DNA Research **15**, 333–346, (2008) doi:10.1093/dnares/dsn024

Sequencing and Analysis of Approximately 40 000 Soybean cDNA Clones from a Full-Length-Enriched cDNA Library

Taishi Umezawa[1,†], Tetsuya Sakurai[2,†], Yasushi Totoki[3], Atsushi Toyoda[4], Motoaki Seki[5], Atsushi Ishiwata[2], Kenji Akiyama[2], Atsushi Kurotani[2], Takuhiro Yoshida[2], Keiichi Mochida[6], Mie Kasuga[7], Daisuke Todaka[7,15], Kyonoshin Maruyama[7], Kazuo Nakashima[7], Akiko Enju[5], Saho Mizukado[1], Selina Ahmed[7], Kyoko Yoshiwara[7], Kyuya Harada[8], Yasutaka Tsubokura[8], Masaki Hayashi[8], Shusei Sato[9], Toyoaki Anai[10], Masao Ishimoto[11], Hideyuki Funatsuki[11], Masayoshi Teraishi[12], Mitsuru Osaki[13], Takuro Shinano[11], Ryo Akashi[14], Yoshiyuki Sakaki[3,4], Kazuko Yamaguchi-Shinozaki[7,15], and Kazuo Shinozaki[1]*

Gene Discovery Research Team, RIKEN Plant Science Center, Koyadai 3-1-1, Tsukuba, Ibaraki 305-0074, Japan[1]; Integrated Genome Informatics Research Unit, RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan[2]; Genome Annotation and Comparative Analysis Team, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan[3]; Sequence Technology Team, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan[4]; Plant Genomic Network Research Team, RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan[5]; Gene Discovery Research Group, RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan[6]; Biological Resources Division, Japan International Research Center for Agricultural Sciences (JIRCAS), 1-1 Ohwashi, Tsukuba, Ibaraki 305-8686, Japan[7]; National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan[8]; Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan[9]; Department of Applied Biological Sciences, Faculty of Agriculture, Saga University, Honjo 840-8502, Saga, Japan[10]; National Agricultural Research Center for Hokkaido Region, 1 Hitsujigaoka, Sapporo, Hokkaido 062-8555, Japan[11]; Experimental Farm, Kyoto University, Takatsuki, Osaka 569-0096, Japan[12]; Graduate School of Agriculture, Hokkaido University, Sapporo, Hokkaido 060-8589, Japan[13]; Division of BioResource, Frontier Science Research Center, University of Miyazaki, Miyazaki 889-2192, Japan[14] and Laboratory of Plant Molecular Physiology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan[15]

(Received 3 June 2008; accepted 10 September 2008; published online 16 October 2008)

Abstract

A large collection of full-length cDNAs is essential for the correct annotation of genomic sequences and for the functional analysis of genes and their products. We obtained a total of 39 936 soybean cDNA clones (GMFL01 and GMFL02 clone sets) in a full-length-enriched cDNA library which was constructed from soybean plants that were grown under various developmental and environmental conditions. Sequencing from 5′ and 3′ ends of the clones generated 68 661 expressed sequence tags (ESTs). The EST sequences were clustered into 22 674 scaffolds involving 2580 full-length sequences. In addition, we sequenced 4712 full-length cDNAs. After removing overlaps, we obtained 6570 new full-length sequences of soybean cDNAs so far. Our data indicated that 87.7% of the soybean cDNA clones contain complete coding sequences in addition to 5′- and 3′-untranslated regions. All of the obtained data confirmed that our collection of soybean full-length cDNAs covers a wide variety of genes. Comparative analysis between the derived sequences from soybean and *Arabidopsis*, rice or other legumes data revealed that some specific genes were involved in our collection and a large part of them could be annotated to unknown functions. A large set of soybean full-length cDNA clones reported in

Edited by Satoshi Tabata

* To whom correspondence should be addressed. Tel. +81 29-836-4359. Fax. +81 29-836-9060. E-mail: sinozaki@rtc.riken.jp.

† Contributed equally to this work.

this study will serve as a useful resource for gene discovery from soybean and will also aid a precise annotation of the soybean genome.

Key words: EST; full-length cDNA; functional annotation; legume; soybean

1. Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the most important crops in the world. The agronomical importance of soybean has been steadily increasing, since it is an important source for protein and vegetable oil for human and animal nutrition. In addition, soybean serves as a valuable renewable agricultural source for industrial products, e.g. lubricating oil, printing ink or biodiesel.[1] Owing to this importance of soybean, its agronomical features should be distinguished at the molecular level to facilitate breeding, gene discovery or industrial applications.

At the present time, access to a large set of genomic information is huge benefit for biological or agronomical research. In the past decade, a large amount of genomic information had been compiled from a number of organisms involving several model plants.[2–5] The vast amount of sequence information, including not only whole-genome sequences but also other information such as transcriptome, proteome or metabolome data, has enabled scientists to extend our understandings for genomic structures, evolution, gene discovery or gene functions, etc.[6]

The genome size of soybean is large (1115 Mb with $2n = 40$) and it likely arose from complex genome duplication events.[7,8] It was suggested that at least one of the original genomes was duplicated prior to the most recent polyploidization event in soybean.[7] Despite such complexity, the draft sequence of soybean genome has been already released (http://www.phytozome.net/soybean). Furthermore, genomic sequencing has been conducted for two model legume plants, *Lotus japonicus* and *Medicago truncatula*, and a large set of *Lotus* genomic sequence data are now available.[9] These data will provide us a great potential to a broad range of plant science as well as legume research.

In addition to an entire genome sequence, a catalog of gene transcripts can also serve as a critical resource for molecular studies. Actually, >390 000 soybean expressed sequence tag (EST) sequences have already been obtained (http://www.ncbi.nlm.nih.gov/dbEST/) from different tissues, organs, seeds and developmental stages of soybean. These data serve as a valuable resource to help describe gene expression profiles and ultimately classify genes by families and functions.[10] Especially, full-length cDNA collections can serve as a powerful tool to facilitate genomic or other omics research efficiently. Several techniques have been established to prepare full-length cDNA enriched libraries from various organisms,[11–13] and the usefulness of full-length cDNAs has been confirmed in various plants such as *Arabidopsis*, rice, poplar, wheat and maize.[14–19] A major advantage of this approach is that the most of clones contain the complete coding sequences as well as the 5′- and 3′-untranslated regions (UTRs). Inclusion of the entire sequence data dramatically facilitates the subsequent sequencing, annotation, and protein expression and other functional assays.[18] Furthermore, a large collection of full-length sequences of cDNA clones also provides a set of protein sequences allowing us to estimate gene functions by searching homology to other proteins, conserved domains or motifs. Thus, the preparation and analysis of full-length cDNA clones maybe closely connected to soybean genomic sequencing projects.

In the present study, we report a large-scale collection of full-length cDNA clones derived from a Japanese soybean cultivar, Nourin No. 2. Soybean was domesticated in East Asia, where various kinds of landraces have been established as a result of adaptation to different environments and the diversification of food cultures. 'Nourin No. 2′ is an elite cultivar which had been developed by cross-breeding and is one of the ancestors of Japanese modern cultivars. It has many typical features of Japanese soybean, for example, a white seed coat, a relatively large seed size and high protein content. We constructed a full-length-enriched cDNA library from Nourin No. 2 and obtained 22 674 non-redundant cDNA sequences from 5′ or 3′ end sequences of 39 936 randomly selected clones from the library. In addition, the entire sequences of 4712 full-length cDNAs were determined. We designed a web-based interface to retrieve sequence information of soybean full-length cDNA clones (http://rsoy.psc.riken.jp/). The information and resources pertaining to full-length soybean cDNAs will be publicly available from the National Bioresource Project for *Lotus*/*Glycine* in Japan (http://www.legumebase.agr.miyazaki-u.ac.jp/).

2. Materials and methods

2.1. Plant materials

Soybean [*G. max* (L.) Merr. cv. Nourin No. 2] was used to construct a full-length cDNA library. In this study, we prepared soybean plants under various developmental and environmental conditions: (1) drought stress, (2) salt stress, (3) chilling stress

(4°C), (4) low temperature (15°C), (5) phosphorous starvation, (6) flooding, (7) nematode infection, (8) flower buds, (9) nodules and (10) developing seeds. These growth conditions and treatments are summarized in Table 1, and the procedures are described as follows.

Drought, salt and chilling stresses. Soybean seeds were sown in vermiculite for 7 days, then seedlings were grown hydroponically under greenhouse conditions (36°0′N, 140°1′E, Tsukuba, Japan). Average temperature was 21°C, and day length was 14 h. Seedlings were cultured in 30 L vessels containing a nutrient solution as previously described.[20] Two-week-old seedlings were exposed to a drought stress by withholding nutrient solution, and salt stress was produced by supplementing plants with a nutrient solution containing 100 mM NaCl. For cold stress, seedlings were transferred into a refrigerate chamber in which temperature was controlled to 4°C. Seedlings were harvested at 0, 1, 5, 10 and 24 h after the initiation of stresses. Each plant sample was stored at −80°C until use.

Low temperature treatment. As previously described, soybean plants were grown in a growth chamber at 22/17°C with 15 h light until the first trifoliate leaves were fully expanded.[21] The only exception was that culture soil Sankyo Engei Baido, containing 374 mg of N, 647 mg of P and 201 mg of K (Hokkai Sankyo, Sapporo), was used. The plants were subsequently grown at 15/15°C for 1 h, 1 and 4 days before total RNA was isolated from the whole shoots.

Phosphorous starvation. Soybean seeds were surface sterilized with 70% ethanol for 30 s with subsequent washing in deionized water. Sterilized seeds were sown in vermiculite for 7 days and seedlings were subsequently grown hydroponically under greenhouse conditions (43°3′N, 141°2′E, Sapporo, Japan). Average temperature was 24°C and day length was 13–14 h. Eight seedlings were cultured in a 56 L vessel, containing a nutrient solution made up of 0.83 mM N (NH_4NO_3), 0 μM (−P solution) or 32 μM (+P solution) NaH_2PO_4, 0.38 mM K (KCl), 0.19 mM K (K_2SO_4), 0.75 mM Ca ($CaCl_2{\cdot}2H_2O$), 0.82 mM Mg ($MgSO_4{\cdot}7H_2O$), 35.8 μM Fe ($FeSO_4{\cdot}7H_2O$), 9.1 μM Mn ($MnSO_4{\cdot}4H_2O$), 46.3 μM B (H_3BO_3), 3.1 μM Zn ($ZnSO_4{\cdot}7H_2O$), 0.16 μM Cu ($CuSO_4{\cdot}5H_2O$) and 7.4 nM Mo [$(NH_4)6Mo_7O_{24}{\cdot}4H_2O$]. The concentration of phosphorus, as well as the pH (5.3 ± 0.1), was adjusted every day. Plants were harvested after 2 weeks and stored at −80°C until further use.

Flooding treatment. The water content in soybean seeds was treated at 15% in the humidity conditioning chamber. The seeds were immersed and germinated in the tubs with sterilized water at a depth of 5 cm. The flooding treatments were imposed for 3 or 5 days at 25°C under dark conditions.

Table 1. Collection of RNA samples for constructing a soybean full-length cDNA library

	Sample name	Treatments or strains	Condition 1	Condition 2	Tissues	Place[a]
1	Drought stress	Removal of media	Hydroponics	Green house	Whole plants	Ibaraki
2	Salt stress	100 mM NaCl	Hydroponics	Green house	Whole plants	Ibaraki
3	Chilling stress	4°C	Hydroponics	Green house	Whole plants	Ibaraki
4	Low temperature	15°C	Pot	Green house	Whole plants	Hokkaido1
5	Pi starvation	Nutrient solution without Pi	Hydroponics	Green house	Whole plants	Hokkaido2
6	Flooding	Imbibition	Hydroponics	Green house	Whole plants	Hokkaido1
7	SCN infected roots	*Heterodera glycines* Ichinohe	Soil	Field	Roots	Hokkaido1
8	Flower buds	Normal condition	Soil	Field	Flower buds	Chiba
9	Roots and nodules	*Bradyrhizobium japonicum* strains A1017 and USDA110	Vermiculite and soil	Field	Roots	Chiba
10	Developing seeds	Normal condition	Soil	Field	Seeds	Saga

We prepared RNA samples from soybean plants distributed by several laboratories in Japan. The regions, conditions and treatments for soybean plants widely ranged as shown here. Additional information was described in Materials and Methods.

[a]The latitude and longitude for each place is as follows: Ibaraki, 36°0′N, 140°1′E; Hokkaido1, 43°1′N, 141°9′E; Hokkaido2, 43°3′ N, 141°2′ E; Chiba, 35°8′N, 139°9′E and Saga, 33°2′N, 130°3′E.

Germinating seedlings were washed with the water and were then used for the isolation of total RNA.

Inoculation of soybean cyst nematode (SCN: Heterodera glycines Ichinohe). Soybean plants were grown in the field infested with SCN race 3 in Sarabetsu, Hokkaido (43°1′N, 141°9′E) on 14 May 2003. Ten inoculated plans were harvested after 50 days from sowing and their roots from the plants were washed free of soil and used for the isolation of total RNA. Average temperature was 15°C, and day length was 15 h.

Flower buds and nodules. Soybean plants were grown in a greenhouse or a field at Matsudo, Japan (35°8′N, 139°9′E; average temperature, 25°C; day length, 14 h). Flower buds were excised from field-grown soybean plants at 50 days after sowing (DAS), and roots and nodules were sampled at 20 and 50 DAS. Alternatively, soybean plants were grown in vermiculite with a Hyponex solution (HYPONeX Japan, Osaka, Japan) under a greenhouse condition. Inoculation with a suspension of *Bradyrhizobium japonicum* strains A1017 and USDA110 (each 1×108 cells/mL) was carried out twice on the first and third day after sowing. Roots and nodules were sampled at 20 and 50 DAS.

For developing seeds, soybean plants were grown in a greenhouse at Saga, Japan (33°2′N, 130°3′E; average temperature, 20–27°C; day length, 14 h). Developing seeds were harvested at 7, 14, 25, 35 and 50 days after flowering. The harvested samples were separated into six pools by seed-length (under 5, 5–7, 7–9, 9–12, 12–15, and over 15 mm), and they were stored at −80°C until use.

2.2. RNA extraction

Total RNA was extracted from soybean plants using TRIzol reagent (Invitrogen, CA, USA) according to the manufacturer's instruction. An alternative protocol was used for developing seeds as follows; frozen samples were crushed to a fine powder with a mortar and pestle in liquid N_2. Each crushed sample was homogenized with 5 volumes (FW/V) of 100 mM Tris–HCl buffer (pH 8.0) containing 90 mM LiCl, 4.5 mM EDTA and 1% SDS, and 2 volumes of water-saturated phenol. After adding 1 volume of 2 M sodium acetate (pH 4.0) and 2 volumes of chloroform, the aqueous phase was separated by centrifugation at 10 000*g* for 10 min. The aqueous phase was re-extracted three times with phenol:chloroform (1:1, v/v) and chloroform. Total RNA was precipitated as a lithium salt by adding 3 volumes of 8 M LiCl. For cDNA library construction, contaminated oligosaccharides were removed from total RNA with the glass fiber-mediated method according to the manufacturer's instructions (RNeasy Plant Kit, QIAGEN).

2.3. Construction of a full-length-enriched cDNA library

Aliquots of total RNA from plant materials (Table 1) were mixed equally, and the RNA mixture was used for the construction of a full-length-enriched cDNA library. Construction of the library was accomplished by the biotinylated CAP trapper method and trehalose-thermoactivated reverse transcriptase as described in the previous reports.[11,22] The resultant double-strand cDNAs were ligated into a λFLC-III vector.[23]

2.4. Both-ends sequencing of soybean cDNA clones

The DNA of each clone was directly amplified from 384 bacterial cultures of a glycerol stock plate by the RCA method[24] using a TempliPhi HT DNA amplification kit (GE Healthcare, UK). End sequencing of 39 936 clones was carried out using ABI 3700 capillary sequencers (Applied Biosystems). The M13-21 primer (5′-TGTAAAACGACGGCCAGT-3′) and the 1233 primer (5′-AGCGGATAACAATTTCACACAGGA-3′) were used for forward and reverse sequencing, respectively.

2.5. Sequence data trimming and assembly

Raw sequence data were base-called using the Phred program[25,26] and the low quality region (Phred quality score <20, and more than 20 bases repeated) which was found at both edges of each raw sequence was discarded. We used the sim4 program[27] for the detection of vector sequences. Sequence data of lengths shorter than 100 bases after this trim process were omitted. In addition, if the repetition of a single nucleotide in a sequence was longer than 10% of its total length, we rejected such sequence. Sequences with high similarity to the soybean cystnematode (SCN: *Heterodera glycines*) gene (BLASTN *e*-value $<1e-50$) were also removed. The ESTs were assembled by the CAP3 program[28] with 40 bp overlap and 90% sequence identity. All EST sequences were submitted to the DNA Databank of Japan (DDBJ) under accession numbers BW650749–BW684913 and DB955456–DB990717.

2.6. Full-length sequencing for soybean cDNAs

We selected 4712 clones from the GMFL01 clone set for full-length sequencing, according to their expression profiles (as described in Results and discussions). One representative clone from each contig was selected, and a total of 4712 clones were re-arrayed from the original plates to new 384-well plates. The DNA of each clone from the re-arrayed plates was amplified as described above. Full-length sequencing

was performed by both primer walking and shotgun methods using ABI 3730 capillary sequencers (Applied Biosystems). In the finishing process, the Phred/Phrap/Consed system[25,26,29] was used to assemble sequences. All full-length sequences were submitted to the DDBJ under the accession numbers AK243693–AK246134 and AK285150–AK287419.

2.7. Full-length cDNA library quality

The quality of the soybean full-length cDNA library was evaluated as follows: (i) a clone had both 5′ and 3′ sequence data, (ii) the *e*-values in a search result of 5′-sequence with fastx34[30] was less than $1e-30$ against the NCBI non-redundant protein data set, (iii) the aligned reading frame was oriented in the plus direction, (iv) the fastx34 alignment of 5′-sequence data was initiated by a methionine residue and (v) a poly $(A)^+$ tail existed in the 3′-sequence. Subsequent to these analyses, clones fulfilling these aforementioned conditions were regarded as full-length soybean cDNAs containing 5′-UTR, CDS and 3′-UTR.

2.8. Scaffold construction

To obtain a non-redundant set of transcripts, we clustered 5′ or 3′ end sequences according to clone names in the CAP3 output. The '.ace' file and the '.singlets' from the CAP3 output were parsed to build scaffolds, which are clusters of sequences representing a unique transcript for which the positional relation and direction of the fragments is implied.

2.9. Annotation of the sequences

After these scaffolds were created, the sequences were queried against several public databases. As a means to estimate similarity to genes from other plants, the sequences were assigned to known information by the BLASTX search (*e*-value $<1e-5$) against protein data sets from TAIR,[31] RAP-DB,[32] JGI Poplar[4] and KOGs.[33,34] The sequences were also submitted to the BLASTN search (*e*-value $<1e-30$) against nucleotide data sets from *L. japonicus* and *M. truncatula* in the NCBI UniGene collections.[35] We subsequently classified the results to show some differences between soybean and other plants. The sequences excluded in all searches were submitted to the InterPro version 4.2 with DBrelease12.1[36] to identify their functional domains. On the other hand, for the detection of novel soybean genes, we used the data sets from the UniGene cluster and complete CDSs of *G. max* in GenBank for the BLASTN analysis. We selected soybean cDNA clones showing their *e*-values $>1e-100$, then they were regarded as novel soybean transcripts.

2.10. UTR detection

For detecting UTR sequences in ESTs, we used the fully sequenced scaffolds of the resulting CAP3 assembly. First, we found potential ORFs in each fully sequenced scaffold with our original perl scripts. We selected the longest of potential ORFs with the sense directions in each scaffold. Then, ESTs which consisted of the scaffold were aligned with sim4 to the selected ORFs, and we obtained the start and end points of CDSs and UTRs in the EST sequences. We omitted some ESTs in which the edge position of the sim4 alignment to the selected ORF was unclear.

3. Results and discussion

3.1. Construction of the soybean full-length-enriched cDNA library

Here, we constructed a soybean full-length-enriched cDNA library using the biotinylated CAP trapper method which is optimized to generate a large percentage of full-length cDNA inserts.[11,22] At first, we prepared 10 kinds of soybean plants in collaboration with several project members from southern to northern parts of Japan (Table 1). Then, we mixed each RNA sample from soybean plants that were grown under various conditions, e.g. drought, salt and cold stresses, Pi starvation, flooding and SCN infection. Finally, the efficiency of our library was 1.1×10^6 pfu with a λFLCIII vector, and we prepared two sets (GMFL01 and GMFL02) of 19 968 cDNA clones (52×384-well plates) from the library.

3.2. The quality of soybean full-length-enriched cDNA library

The outline of the cDNA library construction and data analysis was illustrated in Fig. 1. We prepared two sets of 19 968 clones (GMFL01 and GMFL02) and a total of 39 936 clones were sequenced from 5′ and 3′ ends. Sequences were subsequently trimmed for low quality reads and vector contamination. In addition, we detected and removed 902 SCN transcripts from the obtained sequence data. It is possible that the SCN transcripts could be contaminants from SCN-infected soybean plants (Table 1). After these processes, 37 834 clones (18 670 from GMFL01-set and 19 164 from GMFL02-set) were available for further analyses. We obtained a total of 68 661 EST sequences consisting of 36 512 5′-ESTs and 32 149 3′-ESTs. Both 5′ and 3′ ends of sequences could be assigned for 30 827 clones (14 740 from GMFL01-set and 16 087 from GMFL02-set). However, sequences from only one-

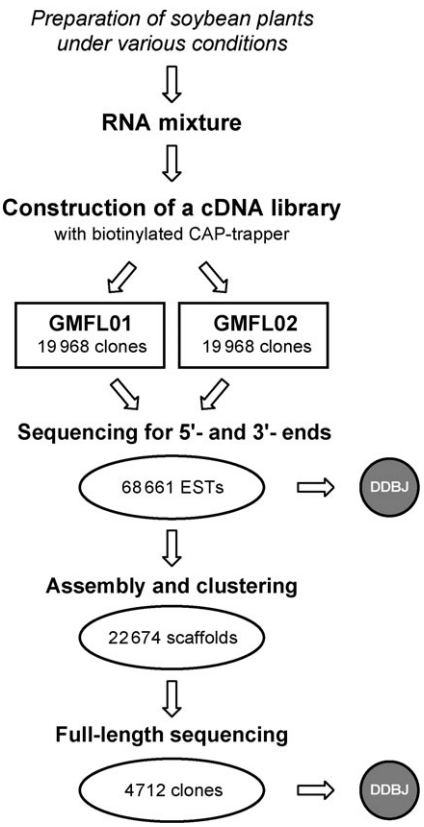

Figure 1. Scheme for construction and data processing of a soybean full-length-enriched cDNA library. We constructed a soybean full-length-enriched cDNA library using a biotinylated CAP-trapper method from multiple sources of soybean plants under various conditions (Table 1). A total of 39 936 cDNA clones (GMFL01 and GMFL02 clone sets) were sequenced and 68 661 both-end sequences were derived. These sequences were deposited to the DDBJ. The sequences were clustered into 22 674 scaffolds. We subsequently selected 4712 clones for full-length cDNA sequences and deposited them to DDBJ.

side sequence could be detected for remaining 7007 clones. All sequences were tagged with unique clone IDs as GMFL01-XX-YYYF/R or GMFL02-XX-YYYF/R. In this nomenclature system, X and Y refer to plate numbers and well numbers in each plate, respectively. The letters F and R indicate forward or reverse sequences, respectively. All sequences were deposited to DDBJ with the accession numbers BW650749–BW684913 and DB955456–DB990717. Clones which had both read sequences showing significant sequence similarity to known proteins were analyzed to confirm whether they contained initiation codons and poly $(A)^+$ tails. The results suggested that 87.7% of the clones contained entire open reading frames in their inserts.

Table 2 summarizes the results from sequence data analysis. The sequences were assembled into 11 036 contigs and 15 255 singletons using CAP3.[28] The CAP3 assembly data were able to further clustered, resulting in the construction of 22 674 scaffolds as non-redundant EST sequences. This set of 22 674 scaffolds was then used as a basic material for further analyses. We built a cluster profile representing the number of clones per scaffold (Fig. 2). These data showed that a large number of scaffolds contain a small number of clones, suggesting that the redundancy of soybean cDNA library was relatively low.

Table 2. Summary of soybean cDNA sequences for assembly and clustering

Groups	Records
Number of initial clones	39 936
Number of available clones	37 834
Number of good sequences	68 661
5′-EST sequences	36 512
3′-EST sequences	32 149
Average trimmed EST length (bp)	526.2
Clones with 5′- and 3′-sequences	30 827
Contigs	11 036
Average contig length (bp)	697.4
Singletons	15 255
Scaffolds	22 674
Max. scaffold size (no. of EST)	199
Average scaffold size (no. of EST)	1.7
Distinct genes	13 526
Putative splicing variants	4325
Full-length sequenced clones	4712
Full-read clones in EST sequences	2580
A total of full-length cDNA sequences	6570

As described above, our cDNA library contains multiple RNA samples from soybean plants that were grown under various conditions (Table 1). The inclusion of RNA samples from multiple growth conditions is advantageous since it increases the variation of transcripts in the library. For example, it is likely that abiotic stress treatments might enhance the proportion of stress-responsive genes in the library. In addition, several specific tissues, such as developing seeds, flower buds and nodules, might increase tissue-specific genes in our collection.

3.3. *Full-length sequences of soybean cDNAs*

If properly constructed, a full-length-enriched cDNA library should contain a high proportion of full-length cDNA clones and maintain complete coding regions, splicing information and 5′- and 3′-UTR sequences. Thus, the entire sequences of full-length cDNA clones will be extremely informative. Here, we determined 4712 full-read cDNA sequences that were selected from the GMFL01 clone set. The majority of

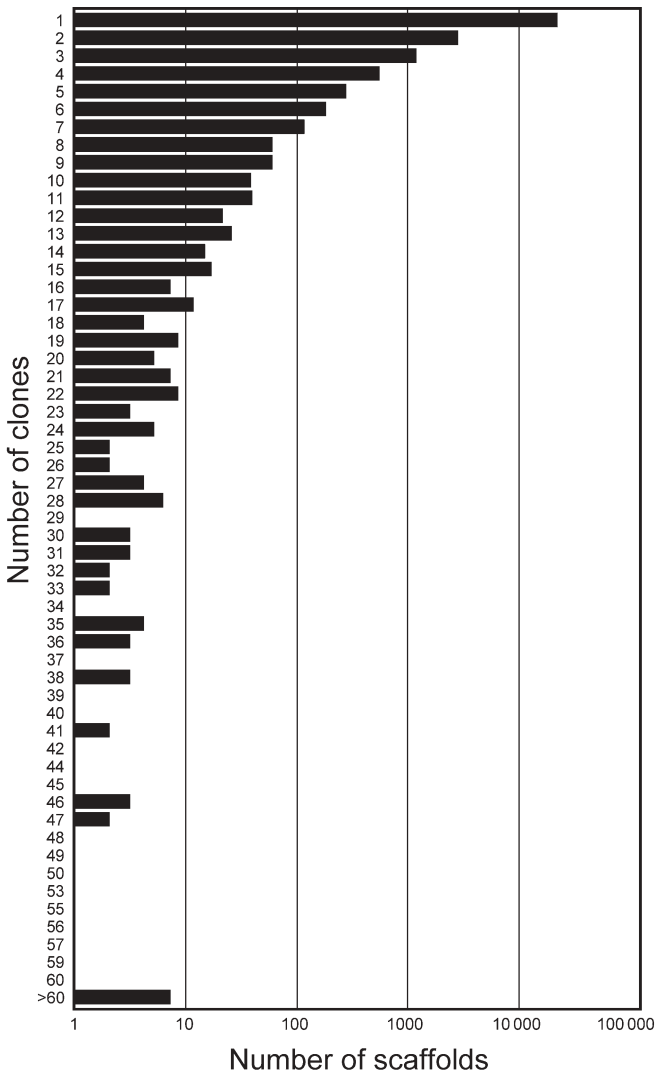

Figure 2. Distribution of numbers of soybean cDNA clones involved in each cluster of sequence assembly. We derived 68 661 sequences from 39 936 soybean cDNA clones, and clustered them into 22 674 scaffolds. Sequence assembly performed by CAP3 reveals a large distribution of the numbers of clones per scaffold.

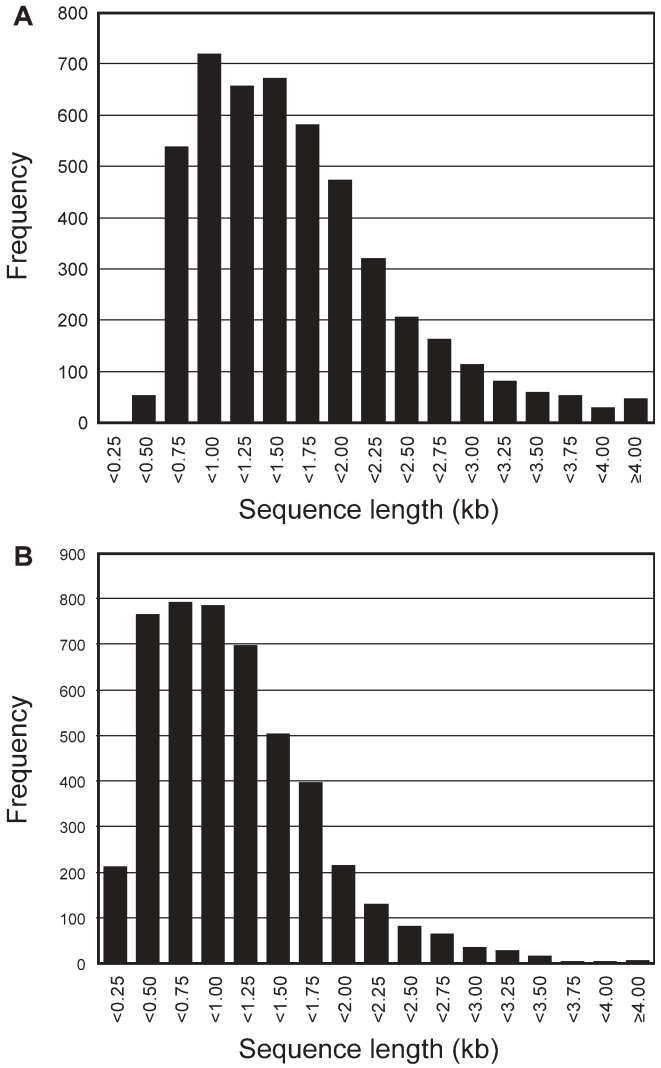

Figure 3. Length distributions of soybean cDNA inserts and ORFs. Sequence length of soybean cDNA inserts (**A**) and its ORFs (**B**) was obtained from a total of 4712 full-length sequences of soybean cDNAs. Used definitions and calculation methods were described in the Materials and methods section.

the sequences consist of stress-responsive genes which were identified by a large set of microarray data from soybean plants under drought, salt, cold stresses and ABA treatments (Todaka et al., unpublished results). The microarray was constructed in the format of a 44K custom oligonucleotide microarray (Agillent Technologies, CA, USA) from a data set of soybean full-length cDNAs and a public EST database (Todaka et al., unpublished results). All sequences of full-length cDNAs were deposited to the DDBJ with accession numbers AK243693–AK246134 and AK285150–AK287419.

Figure 3 shows length distributions of soybean cDNA inserts and ORFs from 4712 full-length cDNA sequences. The full-length sequences ranged from 413 to 7338 bp, and their average size was 1539 bp. Seki et al. (in RIKEN Plant Science Center) determined full-length sequences of 21 005 RIKEN *Arabidopsis* Full-Length (RAFL) cDNA clones the average size of RAFL cDNA inserts was estimated to be 1445 bp[37] (Seki et al., unpublished results). Although the average size of soybean cDNA inserts was slightly longer than *Arabidopsis*, it was similar to other plants, e.g. rice and wheat had the average 1.5 kb cDNA inserts.[17,38] We then extracted the longest ORFs from those full-length sequences and the average size of soybean ORFs was 1042 bp and the median was 933 bp which was similar to *Arabidopsis* (1097 bp) or rice (947 bp).[38] Since these data were in accordance with those from other published data, we have confidence that our

soybean cDNA library successfully captured a wide range of cDNA inserts without any bias. It is likely that our utilization of a phage vector, λFLC, which is optimized to capture long cDNAs[23] maximized our chances for construction of the full-length-enriched soybean library.

The 4712 full-length cDNA sequences were classified into eukaryotic clusters of orthologous groups of proteins (KOGs) in Fig. 4. KOGs include proteins from seven eukaryotic genomes:[33,34] three animals (*Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*), one plant (*Arabidopsis thaliana*), two fungi (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*) and an intracellular microsporidian parasite (*Encephalitozoon cuniculi*). Of these, 3015 (64.0%) were assigned to KOGs by BLASTX (*e*-value $<1e-5$). This result may represent the functional classification of abiotic stress-responsive genes in soybean, because the majority of 4712 clones were selected from microarray analysis using soybean plants under drought, salt,

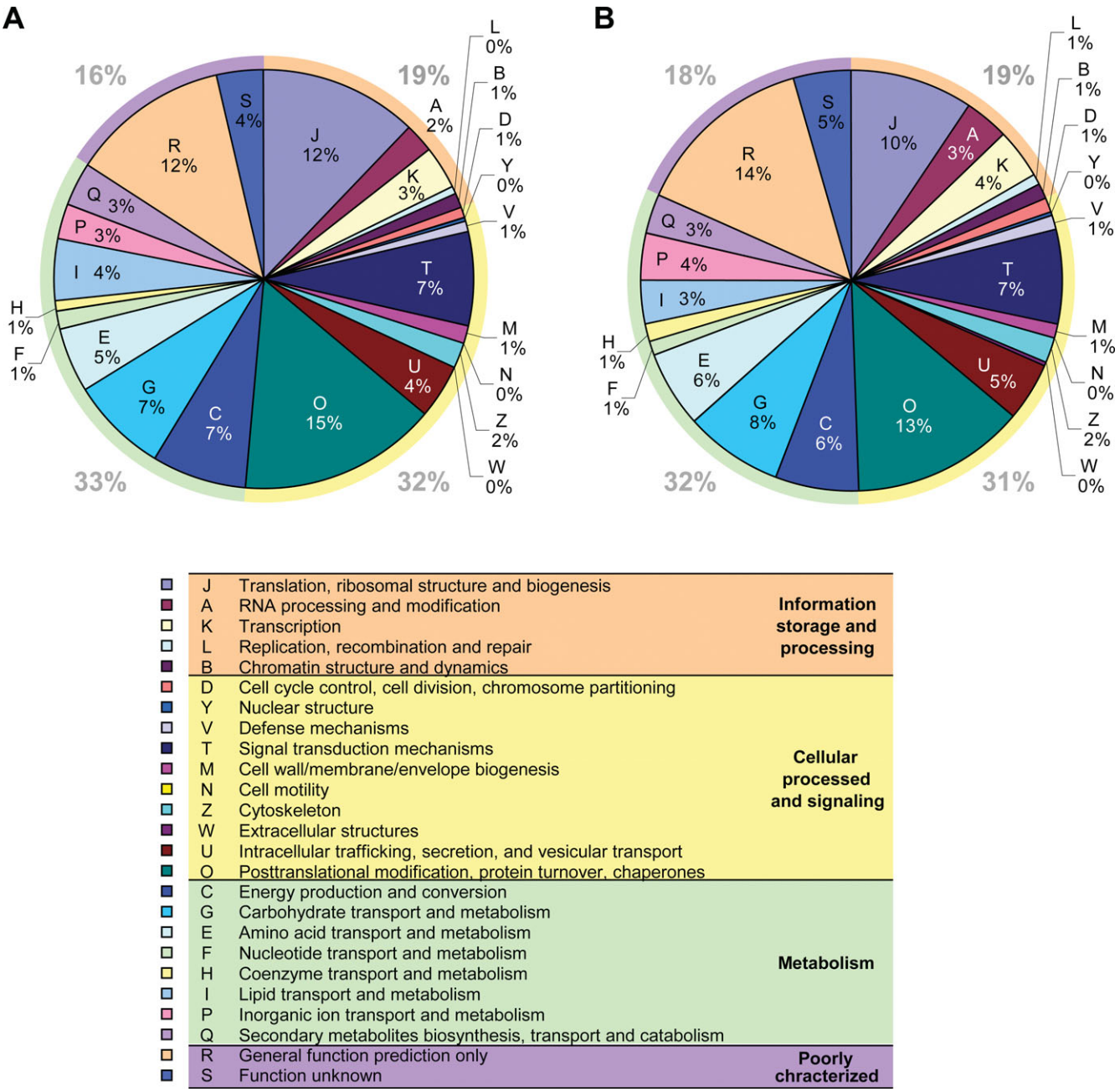

Figure 4. Functional annotation of soybean genes. The 22 674 scaffolds (**A**) and 4712 full-length sequences (**B**) of soybean cDNAs were classified into functional groups by the KOG database.[44] The colors of each functional group are indicated in the table. Graphs are outlined with multi-color frames which represent four subcategories: 'information storage and processing' (light red), 'cellular processing and signaling' (bright yellow), 'metabolism' (greenish brown) and 'poorly characterized' (pink).

cold stresses and ABA treatments as described above. However, Fig. 4B shows that the proportion of each KOG subcategory was similar between 22 674 scaffolds and 4712 full-length cDNA clones. Further analysis is necessary to discuss an expression profile of soybean stress-responsive genes in which signal transduction and metabolic pathways should be activated for cellular responses to abiotic stress.[39] On the whole, the distribution of the functional classification of all scaffolds built by our collected soybean genes was comparable with those identified from a similar study performed with a collection of *Populus nigra*.[16]

In addition to the 4712 full-length sequences, we obtained additional 1858 sequences of full-read cDNAs within 22 674 scaffolds from 5′ and 3′ end sequences assembly. We then compared a total of 6570 full-length sequences with known soybean mRNA sequences containing CDSs that were reported in GenBank as of July 2007. As a result, we found that 5898 sequences were assigned as new full-length sequences of soybean cDNAs. We will continue further sequencing of other full-length cDNAs of our collection in an attempt to identify additional novel full-length sequences.

3.4. Comparative analysis with other plants

By using the BLAST program,[40] soybean cDNA sequences were compared with other plants, *Arabidopsis thaliana*, *Oryza sativa* (rice), *Populus trichocarpa* (poplar), *L. japonicus* and *M. truncatula* (Table 3). The data query was 22 674 non-redundant soybean sequences (scaffolds) derived from 5′ and 3′ ends of soybean cDNA clones. Data subjects were 31 921, 40 041 and 45 555 gene sets of *Arabidopsis*, rice and poplar, respectively (Table 4). The result revealed that 21 047, 19 969 and 21 277 of soybean cDNA clones showed homology to *Arabidopsis*, rice and poplar sequences (BLASTX *e*-value $<1e-5$). We found that 1194 scaffolds (5.3%) did not match with *Arabidopsis*, rice and poplar sequences. Comparative analyses between soybean and *L. japonicus* or *M. truncatula* should be able to confirm whether the 1194 sequences contain legume-specific or soybean-specific genes. However, we found that a large number of soybean sequences did not match with the other legume sequence data sets. This lack of matching probably occurred since the data sets for other legume models have not yet been saturated, e.g. 148 457 records of *L. japonicus* and 232 299 records of *M. truncatula* primarily consisting of ESTs. This problem can be solved in the future when the genomic information is accumulated for soybean, *L. japonicus* or *M. truncatula*. As a result of the comparative analyses, we detected 1085 sequences (4.8%) that did not match with all data subjects (Table 3). Among them, 520 sequences correspond to soybean ESTs and these sequences were therefore identified as true soybean-specific genes. The 520 sequences were submitted to InterPro to obtain functional domain information from soybean-specific genes (Table 5).

Table 3. Comparative analysis of cDNAs between soybean and other plants

Species	No. of records	Hit	No hit	No hit among species	No hit in all searches
Ath	31 921	21 047	1627		
Osa	40 041	19 969	2 705	1194	
Ptr	45 555	21 277	1397		1085
Lja	148 457	13 987	8687	5789	
Mtr	232 299	14 798	7876		

Soybean full-length cDNA sequences from 22 674 scaffolds were submitted to BLASTX search (*e*-value $<1e-5$) against data sets of *Arabidopsis thaliana* (Ath), rice (Osa) or poplar (Ptr), or BLASTN search (*e*-value $<1e-30$) against data sets of *L. japonicus* (Lja) or *M. truncatula* (Mtr). All sequence data were obtained from public databases. The URLs are http://www.arabidopsis.org/ (Ath: 31 921 records), http://rapdb.lab.nig.ac.jp/ (Osa: 40 041 records), http://genome.jgi-psf.org/Poptr1_1/ (Ptr: 45 555 records) and http://www.ncbi.nlm.nih.gov/ (Lja: 148 457 records and Mtr: 232 299 records).

Table 4. List of data sets for comparative analyses with other plants

Data set	Source
A. thaliana proteins	TAIR7 release[a]
O. sativa proteins	RAP1 based on the IRGSP sequence build 3[b]
P. trichocarpa proteins	JGI Populus trichocarpa ver1.1[c]
L. japonicus transcripts	Collected in NCBI (GenBank) as of July 2007 and cleaned from contamination of vector and *Escherichia coli* genomic sequences[d]
M. truncatula transcripts	
G. max transcripts	
G. max mRNA sequences containing CDSs	
Non-redundant proteins	NCBI-nr 28 May 2007 release[d]
Orthologous groups of proteins for eukaryotic	NCBI-KOGs 3 March 2003 release[d]

The version and date of data sets were listed for *Arabidopsis*, rice, poplar, *L. japonicus*, *M. truncatula*, soybean, non-redundant protein sequences and KOGs. Data sets were obtained from public databases as indicated by superscripts.
[a]http://www.arabidopsis.org/.
[b]http://rapdb.lab.nig.ac.jp/.
[c]http://genome.jgi-psf.org/Poptr1_1/.
[d]http://www.ncbi.nlm.nih.gov/.

Table 5. Putative functions of soybean cDNAs which were not homologized to other plants data sets

InterPro Name	InterPro ID	No. of genes
Nodulin	IPR003387	24
Proteinase inhibitor I4, serpin	IPR000215	5
Albumin I	IPR012512	4
Peptidase M, neutral zinc metallopeptidases, zinc-binding site	IPR006025	3
Plant lipid transfer/seed storage/trypsin-alpha amylase inhibitor	IPR003612	3
Aldo/keto reductase	IPR001395	3
Ankyrin	IPR002110	2
Peptidase S1 and S6, chymotrypsin/Hap	IPR001254	2
ATP-dependent DNA ligase	IPR000977	2
Glycine rich	IPR010800	1
Protein of unknown function DUF581	IPR007650	1
Zinc finger, C2H2-type	IPR007087	1
Late embryogenesis abundant protein 3	IPR004926	1
Aminotransferases class-I pyridoxal-phosphate-binding site	IPR004838	1
Immunoglobulin/major histocompatibility complex	IPR003006	1
Phosphotransferase system, HPr serine phosphorylation site	IPR002114	1
Aldehyde dehydrogenase	IPR002086	1
C-5 cytosine-specific DNA methylase	IPR001525	1
Annexin	IPR001464	1
Lipoxygenase	IPR000907	1
Endoplasmic reticulum targeting sequence	IPR000886	1
GPCR, family 2, secretin-like	IPR000832	1
Oxidoreductase, molybdopterin binding	IPR000572	1
Glutelin	IPR000480	1
ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding	IPR000194	1
Glyceraldehyde 3-phosphate dehydrogenase	IPR000173	1
Peptidase, cysteine peptidase active site	IPR000169	1

InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.[36] Among them, 64 sequences were assigned to InterPro ID. The results showed that the total and non-redundant number of InterPro IDs were 66 and 27, respectively. The InterPro ID analysis indicated that the majority of soybean-specific sequences could be assigned to some legume-specific proteins, for example, nodulins and trypsin inhibitors, etc.

We obtained 5′-UTR regions from 8211 sequences and 3′-UTR regions from 7607 sequences using 13 381 clones in the 6570 fully sequenced scaffolds. The average lengths of 5′-UTR and 3′-UTR were 123.13 bp (standard variation $\pm$ 123.31) and 247.52 bp ($\pm$108.14), and the median lengths were 75 and 233 bp, respectively (Fig. 5). In *Arabidopsis* or rice, it was reported that the median lengths of 5′-UTR were 82–88 and 123 bp, respectively [38]. This indicates that 5′-UTR length of soybean was similar to that of *Arabidopsis* rather than that of rice. It is reasonable because both soybean and *Arabidopsis* belong to dicot plants, whereas rice is monocot. Although the detailed reason for this similarity is not known at this time, future comparative studies with additional plants should address whether such shorter 5′-UTR regions are specific to soybean, legumes or dicots.

As a means to identify novel transcripts from our cDNA clones, we compared the 22 674 soybean sequences with known soybean sequences in GenBank to identify new transcripts in our clones. As of July 2007, the GenBank database contains 386 196 records of the soybean nucleotide sequences (ESTs and mRNAs). Subsequent to our sequence analysis, we found that 3022 scaffolds did not match with soybean sequences. As a result, they were regarded as new soybean transcripts. These data also suggest that the soybean full-length cDNA library captures a wide variety of transcripts with relatively low redundancy. In general, redundancy in a cDNA library can be reduced incorporating the

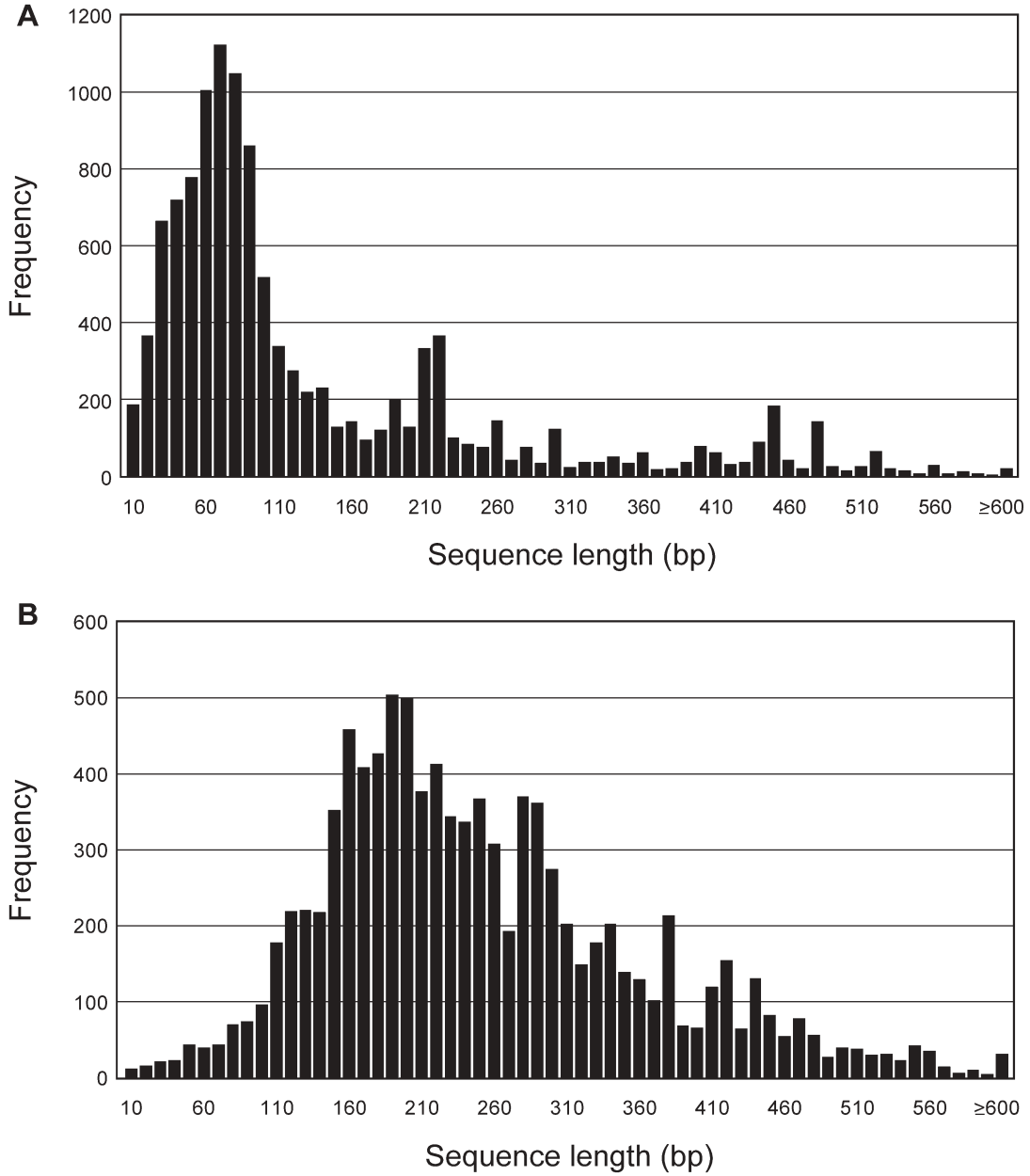

Figure 5. Length distributions of 5′- and 3′-UTR sequences of soybean cDNA clones. The 5′-UTR (**A**) and 3′-UTR sequences (**B**) were derived from 68 661 soybean EST sequences. The definitions and calculation methods were described in the Materials and methods section.

normalization processes.[41] Although our library was not normalized, it successfully produced a large number of cDNAs (93% of soybean UniGene set), suggesting that it should be a good source for obtaining information from non-redundant soybean cDNAs. Various RNA samples, as shown in Table 1, might extend the variation of transcripts in the soybean full-length-enriched library.

3.5. *The value of the soybean full-length-enriched cDNA library and the full-length cDNA collection*

We report here a large number of sequences of soybean cDNAs, and the majority of them contain CDS. The value of this information is expected to increase when it is integrated to whole-genome sequence of soybean. For example, 5′-UTR data in full-length cDNA sequences will help to identify the promoter sequences, or full-length sequences will allow us to outlook for gene structures of paralogs or gene families which resemble each other in soybean genome. Actually, we compared our cDNA sequence data (22 674 scaffolds) with predicted 62 199 transcripts from the soybean draft genome (JGI-DOE; http://www.phytozome.net/soybean), but 20.2% of cDNA sequences showed low similarity. Also, we analyzed a public data set of soybean

Figure 6. The web-based interface to the soybean full-length cDNA database. The query set should be prepared as nucleotide and peptide sequences, or keywords for gene functions/annotations. This website includes a full-set of BLAST programs against five data sets of nucleotide—*G. max* cDNA (RIKEN), *G. max* mRNA (GenBank), *G. max* (UniGene), *L. japonicus* (UniGene) and *M. truncatula* (UniGene)—and four data sets of peptide—UniProt-TrEMBL plants, *Arabidopsis thaliana* (TAIR), *Oryza sativa* (RAP-DB) and *Populus trichocarpa* (JGI). These tools can be accessed from the following URL: http://rsoy.psc.riken.jp/

transcripts (Ver.2 2006-09-28 release, 116 965 data from TIGR Plant Transcript Assemblies DB;[42] http://plantta.tigr.org/), and we found that 16.0% of TIGR's data showed low similarity to the predicted transcripts. These results indicate that the prediction of soybean gene structures is still incomplete now, and our cDNA collection will be a great help for annotation of soybean genome data.

Another effective approach using full-length cDNAs is to analyze gene functions and structures. First, the utilization of full-length cDNAs enables us to demonstrate an expression pattern of single transcript in detail. This is possible because a full-length cDNA represents a single splice variant from each transcription unit. Secondly, they are easy to use for a gene transfer system and lead us toward better understanding of gene functions through gain-of-function or loss-of-function analyses. Recently, a large-scale screening system of transgenic plants using full-length cDNA clones and agrobacterium-mediated transformation had been established, i.e. the FOX-hunting system (Full-length cDNA Over-eXpressing gene hunting system) enables us to survey a lot of gene functions within a short period[43]. In addition, it will be useful for constructing a microarray on which specific oligonucleotide probes for each soybean cDNA clones are printed and the system clearly demonstrated a transcriptome in soybean (Todaka et al., unpublished results).

Our information from soybean full-length cDNA clones could also be useful for generating molecular markers. First, SSCP or CAPS markers can be developed by PCR based on 5′- and 3′-sequences of cDNA clones. Furthermore, full-read cDNA sequences are likely to be an important material for finding SNP or SSR markers. Actually, the development of molecular markers has already been conducted by using our soybean full-length cDNA collection (Dr K. Harada, personal communication). Our collection will be also useful for comparative genomics such as synteny analysis of soybean and other plants, for example, *L. japonicus*. Now we have a lot of information of genomic resources from soybean and *L. japonicus*, which will provide us various opportunities to accelerate understanding of legume systems.

Here we reported on the soybean full-length cDNA resource which will be arranged and maintained by each cDNA clone. We also developed a web-based interface of the soybean full-length cDNA database which will enable researchers to gain easy access to the data (http://rsoy.psc.riken.jp/). The website contains a BLAST program which allows a search

of soybean full-length cDNA clones based on sequence similarity, as well as a keyword search against gene functions or annotations (Fig. 6). The information and resources of the soybean full-length cDNA clones will be distributed from the National Bioresource Project for *Lotus*/*Glycine* in Japan (http://www.legumebase.agr.miyazaki-u.ac.jp/). As described above, full-length cDNAs are high potential resources because they can provide various outputs for post-genomic life sciences. We hope that the soybean full-length cDNA collection will be useful in various situations of legume research.

Acknowledgements: We thank all of the technical staff of the Sequencing Technology Team at RIKEN Genomic Sciences Center, Gene Discovery Research Team, Integrated Genome Informatics Research Unit and Plant Genomic Network Research Team at RIKEN Plant Science Center, and Biological Resources Division at JIRCAS for their assistance.

Funding

This work was supported by grants from JIRCAS Comprehensive Research Project ('Comprehensive studies on development of sustainable soybean production technology in South America'), the Grant-in-Aid for Scientific Research (17018005, 18017004 and 18700106) from MEXT, and RIKEN Plant Science Center, Japan. Full-read sequencing of 4712 cDNAs was supported by National BioResource Project for Lotus/Glycine from MEXT, Japan.

References

1. Hill, J., Nelson, E., Tilman, D., Polasky, S. and Tiffany, D. 2006, Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels, *Proc. Natl Acad. Sci. USA*, **103**, 11206–11210.
2. The Arabidopsis Genome Initiative. 2000, Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*, *Nature*, **408**, 796–815.
3. Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. 2002, A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica), *Science*, **296**, 92–100.
4. Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. 2006, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science*, **313**, 1596–1604.
5. Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002, A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica), *Science*, **296**, 79–92.
6. Tabata, S. 2002, Impact of genomics approaches on plant genetics and physiology, *J. Plant Res.*, **115**, 271–275.
7. Shoemaker, R. C., Schlueter, J. and Doyle, J. J. 2006, Paleopolyploidy and gene duplication in soybean and other legumes, *Curr. Opin. Plant Biol.*, **9**, 104–109.
8. Nelson, R. T. and Shoemaker, R. 2006, Identification and analysis of gene families from the duplicated genome of soybean using EST sequences, *BMC Genomics*, **7**, 204.
9. Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., et al. 2008, Genome structure of the legume, *Lotus japonicus*, *DNA Res.*, 10.1093/dnares/dsn008.
10. Shoemaker, R., Keim, P., Vodkin, L., Retzel, E., Clifton, S. W., Waterston, R., Smoller, D., Coryell, V., Khanna, A., Erpelding, J., et al. 2002, A compilation of soybean ESTs: generation and analysis, *Genome*, **45**, 329–338.
11. Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., et al. 1996, High-efficiency full-length cDNA cloning by biotinylated CAP trapper, *Genomics*, **37**, 327–336.
12. Maruyama, K. and Sugano, S. 1994, Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides, *Gene*, **138**, 171–174.
13. Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. and Sugano, S. 1997, Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library, *Gene*, **200**, 149–156.
14. Jia, J., Fu, J., Zheng, J., Zhou, X., Huai, J., Wang, J., Wang, M., Zhang, Y., Chen, X., Zhang, J., et al. 2006, Annotation and expression profile analysis of 2073 full-length cDNAs from stress-induced maize (*Zea mays* L.) seedlings, *Plant J.*, **48**, 710–727.
15. Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., et al. 2003, Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice, *Science*, **301**, 376–379.
16. Nanjo, T., Futamura, N., Nishiguchi, M., Igasaki, T., Shinozaki, K. and Shinohara, K. 2004, Characterization of full-length enriched expressed sequence tags of stress-treated poplar leaves, *Plant Cell Physiol.*, **45**, 1738–1748.
17. Ogihara, Y., Mochida, K., Kawaura, K., Murai, K., Seki, M., Kamiya, A., Shinozaki, K., Carninci, P., Hayashizaki, Y., Shin, I. T., et al. 2004, Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags, *Genes Genet. Syst.*, **79**, 227–232.
18. Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., et al. 2002, Functional annotation of a full-length *Arabidopsis* cDNA collection, *Science*, **296**, 141–145.
19. Haas, B. J., Volfovsky, N., Town, C. D., Troukhan, M., Alexandrov, N., Feldmann, K. A., Flavell, R. B., White, O. and Salzberg, S. L. 2002, Full-length messenger RNA

sequences greatly improve genome annotation, *Genome Biol.*, **3**, 0029.0021–0029.0012.
20. Umezawa, T., Mizuno, K. and Fujimura, T. 2002, Discrimination of genes expressed in response to the ionic or osmotic effect of salt stress in soybean with cDNA–AFLP, *Plant Cell Environ.*, **25**, 1617–1625.
21. Funatsuki, H., Kawaguchi, K., Matsuba, S., Sato, Y. and Ishimoto, M. 2005, Mapping of QTL associated with chilling tolerance during reproductive growth in soybean, *Theor. Appl. Genet.*, **111**, 851–861.
22. Seki, M., Carninci, P., Nishiyama, Y., Hayashizaki, Y. and Shinozaki, K. 1998, High-efficiency cloning of *Arabidopsis* full-length cDNA by biotinylated CAP trapper, *Plant J.*, **15**, 707–720.
23. Carninci, P., Shibata, Y., Hayatsu, N., Itoh, M., Shiraki, T., Hirozane, T., Watahiki, A., Shibata, K., Konno, H., Muramatsu, M., et al. 2001, Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis, *Genomics*, **77**, 79–90.
24. Dean, F. B., Nelson, J. R., Giesler, T. L. and Lasken, R. S. 2001, Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification, *Genome Res.*, **11**, 1095–1099.
25. Ewing, B. and Green, P. 1998, Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Res.*, **8**, 186–194.
26. Ewing, B., Hillier, L., Wendl, M. C. and Green, P. 1998, Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res.*, **8**, 175–185.
27. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. and Miller, W. 1998, A computer program for aligning a cDNA sequence with a genomic DNA sequence, *Genome Res.*, **8**, 967–974.
28. Huang, X. and Madan, A. 1999, CAP3: a DNA sequence assembly program, *Genome Res.*, **9**, 868–877.
29. Gordon, D., Abajian, C. and Green, P. 1998, Consed: a graphical tool for sequence finishing, *Genome Res.*, **8**, 195–202.
30. Pearson, W. R., Wood, T., Zhang, Z. and Miller, W. 1997, Comparison of DNA sequences with protein sequences, *Genomics*, **46**, 24–36.
31. Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., et al. 2003, The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community, *Nucleic Acids Res.*, **31**, 224–228.
32. Ohyanagi, H., Tanaka, T., Sakai, H., Shigemoto, Y., Yamaguchi, K., Habara, T., Fujii, Y., Antonio, B. A., Nagamura, Y., Imanishi, T., et al. 2006, The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. japonica genome information, *Nucleic Acids Res.*, **34**, D741–D744.
33. Tatusov, R. L., Koonin, E. V. and Lipman, D. J. 1997, A genomic perspective on protein families, *Science*, **278**, 631–637.
34. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., et al. 2003, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.
35. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., et al. 2007, Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.*, **35**, D5–D12.
36. Zdobnov, E. M. and Apweiler, R. 2001, InterProScan—an integration platform for the signature-recognition methods in InterPro, *Bioinformatics*, **17**, 847–848.
37. Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Palm, C. J., Southwick, A. M., Wu, H. C., Kim, C., Nguyen, M., et al. 2003, Empirical analysis of transcriptional activity in the *Arabidopsis* genome, *Science*, **302**, 842–846.
38. Alexandrov, N. N., Troukhan, M. E., Brover, V. V., Tatarinova, T., Flavell, R. B. and Feldmann, K. A. 2006, Features of *Arabidopsis* genes and genome discovered using full-length cDNAs, *Plant Mol. Biol.*, **60**, 69–85.
39. Shinozaki, K., Yamaguchi-Shinozaki, K. and Seki, M. 2003, Regulatory network of gene expression in the drought and cold stress responses, *Curr. Opin. Plant Biol.*, **6**, 410–417.
40. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
41. Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. and Hayashizaki, Y. 2000, Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes, *Genome Res.*, **10**, 1617–1630.
42. Childs, K. L., Hamilton, J. P., Zhu, W., Ly, E., Cheung, F., Wu, H., Rabinowicz, P. D., Town, C. D., Buell, C. R. and Chan, A. P. 2007, The TIGR Plant Transcript Assemblies database, *Nucleic Acids Res.*, **35**, D846–D851.
43. Ichikawa, T., Nakazawa, M., Kawashima, M., Iizumi, H., Kuroda, H., Kondou, Y., Tsuhara, Y., Suzuki, K., Ishikawa, A., Seki, M., et al. 2006, The FOX hunting system: an alternative gain-of-function gene hunting technique, *Plant J.*, **48**, 974–985.
44. Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., et al. 2004, A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes, *Genome Biol.*, **5**, R7.