein (AA -49 to 96) [Pisum sativum]
12	unknown
11	similar to Prunus pectinesterase (gb|X95991). [Arabidopsis thaliana]
11	ribosomal protein L30 [Lupinus luteus]
10	acidic ribosomal protein P3a [Zea mays]
8	anther-specific gene product; putative [Brassica campestris][17]
8	putative ribosomal protein [Arabidopsis thaliana]
8	similar to lipid transfer protein [Brassica rapa][17]
7	putative ribosomal protein S16 [Arabidopsis thaliana]
7	NAP16kDa protein [Arabidopsis thaliana]
	Root
28	jasmonate inducible protein isolog [Arabidopsis thaliana]
18	peroxidase ATP11a [Arabidopsis thaliana]
16	cucumisin [Arabidopsis thaliana]
15	cytochrome P450 monooxygenase [Arabidopsis thaliana]
14	peroxidase ATP8a [Arabidopsis thaliana]
13	putative plasma membrane-cell wall linker proteins [Arabidopsis thaliana][16]
12	flavonol synthase [Arabidopsis thaliana]
11	beta-glucosidase [Arabidopsis thaliana]
9	Dr4 [Arabidopsis thaliana]
8	ABC transporter (PDR5-like) isolog [Arabidopsis thaliana]
	Green siliques
29	APG protein isolog [Arabidopsis thaliana]
26	12S cruciferin seed storage protein [Arabidopsis thaliana][18]
26	gamma-VPE [Arabidopsis thaliana]
24	thioesterase homolog [Arabidopsis thaliana]
24	dihydroflavonol 4-reductase [Arabidopsis thaliana]
17	putative pectinesterase [Arabidopsis thaliana]
14	germin-like protein [Arabidopsis thaliana]
13	12S storage protein CRB [Arabidopsis thaliana][18]
11	unknown
10	putative protein [Arabidopsis thaliana]

References

1. Meinke, D. W., Cherry, J. M., Dean, C. D., Rounsley, S., and Koornneef, M. 1998, *Arabidopsis thaliana*: a model plant for genome analysis, *Science*, **282**, 662–682.
2. Lin, X., Kaul, S., Rounsley, S. et al. 1999, Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*, *Nature*, **402**, 761–768.
3. Mayer, K., Schüller, C., Wambutt, R. et al. 1999, Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*, *Nature*, **402**, 769–777.
4. Bevan, M. 1997, Objective: the complete sequence of a plant genome, *Plant Cell*, **9**, 476–478.
5. Asamizu, E., Nakamura, Y., Sato, S., Fukuzawa, H., and Tabata, S. 1999, A large scale structural analysis of cDNAs in a unicellular green alga, *Chlamydomonas reinhardtii*. I. Generation of 3433 non-redundant expressed sequence tags, *DNA Res.*, **6**, 369–373.
6. Asamizu, E., Nakamura, Y., Sato, S., and Tabata, S. 2000, Generation of 7,137 Non-redundant Expressed Sequence Tags from a Legume, *Lotus japonicus*, *DNA Res.*, **7**, 127–130.
7. Nikaido, I., Asamizu, E., Nakajima, M., Nakamura, Y., Saga, N., and Tabata, S. 2000, Construction of a gene catalogue of a marine red alga, *Porphyra yezoensis*. I. Generation of 10,154 expressed sequence tags, *DNA Res.*, this issue.
8. Höfte, H., Desprez, T., Amselem, J. et al. 1993, An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*, *Plant J.*, **4**, 1051–1061.
9. Cooke, R., Raynal, M., Laudié, M. et al. 1996, Further progress towards a catalogue of all Arabidopsis genes: analysis of a set of 5000 non-redundant ESTs, *Plant J.*, **9**, 101–124.
10. Newman, T., de Bruijn, F. J., Green, P. et al. 1994, Genes galore: a summary of methods for accessing results from

large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones, *Plant Physiol.*, **106**, 1241–1255.
11. Horsch, R. B., Fry, J., Hoffman, N., Neidermeyer, J., Rogers, S. G., and Fraley, R. T. 1988, Plant Molecular Biology Manual, Kluwer Academic Publishers, A5: 1–9.
12. Bonaldo, M. F., Lennon, G., and Soares, M. B. 1996, Normalization and subtraction: two approaches to facilitate gene discovery, *Genome Res.*, **6**, 791–806.
13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–410.
14. Sato, S., Nakamura, Y., Kaneko, T., Katoh, T., Asamizu, E., and Tabata, S. 2000, Structural Analysis of *Arabidopsis thaliana* Chromosome 3. I. Sequence Features of the Regions of 4,504,864 bp Covered by Sixty P1 and TAC Clones, *DNA Res.*, **7**, 131–135.
15. Sato, S., Nakamura, Y., Kaneko, T. et al. 2000, Structural Analysis of *Arabidopsis thaliana* Chromosome 5. X. Sequence Features of the Regions of 3,076,755 bp Covered by Sixty P1 and TAC Clones, *DNA Res.*, **7**, 31–63.
16. Neuteboom, L. W., Ng, J. M., Kuyper, M., Clijdesdale, O. R., Hooykaas, P. J., and van der Zaal, B. J. 1999, Isolation and characterization of cDNA clones corresponding with mRNAs that accumulate during auxin-induced lateral root formation, *Plant Mol. Biol.*, **39**, 273–287.
17. Kim, H. U. and Chung, T. Y. 1997, Characterization of three anther-specific genes isolated from Chinese cabbage, *Plant Mol. Biol.*, **33**, 193–198.
18. Parcy, F., Valon, C., Kohara, A., Misera, S., and Giraudat, J. 1997, The *ABSCISIC ACID-INSENSITIVE3, FUSCA3*, and *LEAFY COTYLEDON1* loci act in concert to control multiple aspects of Arabidopsis seed development, *Plant Cell*, **9**, 1265–1277.