DNA Research **16**, 81–89, (2009) doi:10.1093/dnares/dsn034

Development of 5006 Full-Length CDNAs in Barley: A Tool for Accessing Cereal Genomics Resources

Kazuhiro Sato[1,*], Tadasu Shin-I[2], Motoaki Seki[3], Kazuo Shinozaki[3], Hideya Yoshida[1], Kazuyoshi Takeda[1], Yukiko Yamazaki[2], Matthieu Conte[4], and Yuji Kohara[2]

Research Institute for Bioresources, Okayama University, Kurashiki 710-0046, Japan[1]; National Institute of Genetics, Mishima 411-8540, Japan[2]; Plant Science Center, RIKEN, Yokohama 230-0045, Japan[3] and Crop Research Informatics Laboratory, International Rice Research Institute, PO Box 933, Manila 1099, Philippines[4]

(Received 5 August 2008; accepted 14 December 2008; published online 15 January 2009)

Abstract

A collection of 5006 full-length (FL) cDNA sequences was developed in barley. Fifteen mRNA samples from various organs and treatments were pooled to develop a cDNA library using the CAP trapper method. More than 60% of the clones were confirmed to have complete coding sequences, based on comparison with rice amino acid and UniProt sequences. Blastn homologies (E<1*E*-5) to rice genes and *Arabidopsis* genes were 89 and 47%, respectively. Of the 5028 possible amino acid sequences derived from the 5006 FLcDNAs, 4032 (80.2%) were classified into 1678 GreenPhyl multigenic families. There were 555 cDNAs showing low homology to both rice and *Arabidopsis*. Gene ontology annotation by InterProScan indicated that many of these cDNAs (71%) have no known molecular functions and may be unique to barley. The cDNAs showed high homology to Barley 1 GeneChip oligo probes (81%) and the wheat gene index (84%). The high homology between FLcDNAs (27%) and mapped barley expressed sequence tag enabled assigning linkage map positions to 151–233 FLcDNAs on each of the seven barley chromosomes. These comprehensive barley FLcDNAs provide strong platform to connect pre-existing genomic and genetic resources and accelerate gene identification and genome analysis in barley and related species.

Key words: full-length cDNA; *Hordeum vulgare*; mRNA; gene ontology

1. Introduction

Cultivated barley (*Hordeum vulgare* L.) is a true diploid with genome size estimated to be ca. 5000 Mb.[1] In order to approach this large genome, several projects have generated significant numbers of expressed sequence tags (ESTs) (ca. 500 000) (see HarvEST database http://harvest.ucr.edu/). These large numbers of ESTs may represent most of the barley genome's transcripts.

Sato et al. (submitted for publication) assigned linkage map positions to 2890 non-redundant 3′ ESTs, providing the densest, reliable barley map available. Other projects have also mapped more than 1000 barley ESTs[2] (see also http://harvest.ucr.edu/), but these are consensus maps. Barley ESTs were also mapped on chromosome deletion stocks to estimate their physical locations.[3,4] These mapped ESTs will promote the analysis of barley genome structure and are an essential foundation for genome sequencing based on high quality genome libraries.[5,6]

Quality-controlled barley EST sequences were used to develop a GeneChip oligo-microarray[7] for analyzing global expression of transcripts in different organs and/or various growth stages.[8] However, EST-based microarrays often lack complete gene

Edited by Satoshi Tabata

* To whom correspondence should be addressed. Tel. +81 86-434-1244. Fax. +81 86-434-1244. E-mail: kazsato@rib.okayama-u.ac.jp

annotation due to the lower homology between partial sequences of cDNAs (ESTs) and the reference-sequenced plant genomes (e.g. rice and *Arabidopsis*). Full-length (FL) cDNA sequences are essential for annotation of genome sequences via transcript mapping.

There are several procedures for developing an FLcDNA library. Of those, the biotinylated CAP trapper method gives a high level of complete coding sequences (CDSs) in FLcDNAs.[9,10] Using this technique, a significant amount of plant FLcDNA sequences was generated. The first comprehensive (14 668) set of FLcDNA sequences was published for *Arabidopsis thaliana*.[11] These FLcDNAs traced to 19 different mRNA samples, covering most of the transcripts in this model plant species for gene annotation. The rice (*Oryza sativa*) FLcDNA project was the second in plants and released 28 469 sequences.[12] In both cases, released genome sequences were already available so that the FLcDNAs assisted in mapping transcripts.[13]

The number of organs in plants is limited, compared with animals, which have various organs with specific profiles of gene expression. The 'body map' described the spectrum of transcripts from each organ of the human body collected from organ-specific mRNA samples.[14] The FLcDNA projects in both *Arabidopsis*[11] and rice[12] used stress conditions rather than organs to achieve higher transcript coverage. The stress induction of transcripts is a frequently used approach in plant EST projects including barley (http://pgrc.ipk-gatersleben.de/cr-est/liball.php) and poplar.[15] Even if the stress conditions are similar, responses to specific stresses could be different among plant species.

Barley has special features compared with other plant species. It was one of the earliest crops domesticated in the Near East,[16] and it is well adapted to semi-arid conditions. It was also known to be more tolerant to salt than wheat in ancient Mesopotamia,[17] but it is the cereal crop most sensitive to aluminum toxicity under acid soil conditions.[18]

Within the evolutionary tree of the grass family (Poaceae), which involves many important cereal species, e.g. rice and maize, barley (*H. vulgare* L.) belongs to the tribe Triticeae. This group includes important crop species such as wheat (*Tiriticum aestivum* L.) and rye (*Secale cereale* L.).[19] The genetic relatedness between barley and other Triticeae species, especially wheat, is well confirmed based on both genetic nucleotide sequences and intergeneric hybridization.[20] Triticeae crop species may have a common diploid ancestor with seven pairs of chromosomes, as was well demonstrated by the direct use of primers from barley ESTs to develop a diploid wheat genetic map.[21] The relatively high genomic similarity between barley and rice is known since the early synteny analyses based on restriction fragment length polymorphism markers,[22,23] and it is used to isolate genes of importance in barley.[24,25] Thus, barley cDNA sequences are expected to show high similarity with wheat cDNA sequences and reasonably high similarity with rice cDNA sequences.

In the present study, we collected a significant number of barley FLcDNAs by using the biotinylated CAP trapper method.[9,10] The FLcDNA sequences were compared with rice and *Arabidopsis* genes, and we evaluated the spectrum of transcripts represented by Gene Ontology (GO) mapped by InterProScan. The FLcDNA sequences are also compared with transcripts from barley and wheat in order to obtain access to the genomic and genetic resources available in these species.

2. Materials and methods

2.1 Plant materials

Cultivated barley (*H. vulgare* L.) cv. Haruna Nijo was used to isolate all the RNA samples used in this study. The types of samples are listed in Table 1.

For heat and cold stress treatments, plants were grown on water agar in a growth chamber at 20°C with a 16 h photoperiod and a light intensity 320 μmol/m^2/s. The first leaf stage plants were moved to treatment chambers with fluorescent light and exposed to either 40°C (heat treatment) for 24 h or -1°C for 24 h (cold treatment).

All the other stress-treated plants were grown in hydroponic culture. Seed samples were placed on the moist filter paper in Petri dishes at 20°C in the dark for 3 days. Seedlings were then mounted on plastic frames with strips of polyurethane foam. Frames were placed over 35 L plastic tanks containing a nutrient solution consisting of the following components (μM): Ca, 1000; Mg, 400; K, 1000; NO_3, 3400; NH_4, 600; PO_4, 100; SO_4, 401.1; Cl, 78; Na, 40.2; Fe, 20; B, 23; Mn, 9; Zn, 0.8; Cu, 0.30 and Mo, 0.1. Iron was supplied as Fe-EDTA prepared from equimolar amounts of $FeCl_3$ and Na_2EDTA. Throughout the experiment, solutions were constantly aerated. Plants were grown in a growth chamber at 20°C with 16 h photoperiod and a light intensity of 320 μmol/m^2/s. After 3 days in the nutrient solution, the solution was completely changed, as described below for each stress. In the Al stress treatment, plants were exposed to 30 μM of $AlK(SO_4)_2 \cdot 12H_2O$, which was added to the complete nutrient solution, adjusted to pH 4.3. In the NaCl stress treatment, 0.1 M of NaCl was added to the complete nutrient solution, adjusted to pH 6.0. For the drought treatment, plants were moved from the solution culture to dry filter paper in the same growth chamber.

Table 1. Tissues and stages used for generating an FLcDNA library of barley cv. Haruna Nijo

Stress-treated samples			
Treatment	Organ	Treatment period	Condition
$AlK(SO_4)\cdot 12H_2O$ (30 μM)	Seedling root	6 h	Hydroponic, light
NaCl (0.1 M)	Seedling leaf	6 h	Hydroponic, light
NaCl (0.1 M)	Seedling root	6 h	Hydroponic, light
cold (−1°C)	Seedling leaf	24 h	Agar, light
Heat (40°C)	Seedling leaf	24 h	Agar, light
Wound (5 cm cut)	Seedling leaf	12 h	Hydroponic, light
Drought on filter paper	Seedling leaf and root	2 h	Hydroponic, light
Organ samples			
Stage	Organ	Day of sampling	Condition
Germinating seed	Entire plant	2nd day	20°C
Germinating seed	Embryo	2nd day	20°C
Seedling	Shoot	5th day	20°C, dark
Heading	Upper three leaf blades	120th day	Filed grown
Booting	Young spike (3–5 cm)	120th day	Filed grown
Vegetative stage	Culm	60th day	Filed grown
Vegetative stage	Root	60th day	Filed grown
Maturing	Spike	140th day	Filed grown

For the wounding stress, seedling leaves were cut for 5 cm from the top to the bottom of the leaf blade.

Organ-specific samples were collected at different plant growth stages. Germinated seed samples were collected from entire plants 2 days after germination. Shoots and embryos were collected from 5-day-old seedlings grown on moist filter paper in a Petri dish at 20°C in the dark. Both whole root and whole shoot samples were collected from 60-day-old plants grown under standard field conditions in Okayama University. Leaf blades of the upper three leaves and young spikes (3–5 cm) were collected at the stage of flag leaf emergence. Spikes at a maturing stage (20 days after flowering) were also collected.

2.2 *RNA preparation and cDNA library construction*

Total RNA was prepared from each of the samples and mixed as described in Table 1, for a total amount of 4 mg. Each sample was ground with a mortar and pestle in the presence of liquid nitrogen. The ground powder was then mixed with 5 volumes of solution (4 M guanidine thiocyanate, 25 mM trisodium citrate dehydrate, 0.5% sodium *N*-lauroyl sarcosynate, 0.1 M 2-mercaptoethanol). The cellular debris was pelleted out in microtubes (14 000 rpm for 10 min at 4°C). The supernatant was layered on top of 1.1 mL of 5.7 M CsCl cushion solution (5.7 M CsCl, 0.1 M EDTA) to create a step gradient and centrifuged for 16 h in a SW-60Ti rotor (Beckman, CA, USA) at 35 000 rpm in 20°C. The RNA pellet was dissolved in 10 mM Tris–HCl (pH 7.5), 5 mM EDTA (pH 7.5). The supernatant was mixed with an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) and centrifuged at 14 000 rpm for 5 min at 4°C. The supernatant was transferred to a new tube, and the lower phase was mixed with 10 mM Tris–HCl (pH 7.5), 5 mM EDTA (pH 7.5) and centrifuged at 14 000 rpm for 5 min at 4°C. The supernatant and the previous supernatant were mixed with the equal volume of chloroform and centrifuged at 14 000 rpm for 5 min at 4°C. The upper phase was collected and mixed with 1/3 volume of 8 M LiCl. The RNA was precipitated at 4°C for 30 min and centrifuged at 14 000 rpm for 30 min. The pellet was washed with 70% ethanol and centrifuged at 14 000 rpm for 10 min at 4°C. The pellet was dried with centrifugal concentrator and dissolved in diethylpyrocarbonate-treated water. The tube was shaken using a tube mixer for 10 min, and the absorbance was measured. RNA samples were stored at −80°C until use.

An FLcDNA library was constructed essentially as reported previously[9,10] by biotinylated CAP trapper using trehalose-thermoactivated reverse transcriptase.[26] The mRNA isolated from RNA samples was quality checked and used for first-strand cDNA synthesis. After oxidation, biotinylation and RNase digestion of first-strand cDNA/mRNA hybrids, FLcDNA/RNA hybrids were captured on magnetic beads. RNA was removed by alkaline treatment to collect first-strand FLcDNA. The oligo(dG)-tailed first-strand cDNA was used for second-strand cDNA synthesis. The cDNA was restricted with *Bam*HI and *Xho*I. After purification, cDNA was cloned into a pFLC-III vector.

2.3 DNA extraction and sequencing

Plasmid DNA was extracted with a multiscreen plasmid extraction kit (Millipore) and then purified by precipitation with polyethylene glycol. DNA sequences were determined using the dye terminator cycle sequencing method with ABI 3700 sequencer. DNA clones were subjected to single-pass sequencing from both 3′- and 5′-ends of the cDNA. 3′ ESTs were assembled by phrap (http://www.phrap.org/) to develop contigs and to identify singlets. From each member of each contig, the corresponding 5′ sequences were assembled to align the sequences on 5′-end. When more than two 5′ assembled sequences were grouped, they were assumed to be different transcripts. The clone with the most extended sequence at the 5′-end was assumed as a representative clone to be sequenced. Individual 3′ singlets with 5′ sequences were also assumed to be non-redundant FLcDNA clones to be sequenced. All the representative clones were cycle sequenced by ABI 3700. Primer walking was used to sequence larger insert clones.

2.4 Procedure for sequence annotation

InterProScan version 4.3 (http://www.ebi.ac.uk/Tools/InterProScan/) was installed on an eight-set PC server, using subprograms of coils, blastprodom, superfamily, seg, scanregexp, profilescan, hmmtigr, hmmsmart, hmmpir, hmmpfam, gene3d and fprintscan. The versions of InterPro databases were release 14 for a total of 5006 sequences and release 16.1 for the selected 555 sequences. A PC cluster was established by the software OSCAR 5.0 (http://oscar.openclustergroup.org/), and the FLcDNA sequences were distributed on each server. The output from InterProScan was analyzed to obtain GO categories of each sequence. GO terms in the second hierarchy of the GO database and the GO edit.obo file (date: 19:12:2007 10:07) were used as a top parent of each category. GO terms in the second hierarchy of the GO database described in the edit.obo file (date: 19:12:2007 10:07), listed in the 'GO term' of Table 2, were used as a top parent of each category. Categories of a GO term were defined as a set of the top parents that are accessible from the GO term through the GO graph structure. A set of categories of all GO terms obtained by InterProScan was calculated for each FLcDNA, and the number of clones in the category was counted for all of the categories. Blast homology was analyzed on an in-house blast server installed with the software package Dynaclust (Dynacom Co.).

3. Results and discussion

3.1 Clone selection and insert sizes of FLcDNAs

A total of 45 897 5′ reads and 47 143 3′ reads were sequenced from both ends of cDNA clones. All the 3′-end sequences were assembled by phrap

Table 2. InterProScan analysis and molecular function GO for 5006 FL barley cDNA clones and 555 selected clones showing low blastn homology (E>1*E*-5) with both rice and *Arabidopsis* genes

Function	GO term	All FLcDNA		Low homology FLcDNA to rice and *Arabidopsis*	
		No. of clones	%	No. of clones	%
Total		5006	100.0	555	100.0
InterProScan results		4980	99.5	535	96.4
No GO clones		1824	36.4	375	67.6
Categories[a]					
Binding	GO:0005488	1632	32.6	86	15.5
Catalytic activity	GO:0003824	1586	31.7	61	11.0
Transporter activity	GO:0005215	243	4.9	13	2.3
Structural molecule activity	GO:0005198	206	4.1	11	2.0
Transcription regulator activity	GO:0030528	107	2.1	8	1.4
Antioxidant activity	GO:0016209	52	1.0	2	0.4
Enzyme regulator activity	GO:0030234	48	1.0	15	2.7
Translation regulator activity	GO:0045182	33	0.7	1	0.2
Molecular transducer activity	GO:0060089	33	0.7	2	0.4
Nutrient reservoir activity	GO:0045735	14	0.3	1	0.2
Motor activity	GO:0003774	5	0.1	0	0.0
Metallochaperone activity	GO:0016530	1	0.0	0	0.0

[a] Categories with namespace of molecular function are displayed.

(http://www.phrap.org/), and a total of 4853 contigs and 1613 singlets were identified from the assembly. For each of the 4853 3′ contig members, respective 5′-end sequences were assembled, although 69 3′ contigs did not have any 5′-end sequences. Singlets were also checked for the availability of 5′-end sequences. A total of 4596 contigs and 1459 singlets were available for 5′-end sequences. These 6055 cDNA clones were served as a basis for complete sequencing of inserts, 5006 of which were successfully sequenced. These cDNA sequences provide for an estimated 15% of the genes of barley. This estimate is based on the number of non-redundant barley transcripts (32 690) by CAP3 assembly,[27] using 3′-end sequences of FLcDNA libraries (Sato, unpublished data).

After trimming the vector sequences, insert sizes of 5006 clones ranged from 167 to 6780 bp, with an average size of 1474 bp (Fig. 1), which is reasonably large compared with rice (30–16 311 bp; average =1107 bp) (rep_orf_nuc.fa: $N = 30\,192$, http://rapdb.dna.affrc.go.jp/). This indicates that the most of the cDNA clones were sufficiently large to include open reading frame (ORF) sequences.

3.2 *Comparison with rice and Arabidopsis gene*

The homologies of the FLcDNA sequences with rice (rap2_rep_nuc: $N = 31\,439$) and *Arabidopsis* (TAIR7_seq_20070320: $N = 37\,022$) were determined by blastn[28] (Fig. 2). The threshold of E <1*E*-5 revealed 4451 and 1798 FLcDNAs with homology to rice and *Arabidopsis*, respectively. With the threshold of E <1*E*-30, 3909 FLcDNAs showed homology to rice. The numbers of FLcDNAs showing homology to both species, only rice and only *Arabidopsis* were 1788 (36%), 2663 (53%) and 10 (0.002%), respectively. Other 555 (11%) FLcDNAs did not show homology to either rice or *Arabidopsis*.

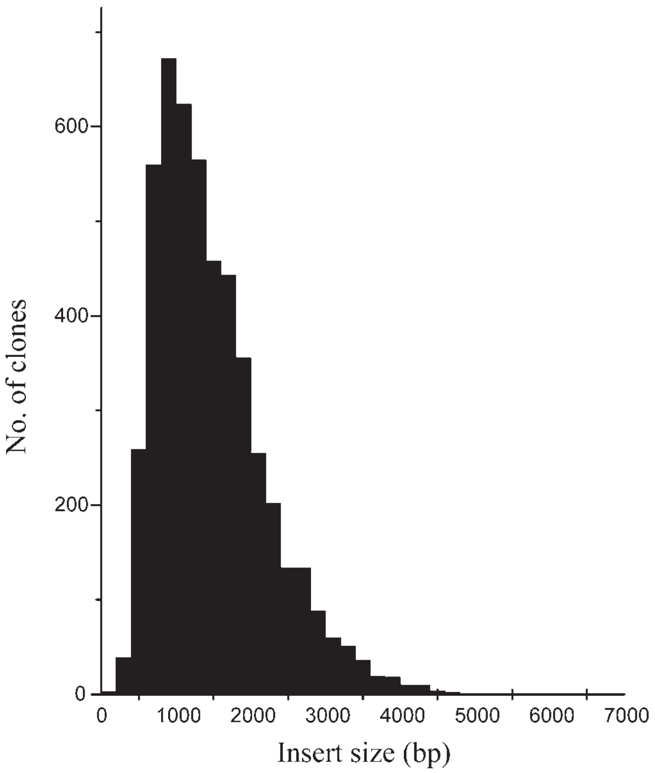

Figure 1. Distribution of insert sizes of 5006 FL barley clones. Average insert size is 1474 bp, with the range of 167–6780 bp.

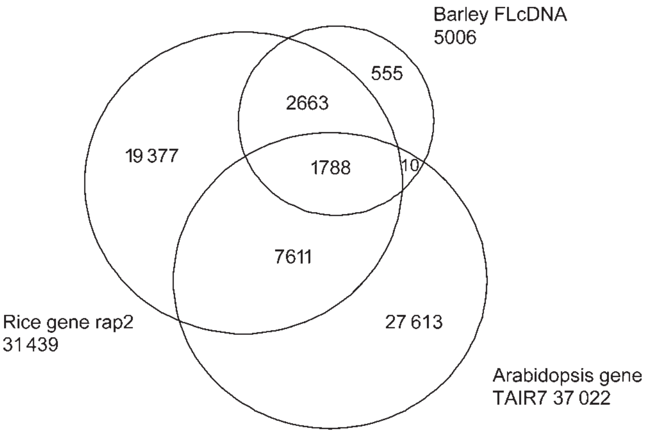

Figure 2. Blastn search among barley FLcDNAs, rice genes and *Arabidopsis* genes (E<1*E*-5).

Subsequently, homologies (E <1*E*-5) to rice representative of ORF amino acid sequences (1397) (rap2_rep_orf_aa: $N = 30\,192$) and *Arabidopsis* CDS (2172) (TAIR7_pep_20070425: $N = 31\,924$) were also analyzed by blastx.[29] The threshold of E <1*E*-30 revealed 3853 and 3679 FLcDNAs with homology to rice and *Arabidopsis*, respectively.

All the homologous barley FLcDNAs were mapped to the rice pseudomolecule (IRGSP built 3) to show the homology to rice locus, mRNAs and CDS. The Gbrowse data set is accessible online from http://map.lab.nig.ac.jp:8090/cgi-bin/gbrowse/Oryza_vs_Hordeum/.

3.3 *Estimation of ORF completeness in barley FLcDNAs*

The completeness of ORFs in each FLcDNA sequence was estimated by the blastx analysis, using rice ORF and UniProt (http://beta.uniprot.org/) sequences. The results are available at Supplementary Table S1. An ORF was judged as FL when an FLcDNA satisfies the following criterion: (i) strand =1, (ii) subject start position <20 bp and (iii) subject start position ≤ FLcDNA start position. Based on this criterion, 2912 and 304 FLcDNAs were estimated to have complete ORFs compared with rice ORF and UniProt sequences, respectively. Thus, at least 60% of the FLcDNAs probably have complete ORFs.

Since only 3853 of the FLcDNAs have high (E <1*E*-30) homology with rice ORF amino acid sequences, many of the FLcDNAs were lacking the homologous rice genes to compare with. Therefore, the estimate of 60% may underestimate the ORF completeness of the FLcDNAs. A preliminary estimate based on random 5′ sequencing of 96 clones picked from the cDNA library suggested 92% of completeness based on homology search with the GenBank nt database (data not shown). The precise estimate of ORF completeness will have to wait until barley genome sequence is available.

3.4 *GO annotation and orthologous gene comparison for FLcDNA sequence*

InterProScan analysis of the 5006 FLcDNA clones produced results for 4980 clones (Supplementary Table S2). Of the 4980 clones, 3156 had GO results. GO-positive sequences were categorized using their GO terms and categories with the namespace of molecular function (Table 2). ‘Binding’ (32.6%) and ‘catalytic activity’ (31.6%) were the most frequent functions of the FLcDNAs. For all other functions, the percentage of FLcDNAs was less than 5%. The 555 FLcDNA clones that showed no significant homology to either rice or *Arabidopsis* genes were separately categorized by GO annotation: 70% had no GO. For positives, the category spectrum was similar to that for the total set of clones.

GO category (molecular function) comparison between the barley FLcDNAs (Table 2) and rice rap2 (http://rapdb.dna.affrc.go.jp/RAP2_statistics.html) showed similar spectra. However, the frequencies of GO molecular functions (GO slim) in *Arabidopsis* (http://www.arabidopsis.org/tools/bulk/go/index.jsp) were different from those in barley and rice. Considering the genomic similarity between barley and rice, it may be reasonable to say that the barley FLcDNAs comprised clones, reflecting the spectrum of functions of all genes in grass species.

Orthologous relationships between barley FLcDNAs and rice/Arabidopsis proteins were compared using the GreenPhyl Iterative Ortholog Search Tool (i-GOST) with GreenPhylDB[30] (http://greenphyl.cirad.fr), which contains 6421 mutigenic families half automatically clustered including 492 TAIR (http://sss.arabidopsis.org/), 1903 InterPro (http://www.ebi.ac.uk/interpro/) and 981 KEGG (http://www.genome.jp/kegg/) families. Of the 5028 possible amino acid sequences derived from the 5006 FLcDNAs, 4032 (80.2%) were classified into 1678 GreenPhyl families (Supplementary Table S3). A total of 997 protein families were identified at GreenPhylDB clustering level 1 (Inflation 1.2).[31] Other protein families were identified at under clustering level 2–4 (Inflation 2–5). Category levels were due to clustering stringency (http://greenphyl.cines.fr/html/cluster.htm). The abundant families are listed in Table 3. The most abundant were kinase and/or LRR superfamilies, with 84 amino acid sequences from FLcDNA.

Table 3. Abundant orthologous families of barley FLcDNA analyzed by GreenPhyl Ortholog Search Tool with GreenPhylDB (http://greenphyl.cirad.fr), which contain all rice and *Arabidopsis* gene families

No. of FL cDNAs	GreenPhylDB ID	Family
84	20828	Kinase and/or LRR superfamily
61	20842	RNA-binding family (RNP-1)
53	20833	Kinase superfamily
40	20878	Ras GTPase family
40	20839	Cytochrome P450 family
29	20863	Peroxidase family
29	20843	WD40 repeat family
26	20918	Ubiquitin-conjugating family. See at level 2
25	20834	Pentatricopeptide (PPR) repeat-containing protein family
22	20884	AAA-type ATPase family
20	20883	Sugar transporter family
20	20862	Dehydrogenase/reductase (SDR) family
19	21037	Chlorophyll a/b binding family
19	21014	Proteasome family
19	20866	Heat shock protein DnaJ family
19	20853	ABC transporter family
18	20954	Cellular retinaldehyde binding/alpha-tocopherol transport family (SEC14)
18	20872	2OG-Fe(II) oxygenase family
18	20841	Zinc finger (C3HC4-type RING finger) family
17	20909	Thioredoxin family
16	20993	O-methyltransferase family
16	20908	Mitochondrial substrate carrier family
14	20849	Glycosyltransferase-family 1
13	21544	Glyceraldehyde-3-phosphate dehydrogenase family
13	20928	Glycosyl hydrolase family 1
12	20994	Peptidyl-prolyl *cis–trans* isomerase cyclophilin-type family
12	20925	Elongation factor family. See at level 2
12	20922	Papain family. See at level 2
11	21176	Histone H2A family
11	21044	Actin and actin-like family
11	20991	Alcohol dehydrogenase, zinc-containing family 2

Continued

Table 3. Continued

No. of FL cDNAs	GreenPhylDB ID	Family
11	20891	Proton-dependent oligopeptide transport (POT) family similar to LeOPT1 family
10	21419	Histone H4 family
10	21103	Tubulin family
10	21046	Chaperonin Cpn60/TCP family
10	21021	Heat shock protein 70 family
10	20952	Major intrinsic family (MIP)
10	20906	Serine carboxypeptidase S10 family

3.5 Comparison with published barley EST sequences

The 5006 FLcDNAs were searched using blastn, against all the barley EST sequences in GenBank. Of these, 4753 (95%) showed high homology (E <1*E*-30) and 253 showed homology below the threshold. The blastn homology with rice genes showed that 152 of these 253 FLcDNAs showed homologies (E <1*E*-5) to rice gene sequences, indicating that these genes may be expressed at low levels and their detection is due to our systematic program of mRNA sampling. The other 101 FLcDNAs did not show homologies to any rice genes (all are included in the 555 genes noted in Table 2), *Arabidopsis* genes or barley ESTs. The blastn search with IRGSP build 4 rice pseudomolecule (http://rapdb.dna.affrc.go.jp/) revealed that only three of these cDNA sequences showed homologies to rice genome. These results indicate that most of these genes may be novel transcripts specific to barley.

3.6 cDNA sequences as tools for gene isolation and genome analysis in the Triticeae

There were high levels of homology with other Triticeae EST-based resources. At E <1*E*-30, the numbers of significant hits and percentages were as follows: Affymetrix Barley 1 GeneChip (4060; 81%), wheat gene index (4199; 84%) and EST sequences on the genetic map of Sato et al. (submitted for publication) (1328; 27%), which comprised 2890 non-redundant sets of 3′ barley ESTs. The blastn scores and target sequence information are available in Supplementary Table S4. Genetic map positions of FLcDNAs are available in Supplementary Table S5 and cMAP viewer online at http://map.lab.nig.ac.jp:8085/cmap/. The 1328 mapped cDNA sequences are well-distributed on each chromosome (160, 233, 211, 151, 214, 168 and 191 cDNAs on chromosomes 1H to 7H, respectively). Mapped FLcDNAs will be useful for cloning genes, especially when the mapped loci are also represented on an expression profiling array such as the Affymetrix Barley 1 GeneChip. As shown in Supplementary Table S4, 4060 (81%) cDNA sequences are formatted on Barley 1 GeneChip and 1300 of them are assigned genetic map positions. As an example, expressed probes on the Barley 1 GeneChip with genetic map positions in barley were used to identify orthologous transporter gene in rice.[18] This strategy led to identifying the corresponding FLcDNAs in barley and ultimately cloning and characterizing the function of Aluminium tolerance gene in barley.

There is general colinearity and content of the barley and wheat genomes.[22,23] Therefore, it is not surprising that 84% of the wheat genes showed a high level of sequence similarity with the barley FLcDNAs. The homology between barley and diploid wheat (*Triticum monococcum* and *Triticum boeoticum*) is complete, except for the reciprocal translocation between chromosomes 4A and 5A[21] (see online at http://map.lab.nig.ac.jp:8085/cmap/). There are some excellent examples of how the complementary use of wheat and barley genetic resources, based on homeology, can be of benefit to gene discovery in both species.[25,32]

3.7 Conclusions

The 5006 barley FLcDNAs are the first published sequences in the Triticeae and third largest resource in plants, after rice and *Arabidopsis*. The present barley FLcDNA sequences are of as high quality as those reported for rice and *Arabidopsis*. These sequences provide access to nucleotide and amino acid sequences in barley and other related species, especially wheat.

Since a whole genome sequence is not yet available for barley, evaluation of these FLcDNA sequences by alignment with genome sequence is not possible, as for *Arabidopsis*[11] and rice.[12] However, the FLcDNAs will be immediately useful after the release of the genome sequence of barley. The efficiency was demonstrated by Sato (unpublished data), who sequenced 400 Haruna Nijo BAC clones mapped on the barley chromosome 3H. The FLcDNAs in this study can be aligned on these BAC sequences for gene identification.

The mapped ESTs can be assigned physical coordinates using cytogenetic stocks, such as barley—wheat addition lines. For example, barley EST markers on the barley chromosome deletion stocks can estimate the physical location of these ESTs.[3,4] Moreover, cDNAs with large insert sizes can be directly mapped to the chromosomes by fluorescent *in situ* hybridization. Thus, the combination of multiple genomic resources, including EST maps and

FLcDNAs, will assist in finally revealing the complete structure and function of the barley genome.

Sequence data from this article have been deposited with the DDBJ/EMBL/GenBank Data Libraries under accession nos AK248134–AK253139. The online database with annotation is available at http://www.shigen.nig.ac.jp/barley/.

Acknowledgements: We would like to thank Patrick M. Hayes (Oregon State University, USA) for critical reading of the manuscript.

Supplementary Data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

Funding

The sequencing analysis in this research was supported by a grant from the National Bioresource Project—Barley, Japan. The bioinformatics analysis was supported by the Program of Promotion of Basic Research Activities for Innovative Biosciences.

References

1. Arumuganathan, K. and Earle, E. D. 1991, Nuclear DNA content of some important plant species, *Plant Mol. Biol. Rep.*, **9**, 208–218.
2. Stein, N., Prasad, M., Scholz, U., et al. 2007, A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics, *Theor. Appl. Genet.*, **114**, 823–839.
3. Nasuda, S., Kikkawa, Y., Ashida, T., Islam, A. K., Sato, K. and Endo, T. R. 2005, Chromosomal assignment and deletion mapping of barley EST markers, *Genes Genet. Syst.*, **80**, 357–366.
4. Ashida, T., Nasuda, S., Sato, K. and Endo, T. R. 2007, Dissection of barley chromosome 5H in common wheat, *Genes Genet. Syst.*, **82**, 123–133.
5. Yu, Y., Tomkins, J. P., Waugh, R., et al. 2000, A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes, *Theor. Appl. Genet.*, **101**, 1093–1099.
6. Saisho, D., Myoraku, E., Kawasaki, S., Sato, K. and Takeda, K. 2007, Construction and characterization of a bacterial artificial chromosome (BAC) library for Japanese malting barley 'Haruna Nijo', *Breed Sci.*, **57**, 29–38.
7. Close, T. J., Wanamaker, S. I., Caldo, R. A., et al. 2004, A new resource for cereal genomics: 22K barley GeneChip comes of age, *Plant Physiol.*, **134**, 960–968.
8. Druka, A., Muehlbauer, G., Druka, I., et al. 2006, An atlas of gene expression from seed to seed through barley development, *Funct. Integr. Genomics*, **6**, 202–211.
9. Carninci, P., Kvam, C., Kitamura, A., et al. 1996, High-efficiency full-length cDNA cloning by biotinylated CAP trapper, *Genomics*, **37**, 327–336.
10. Carninci, P., Nishiyama, Y., Westover, A., et al. 1998, Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA, *Proc. Natl Acad. Sci. USA*, **95**, 520–524.
11. Seki, M., Narusaka, M., Kamiya, A., et al. 2002, Functional annotation of a full-length *Arabidopsis* cDNA collection, *Science*, **296**, 141–145.
12. Kikuchi, S., Satoh, K., Nagata, T., et al. 2003, Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice, *Science*, **301**, 376–379.
13. Satoh, K., Doi, K., Nagata, T., et al. 2007, Gene organization in rice revealed by full-length cDNA mapping and gene expression analysis through microarray, *PLoS ONE*, **2**, e1235.
14. Okubo, K., Hori, N., Matoba, R., et al. 1992, Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, *Nat. Genet.*, **2**, 173–179.
15. 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.
16. Zohary, D. and Hopf, M. 2000, Barley: *Hordeum vulgare*, In: *Domestication of Plants in the Old World*, Third Ed., Oxford University Press Inc.: New York, pp. 59–68.
17. Harlan, J. R. 1995, Barley, In: Smartt, J. and Simmonds, N. W. (eds.), *Evolution of Crop Plants*, 2nd Ed., Longman Scientific and Technical: Harlow, UK, pp. 140–147.
18. Furukawa, J., Yamaji, N., Wang, H., et al. 2007, An aluminum-activated citrate transporter in barley, *Plant Cell Physiol.*, **48**, 1081–1091.
19. Bothmer, R., von Sato, K., Komatsuda, T., Yasuda, S. and Fischbeck, G. 2003, The domestication of cultivated barley, In: Bothmer, von Hintum, R., van Knüpffer, H. and Sato, K. (eds.), *Diversity in Barley (*Hordeum vulgare*)*, Amsterdam, The Netherlands: Elsevier Science B.V., pp. 9–27.
20. Islam, A. K. M. R. and Shepherd, K. W. 1992, Production of wheat-barley recombinant chromosomes through induced homoeologous pairing. 1. Isolation of recombinants involving barley arms 3HL and 6HL, *Theor. Appl. Genet.*, **83**, 489–494.
21. Hori, K., Takehara, S., Nankaku, N., Sato, K., Sasakuma, T. and Takeda, K. 2007, Linkage map construction and QTL detection based on barley ESTs in A genome diploid wheat, *Breeding Sci.*, **57**, 39–45.
22. Moore, G., Devos, K. M., Wang, Z. and Gale, M. D. 1995, Cereal genome evolution. Grasses, line up and form a circle, *Curr. Biol.*, **5**, 737–739.
23. Devos, K. M. 2005, Updating the 'crop circle', *Curr. Opin. Plant Biol.*, **8**, 155–162.
24. Komatsuda, T., Pourkheirandish, M., He, C., et al. 2007, Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene, *Proc. Natl Acad. Sci. USA*, **104**, 1424–1429.
25. Yan, L., Loukoianov, A., Blechl, A., et al. 2004, The wheat VRN2 gene is a flowering repressor down-regulated by vernalization, *Science*, **303**, 1640–1644.

26. Carninci, P., Westover, A., Nishiyama, Y., et al. 1997, High efficiency selection of full-length cDNA by improved biotinylated cap trapper, *DNA Res.*, **4**, 61–66.
27. Huang, X. and Madan, A. 1999, CAP3: a DNA sequence assembly program, *Genome Res.*, **9**, 868–877.
28. 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.
29. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
30. Conte, M. G., Gaillard, S., Lanau, N., Rouard, M. and Perin, C. 2008, GreenPhylDB: a database for plant comparative genomics, *Nucleic Acids Res.*, **36**, D991–D998.
31. Conte, M. G., Gaillard, S., Droc, G. and Perin, C. 2008, Phylogenomics of plant genomics: a methodology for genome-wide searches for orthologs in plants, *BMC Genom.*, **9**, 183.
32. Yan, L., Fu, D., Li, C., et al. 2006, The wheat and barley vernalization gene VRN3 is an orthologue of FT, *Proc. Natl. Acad. Sci. USA*, **103**, 19581–19586.