

Structural Analysis of a *Lotus japonicus* Genome. I. Sequence Features and Mapping of Fifty-six TAC Clones Which Cover the 5.4 Mb Regions of the Genome

Shusei SATO, Takakazu KANEKO, Yasukazu NAKAMURA, Erika ASAMIZU, Tomohiko KATO, and Satoshi TABATA*

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

(Received 20 November 2001)

Abstract

A total of 56 TAC clones with an average insert size of 100 kb were isolated from a TAC library of the *Lotus japonicus* genome based on the expressed sequences tags (ESTs), cDNA and gene information, and their nucleotide sequences were determined according to the shot-gun based strategy. The total length of the sequenced regions is 5,473,195 bp. By comparison with the sequences in protein and EST databases and analysis with computer programs for gene modeling, a total of 605 potential protein-encoding genes with known or predicted functions, 69 gene segments, and 172 pseudogenes were identified. The average density of the genes assigned so far is 1 gene/8120 bp. Introns were identified in approximately 78% of the potential genes. There was an average of 3.8 introns per gene and the average length of the introns was 375 bp. DNA markers were generated based on the nucleotide sequences obtained, and each clone was mapped onto the linkage map using the F₂ mapping population derived from a cross of *L. japonicus* Gifu B-129 and Miyakojima MG-20. The sequence data, gene information and mapping information are available through the World Wide Web at <http://www.kazusa.or.jp/lotus/>.

Key words: *Lotus japonicus*; genomic sequence; TAC genomic library; gene prediction; linkage mapping

1. Introduction

The accumulation of a large quantity of information on the genome and gene structures, together with thousands of expressed sequence tags (ESTs), is drastically changing the strategy of plant genetics. The most recent notable accomplishment in this regard is the completion of sequencing the *Arabidopsis thaliana* genome.¹ Now that the information on the entire genome structure is available, the systematic functional analysis of *A. thaliana* genes will certainly be a popular and promising field. Nevertheless, other plant species have their own characteristics and advantages which make them useful for the study of individual phenomena. And, the application of knowledge from *A. thaliana* to other plant species and vice versa will increase our understanding of the genetic systems in all plants.

Legumes comprise 18,000 diverse species with a variety of characteristics. Many of them have long been targets of plant breeding because of their agronomic importance. In addition, a few species have been chosen as “model

legumes” to perform genetic and physiological studies on legume-specific phenomena such as plant-microbe interactions and symbiotic nitrogen fixation. *Lotus japonicus* is a typical model legume with the following characteristics: short life cycle (2–3 months), self-fertility, diploidy ($n = 6$), and small genome size (472.1 Mb).^{2–4} Mutants in various processes of symbiosis and nitrogen fixation have been isolated,^{5–7} and genes specifically expressed during these processes have been characterized utilizing the established transformation system.^{8–10} A large number of ESTs have been accumulated^{11–13} and high-density linkage maps of all 6 chromosomes have been generated.¹⁴

To understand the genetic system of legume species and to facilitate isolation and characterization of genes responsible for legume-specific phenomena, we initiated large-scale sequencing of the *L. japonicus* genome. We started sequencing from multiple points of the genome by selecting the genomic clones carrying the genes corresponding to a variety of ESTs, cDNAs and gene segments. The sequenced clones were then genetically localized onto the linkage map using the markers generated utilizing the sequence information obtained. In the present paper, we report sequence features of a total of

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

5,473,195 bp of the *L. japonicus* genome, which were carried by 56 genomic clones genetically mapped onto the 6 chromosomes.

2. Materials and Methods

2.1. Genomic libraries

High-molecular-weight DNA was isolated from *L. japonicus* accession MG-20 according to the method described by Zhang et al.¹⁵ The DNA was partially digested with *Mbo*I or *Hind*III and size fractionated in the 100- to 150-kb size range by pulsed-field gel electrophoresis. The recovered DNAs were ligated with *Bam*HI- or *Hind*III-digested pYL7AC7.¹⁶ The ligated DNAs were then used for transformation of *E. coli* DH10B by electroporation, and transformants were selected on LB agar plates containing 25 µg/ml kanamycin and 5% sucrose. The average insert sizes were 87 kb, 96 kb, 105 kb, and 106 kb for four independent preparations, the total of which is 7.7 haploid genome equivalents. The TAC libraries thus generated were arrayed in ninety-three 384-well microtiter plates, and 48 DNA pools, each containing 384 clones, were subjected to PCR screening.

2.2. Clone isolation

EST-associated TAC clones were isolated as follows. Oligonucleotides were synthesized on the basis of nucleotide sequences of the public ESTs and cDNAs from *L. japonicus*, a LjENBP1 gene (provided by J. Stougaard), and a symbiosis-induced lipase-like gene (provided by M. Parniske), as listed in Table 1. Using the synthesized primers, the TAC libraries were screened for each EST sequence by PCR according to the three-dimensional pooling method. The nucleotide sequences of the PCR products were determined and compared with those of the corresponding ESTs to confirm the authenticity of screening.

2.3. DNA sequencing and data assembly

The nucleotide sequence of each TAC insert was determined according to the bridging shotgun method described previously.¹⁷ Briefly, the TAC DNAs were subjected to sonication followed by size-fractionation on agarose gel electrophoresis. Fractions of approximately 1.0 kb and 2.5 kb were cloned into M13mp18 and named libraries of element clones and bridge clones, respectively. The element clones were propagated on microtiter dishes and single-stranded DNA was prepared for the sequencing reaction. For the bridging clones, each insert was amplified by PCR and used as a template.

Sequencing was performed using the cycle sequencing kits (Dye-terminator Cycle Sequencing kit of Applied Biosystems, USA) with DNA sequencer type 377XL (Applied Biosystems, USA) according to the protocol

recommended by the manufacturer. The single-pass sequences deduced for one strand of element clones and those for both ends of bridge clones, the total of which corresponded to about 6 times the equivalent of an insert, were assembled using Phred-Phrap programs (Phil Green, Univ. Washington, Seattle, USA). After extension of the termini of each contig by the primer extension method followed by re-connection, most of the TAC inserts were assembled into 1 to 3 contigs with more than 95% coverage of either both strands or multiple reads on one strand. A lower threshold of acceptability for the generation of consensus sequences was set at a Phred score 20 for each base.

2.4. Computer-assisted data analysis

For assignment of the protein-coding regions and gene modeling, similarity search and computer prediction were performed as described previously.¹⁷ Briefly, similarity search against the non-redundant protein sequence database nr (compiled by NCBI) was carried out using the BLASTX program.¹⁸ In parallel, the positions of potential protein-encoding regions were predicted with the Grail,¹⁹ GENSCAN²⁰ and NetGene2²¹ computer programs. The transcribed regions were assigned by comparison of the nucleotide sequences with *L. japonicus* ESTs both in-house and in the public databases using the BLASTN program.¹⁸ All the results obtained were compiled with the aid of our new web-based tool, named KAPSEL (Kazusa Annotation Pipeline SystEm for *Lotus* genome sequencing project; manuscript in preparation), then assignment of the potential protein coding genes was carried out by taking both similarity to known genes and computer prediction into consideration.

The RNA-coding regions were assigned on the basis of sequence similarity to the reported structural RNAs. For tRNA genes, prediction by the tRNAscan-SE program²² was also taken into account.

2.5. DNA marker generation and linkage analysis

For generation of DNA markers, simple sequence repeats (SSRs) such as (AT)_n, (GT)_n, and (AAT)_n of ≥ 15 bp were searched on the TAC nucleotide sequences determined. Primer pairs were then designed using the Consed program²³ on the flanking sequences of each SSR so that amplified fragment sizes were in the range 100 bp to 220 bp. PCR was performed in a total volume of 10 µl containing 20 ng of *L. japonicus* genomic DNA, 1 × PCR buffer (TaKaRa Shuzo, Japan), 1 unit of TaKaRa Taq (TaKaRa Shuzo, Japan), 0.2 mM dNTPs and 0.5 µM of each designed primer. Reactions were run on a PE9600 with the following program: 2 min at 94°C, then 30 cycles of 45 sec at 94°C, 45 sec at 55°C and 2 min at 72°C, followed by a final 10-min extension at 72°C. PCR products were resolved on a 3% MetaPhor agarose gel (BMA, USA), and the primer sets giving a polymorphism be-

Table 1. List of TAC clones analyzed in this study.

Clone name	Accession No.	Length (bp)	Chr	Annotation of the gene used for screening	Accession No.
LjT09C23	AP004466	103,397	1	early nodule-specific protein	AV772180
LjT06K11	AP004467	106,023	2	alfa-carboxyltransferase precursor	AV779984
LjT09A24	AP004468	88,828	4	6-phosphogluconate dehydrogenase	AV772171
LjT15D01	AP004469	92,631	4	a Lea protein with hydrophobic domain, high pI value(11.6)	AV777717
	AP004470	28,108			
LjT15H09	AP004471	97,710	3	aminoalcoholphosphotransferase	AV772217
LjT17P02	AP004472	72,374	4	nodulin-like protein	AV780525
LjT10J15	AP004473	92,741	4	chlorophyll a/b binding protein	AV771270
LjT13B22	AP004474	77,319	1	ferredoxin-dependent glutamate synthase	AV781337
LjT01H13	AP004475	87,900	1	rac GTPase activating protein 1	AV769057
LjT12K07	AP004476	33,705	2	delta-1-pyrroline-5-carboxylate synthase	AV780844
	AP004477	44,662			
LjT03H13	AP004478	45,319	1	nodulin-26	AV774115
	AP004479	41,820			
	AP004480	11,644			
LjT14O07	AP004481	74,668	6	sucrose synthase (nodulin-100)	AV771445
LjT17C05	AP004482	129,532	6	glutamine synthetase nodule isozyme	AV768975
LjT13O04	AP004483	122,116	1	14-3-3-like protein	AV772587
LjT06H17	AP004484	121,127	1	glutamate-1-semialdehyde 2,1-aminomutase precursor	AV769262
LjT08D14	AP004485	86,539	2	Mg-chelatase subunit D	AV776889
LjT05B20	AP004486	73,833	5	P-protein subunit of glycine decarboxylase enzyme complex	AV777864
	AP004487	72,735			
LjT02F05	AP004488	80,920	2	similarity to SCAMP37	AV771064
LjT04I02	AP004489	83,700	2	GMFP4	AV764976
LjT11G09	AP004490	114,576	3	metallothionein	AV772440
LjT17G08	AP004491	91,039	1	isocitrate dehydrogenase [NADP]	AV772234
LjT08D16	AP004492	51,575	4	phragmoplastin 12 - soybean	AV771202
	AP004493	36,100			
LjT09H17	AP004494	103,129	4	proline-rich protein, 14K - kidney bean	AV765642
LjT08B24	AP004495	87,135	1	protein disulfide isomerase precursor (PDI)	AV772273
LjT11C13	AP004496	53,909	1	serine/threonine protein phosphatase PP2A catalytic subunit	AV764832
	AP004497	34,910			
LjT06A20	AP004498	82,599	4	sucrose-phosphate synthase	AV769733
LjT16N13	AP004499	97,191	2	argonaute (AGO1)-like protein	AV778325
LjT05P21	AP004500	106,632	1	brassinosteroid insensitive 1	AV778300
LjT07E11	AP004501	1,653	1	EIN2	AV780613
	AP004502	80,917			
LjT02M03	AP004503	83,232	5	ERECTA	AV780537
	AP004504	35,858			
LjT10E18	AP004505	106,041	3	homeodomain-leucine zipper proteininterfascicular fiberless 1	AV780002
LjT03K03	AP004506	62,624	1	UVB-resistance protein UVR8	AV778994
	AP004507	1,231			
	AP004508	22,758			
LjT16K17	AP004509	84,322	6	cellulose synthase catalytic subunit	AV766783
	AP004510	10,720			
LjT17H19	AP004511	108,987	6	senescence-associated protein homolog	AV767289
LjT10L16	AP004512	85,815	4	thaumatin-like protein - turnip	AV765136
LjT01O22	AP004513	54,726	5	CONSTANS-like protein 1	AV769706
	AP004514	32,095			
LjT04F23	AP004515	65,620	4	MADS-box protein MADS3	AV770847
	AP004516	20,786			
LjT14B06	AP004517	113,139	6	MdMADS8	AV780842
LjT07K08	AP004518	27,970	4	AGAMOUS like protein	AV777396
	AP004519	71,321			
LjT05P01	AP004520	110,811	5	phytochrome A	AV779905
LjT17K09	AP004521	92,504	3	cell wall invertase II; beta-furanofructosidase	AV769089
	AP004522	4,160			
LjT03J05	AP004523	86,209	1	zinc-finger protein	AV766234
LjT08O18	AP004524	20,200	5	abscisic stress ripening protein 2	AV774006
	AP004525	86,497			
LjT05E07	AP004526	94,516	6	flavonoid 3-O-galactosyl transferase	AV773576
LjT03B03	AP004527	15,020	6	unknown	AB074987
	AP004528	77,452			
LjT01K12	AP004529	78,249	2	unknown	AB074988
	AP004530	27,342			
LjT13M14	AP004531	94,884	3	unknown	AB074989
LjT09L22	AP004532	89,936	1	putative protein	AB074990
LjT14G02	AP004533	76,730	3	unknown	AV780712
LjT14P20	AP004534	124,300	4	plastidic aldolase	AV411259
LjT15N19	AP004535	6,171	4	PGP224 protein, <i>Penunia hybrida</i>	AV418999
	AP004536	78,574			
LjT04C07	AP004537	33,254	1	tubulin beta-1 chain	AV425326
	AP004538	28,884			
	AP004539	968			
	AP004540	7,686			
	AP004541	45,364			
LjT08G20	AP004542	54,472	2	Nin gene	AJ238956
	AP004543	38,490			
	AP004544	1,481			
LjT26E16	AP004545	100,595	1	LjENBP1 gene	
LjT43N05	AP004546	97,268	1	unknown	AV425682
LjT31L24	AP004547	65,103	6	symbiosis induced lipase-like gene	
	AP004548	19,059			
	AP004549	19,055			

tween the parents of the mapping population, accessions Miyakojima MG-20 and Gifu B-129, were selected and used for scoring 127 F₂ mapping populations.¹⁴

In cases where no simple sequence repeat length polymorphism (SSLP) was found, single nucleotide polymorphisms (SNPs) between the parental genotypes were searched. Oligonucleotides were designed based on the sequence information of the TAC clones, mostly from intergenic regions, and the corresponding regions of the genome of the accession Gifu B129 were amplified by PCR, followed by sequence analysis. If SNPs were identified by comparing the sequences between two parents, they were converted into derived cleaved amplified polymorphic sequence (dCAPS) markers facilitated by the dCAPS finder program.²⁴ PCR reactions were performed using the same condition used for amplification of SSR markers. Aliquots of the PCR product (10 μ l) were digested for 2 hr in 15 μ l (total volume) with 2–5 units of the appropriate restriction endonuclease, and the reaction mixture was analyzed on a 3% MetaPhor agarose gel to detect polymorphisms.

Analysis of segregation data for SSR and dCAPS markers and linkage map integration were carried out using the F₂ mapping population of accessions Miyakojima MG-20 and Gifu B-129, as described in the accompanying paper.¹⁴

3. Results and Discussion

3.1. Isolation and sequencing of TAC clones

TAC clones which contain the DNA regions corresponding to 56 ESTs, cDNAs and a gene segment listed in Table 1 were isolated by screening the three-dimensional DNA pools of TAC genomic libraries of *L. japonicus* MG-20 by means of PCR. The nucleotide sequence of each TAC insert was deduced according to the modified shotgun method, as described in Materials and Methods. Regions were left as gaps where the data quality does not meet the requirements for the generation of consensus sequences, mostly due to the presence of heavily repetitive sequences and secondary structures. After confirmation of the relative positions and directions of the contigs by amplification of gap regions by PCR, the sequence of each TAC insert was registered as a multiple contig.

The length of the nucleotide sequence of each TAC insert finally determined is listed in Table 1. The total length of the DNA regions sequenced in this study was 5,473,195 bp. The average GC content was 36%.

3.2. Assignment and structural features of potential protein- and RNA-coding regions

The assignment of potential protein coding regions and gene modeling were performed by a combination of similarity search and computer prediction, as described in Materials and Methods. The possibility still remains

that additional genes may be discovered among the intergenic regions in the future. The genes thus assigned were denoted by numbers with the clone names followed by sequential numbers from one end of the insert to the other. When a clone was divided into multiple contigs, contig numbers (a, b, –) were inserted after the clone name. The gene organization in each TAC clone, the structure of each gene, and gene information are presented in our *L. japonicus* genome database at <http://www.kazusa.or.jp/lotus/>. To sum up, complete structures of 605 potential protein-encoding genes were deduced in regions of the genome totaling 5,473,195 bp. In addition, partial structures of 69 potential protein-encoding genes at the terminal regions of the clones and the contigs, and 172 pseudo genes which contain either frameshifts or termination codons in the original coding regions, were identified.

Six hundred and five potential protein-encoding genes were estimated to amount for approximately 1.1% of the total gene constituents by dividing the total genome size (472 Mb) by the size of the regions sequenced, but this percent is probably an underestimate for the following reasons. Firstly, the clones sequenced in this study were likely to be derived from gene-rich regions of the genome because they were selected based on the ESTs, namely the expressed sequences. Secondly, Ito et al. reported that significant portions of the prometaphase chromosomes of *L. japonicus* are either heavily or moderately condensed,⁴ strongly suggesting the presence of regions rich in repetitive sequences and poor in protein-encoding genes in the genome. Therefore, we speculate that the genes identified in this study represent more of the gene constituents than simply calculated.

The general features of the 605 genes in *L. japonicus* whose complete structures were deduced, along with those of *A. thaliana*, are listed in Table 2. The average length of the coding exons (266 bp) and the average number of introns per gene (3.8 introns) were quite similar between the genes of the two plant species. However, the average length of genes including introns (2712 bp) was longer in *L. japonicus* than in *A. thaliana* (1918 bp) due to a longer average intron length in *L. japonicus*. The average gene density of the sequenced regions of the *L. japonicus* genome was one gene in every 8120 bp, one-half that of *A. thaliana*. It should be noted that this may be an overestimate because, as described previously, it is quite likely that the clones containing the genomic regions of higher gene contents were preferentially selected and sequenced in this study.

RNA-coding regions were assigned on the basis of sequence similarity to the reported structural RNAs, and also by prediction with the tRNAscan-SE program²² for tRNA genes. As a result, a total of seven tRNA genes corresponding to seven amino acid species, tRNA-Pro(AGG), Arg(CCT), Met(CAT), Lys(CTT), Ala(CGC), Leu(CAA), and Glu(TTC), were identified.

Table 2. Structural features of the assigned protein-encoding genes.

Features	605 genes ^{a)}	6,451 genes ^{b)}
Gene length (bp) including introns	168-19,890 (2,712)	78-17,203 (1,918)
Product length (amino acids)	16-1,816 (426)	25-4,706 (427)
Genes with introns	469 (78%)	4,906 (76%)
Number of intron/gene	0-37 (3.8)	0-48 (4.0)
Exon length (bp)	3-5,451 (266)	2-5,966 (256)
Intron length (bp)	30-5,687 (375)	23-2,989 (157)
GC content of exons	45%	44%
GC content of introns	33%	32%

Structural features of the potential protein coding genes assigned so far are listed: The 605 genes are assigned in this study^{a)} and the 6,451 genes^{b)} previously assigned potential protein genes in our *A. thaliana* genome sequencing project. Average values are shown in parentheses.

Table 3. Functional classification of the assigned protein-encoding genes.

		%
Amino acid biosynthesis	6	1.0
Biosynthesis of cofactors, prosthetic groups, and carriers	4	0.7
Cell envelope	8	1.3
Cellular processes	2	0.3
Central intermediary metabolism	3	0.5
Energy metabolism	14	2.3
Fatty acid, phospholipid and sterol metabolism	4	0.7
Photosynthesis and respiration	4	0.7
Purines, pyrimidines, nucleosides, and nucleotides	2	0.3
Regulatory functions	20	3.3
DNA replication, recombination, and repair	1	0.2
Transcription	3	0.5
Translation	9	1.5
Transport and binding proteins	9	1.5
Other categories	98	16.2
Subtotal of genes similar to genes of known function	187	30.9
Similar to hypothetical protein	200	33.1
Subtotal of genes similar to registered genes	387	64.0
No similarity	218	36.0
Total	605	100.0

They are denoted by numbers with the clone names followed by “r” and sequential numbers.

3.3. Functional assignment and characteristic features of potential protein-encoding genes

3.3.1. Functional assignment

Similarity search of the 605 potential protein-encoding genes against the nr databases indicated that 187 (31%) were homologues of genes of known function, 200 (33%) showed similarity to hypothetical genes, mostly of those in *A. thaliana*, and the remaining 218 (36%) showed no

significant similarity to any registered genes.

The potential protein-encoding genes whose function could be anticipated were grouped into 14 categories with respect to different biological roles, according to the principle of Riley.²⁵ The numbers of genes in each category are summarized in Table 3, and the name of each gene is listed in the *L. japonicus* genome database at <http://www.kazusa.or.jp/lotus/>.

3.3.2. Expression of potential protein-encoding genes

The transcriptional level of each of the potential protein-encoding genes was roughly monitored by count-

Table 4. List of SSR markers.

clone name	marker name	SSR pattern		product size (bp)			Fw primer (5' b 3)	Rv primer (5' b 3)	Chr	extension (s) (bp)
		motif	number of repeat	MG-20	B-129	heteroduplex				
LjT08C 23	TM0001	AT	10	136	144	-	TCCTGTTGATCCTCATTATCC	CTCCTCATTATATATAAATTGTCA	1	10
LjT13B 22	TM0009	CT	10	149	105	-	CCTGTCTACCAAACGCTAC	ACTTTCAATTAACATAAAGCAG	1	-
LjT01I 13	TM0010	CT	12	187	205	-	AAACCAGTGTTCAGCTTGT	TTATATATCTCTCTCGGCC	1	-
LjT03H 13	TM0012	CT	20	144	154	-	GAGAGGAATGTCGTAGGAAG	GCTTCTTCTTCAATCTCTGC	1	10(B-129), 18(MG-20)
LjT13O 04	TM0016	AT	16	186	220	-	TTTAGTGCCTTGTAGGTGAG	AAATCTACATAAACTTCAGTG	1	-
LjT08I 17	TM0017	AT	15	149	141	-	CGAATTGTATGGTATTGTATG	AGAATGCTCAAATAACAATC	1	10, 10(B-129), 16(MG-20)
LjT17G 08	TM0023	AT	12	185	179	200	CATAAGCACAACAATTCATAG	GTTGTGTTCAAAGTTAGGG	1	-
LjT08B 24	TM0027	CT	33	178	158	-	AGGATAATTACATTCACCTC	TCTTGCAATATCTATGACTGG	1	-
LjT11C 13	TM0029	CT	15	152	148	155	CCTATATAACCTTATTCAAATTGG	ACGAAAACAAAACCCCTGCTG	1	-
LjT07P 21	TM0032	AT	17	168	160	-	CTTACCACCTTAAGCTTGCTG	GTTTAAATTTCTCCCATTTTC	1	10(MG-20), 20(B-129)
LjT07E 11	TM0033	AAT	18	189	159	200	AGTGTACTCGGATTAACACC	TTAAACATTTGAAATAGATGTAAC	1	-
LjT03K 03	TM0036	ATC	9	171	180	190	GATGTGACGGTGAATATTG	AGAGAGAAGTGGAGCTTACG	1	-
LjT07G 05	TM0050	GT	11	149	165	180	ATTTGTGGGATACATFGAC	GCAGGTATCCATCAATTCTC	1	-
LjT08L 22	TM0063	AT	14	163	155	-	AAATTGAAAAGTTGGGACGG	GCATGAGAGAATGAGACCTATAAG	1	-
LjT04C 07	TM0088	AAT	18	181	190	215	CTCCACCTTATAGAGGGTATG	AAAAGAGTGAAGTTGAAAGC	1	-
LjT28E 16	TM0103	AT	23	192	182	-	TGACAAGAGCTTCATAAGAG	GATGAAGTACAGACACCCGAC	1	-
LjT08K 11	TM0002	CT	15	159	179	-	AGCGATCTACATTCAGAG	AGCGTCTCTCAGTGTG	1, 2*	-
LjT12K 07	TM0011	AT	44	222	250	-	TGAACCGATATCCATATCTAT	TAATTTGAAAGTTGGGGAC	1, 2*	-
LjT08D 14	TM0018	ATG	11	153	147	159	GTTTGAGCAAGTTAGAGGTG	CGGATAGAAAGGTAGAAGAG	2	-
LjT07F 05	TM0020	CA	16	180	190	194	GCAGGCTGTGTTAAAGCATC	TTCTCATGACAGTCATCAAAC	2	140
LjT04I 02	TM0021	CT	16	157	151	-	GGTCATCTTGTGATAGTAAGTAA	CTGTTGTATCAAGCCCAAG	2	-
LjT15H 09	TM0005	GT	16	159	171	175	ACAAAACACAGAAGCTTTTGG	CACTACTCATTTACGCCGAC	3	-
LjT11G 09	TM0022	AT	24	188	148	-	CATTACTAGTCTATGTTTCC	TAAAGTCCATTCATATTGC	3	10(MG-20), 12(B-129), 13(B-129)
LjT10E 18	TM0035	AAG	17	110	95	-	TACATAAAGCAGGGGATG	CCACTTCCACTGTGCTTCTG	3	70
LjT17K 09	TM0049	AT	14	141	159	-	TGGGTTAGCTTACCTGTTTC	ATGTCCTGATCAAATGTTTC	3	-
LjT13M 14	TM0039	AT/CT	7/5	160	166	178	TCCTTCATTCATAACC	TGAGAAGAGAATGAAAAGCC	3	-
LjT14G 02	TM0080	AT	14	137	131	-	AACAAAATACTAAACTATAGCAAAG	CGTCCCACACTCTCTTTAC	3	-
LjT06A 20	TM0030	AAT	18	142	121	-	GTCCAACACGGGAGAATGAG	CATAGAAACCTAAGCATGAGTC	4	-
LjT04F 23	TM0044	ATG	9	139	142	155	TGCTATGATCAGTGTGAAG	TCCAACCTTTATGTTATTAGC	4	-
LjT07K 08	TM0046	CT	16	155	143	-	ATCTAACCAAAACGTGCTTC	TTCTTGCCTTTCTCTGTGG	4	-
LjT14P 20	TM0037	AAAT	7	141	129	-	AGCTGTCGATGATCAGAAT	AAAAGGGTTCAAATAGAATAG	4	-
LjT15N 19	TM0037	AT	11	183	179	200	TTGTGTTGGATGATGTAGC	CTTAACTTTAAAGTTGAAAGCTGG	4	-
LjT10M 08	TM0034	AAT	12	183	192	-	TCACTATGCTTAGATCACAC	GGTTGGTTGTATTGCGTGC	5	-
LjT10I 22	TM0043	AT	12	164	154	-	AAGAAAAGTGAAGTGTGCGC	GGCCAATAATAAGATTGAGC	5	-
LjT05P 01	TM0048	AT	12	162	146	-	TTAATTAGATTGGGAGGTG	CGTAAATAATAGCATTGTCC	5	-
LjT08O 18	TM0052	CT	21	183	177	185, 205	GATATCAGCTGAGTCACTGG	CCACATATGATGATCATTTC	5	-
LjT14O 07	TM0013	AT	20	206	216	-	GTTGTACAGCAATATGTC	CCATATCATATCATATCATATC	6	18(B-129), 16(MG-20)
LjT17C 05	TM0014	AAC	10	159	153	-	AACCAACAGACAGATTATCG	TCATCATGATAATCATCAAC	6	-
LjT17H 19	TM0041	AT	8	173	183	203	GCAGAGAAGAAACGGCTTCG	TCTTTGATCATATAATCACAACC	6	15, 12(B-129), 14(MG-20)
LjT14B 06	TM0045	CT	13	149	141	155	CTCTTATGTTCTTTCAAGC	AGCAACATCACATCTTACC	6	-
LjT07E 07	TM0055	AAG	7	147	153	160	TATAACCGCTCTCACACAC	GATTAACGAAACCGCAGTAG	6	-
LjT03B 03	TM0057	CT	14	138	132	-	CAAGATAAATGCAGATGCG	CTTTCTATAAAGCAGTAAACTGG	6	12(B-129), 12(MG-20)
LjT31L 24	TM0028	AT	14	134	144	-	TCTTTGTTAATTCAGATTGAC	AACTATCCATTAATGTTTCTC	6	140

* markers on the region of translocation between MG-20 and B-129.

a) mapped as an MG-20 dominant marker.

Table 5. List of dCAPS markers.

clone name	marker name	Enzyme	product size (bp)		Fw primer (5 to 3)	Rv primer (5 to 3)	Chr
			MG-20	B-129			
LjT49N 05	TM0178	Hin dIII	145	117+ 28	TCATTCACACTTCCTTTTTATAAAGCT	AGGATTCGAACCTCGAGGAG	1
LjT16N 13	TM0031	Pst I	122+ 27	149	AATGGAAGGATCCAAGCTCCAACCTGCA	AAGTGAAGTTGTTTTATCTGAGG	1, 2*
LjT01K 12	TM0058	Hin dIII	171	142+ 29	ACTAAGTTGCTTGTAACTTATTGTCAAG	GTCCCATTAATAATGCCCTTCC	1, 2*
LjT08G 20	TM0102	Nde I	134+ 29	163	CAGGCTGCAAAACATTAATTGGATACATA	AACTACAATGTCTCCAATGC	1, 2*
LjT09A 24	TM0003	Ssp I	170	145+ 25	GAGGCTGAGGGCAGACAGAGAAATA	GCATCAACTTGTAGTTCCTTC	4
LjT15D 01	TM0004	Bsp I407I	137	108+ 29	AAAAGATTACATAAAAATGTTTGTGATC	ACATGAATGCTGTCCTGTCC	4
LjT17P 02	TM0006	Hha I	188+ 29	217	CAACTTTTGGCCCTGTCTGAACAATAGC	AGCACTGCTCAATCAAAGC	4
LjT10J 15	TM0007	Kpn I	187	159+ 28	TCGCAATTTGATTTTTGAGCCGGTGGT	TAAATAGCGGCCGAAATAGC	4
LjT08D 16	TM0025	Hin II	128+ 29	157	GAAGTTAGTTAGTTAGTTAGTTAGTGACT	TGACAGCAAGAAAGCATCG	4
LjT09I 17	TM0026	Hin dIII	223	194+ 29	TTTTTATAAATATTACCATATTTAAAAGC	GGAGAGTTTGTTCCTCAAGAC	4
LjT10L 16	TM0042	Dra I	111+ 27	138	TAAGAGTTGGGACATATGGATTGTTTAA	GTAAATTTATGTGTATATTGCC	4
LjT08B 20	TM0019	Sph I	137+ 25	162	TCTGCACCCCTTCTCCAGCATCGCA	AAGAGCAGTGAATTATCATCG	5
LjT16K 17	TM0037	Acl I	126+ 25	154	GTCATATTTTCAATTGTAATTTAACGT	CAAGCAAACCTCATAACCTG	6

* markers on the region of translocation between MG-20 and B-129.

ing the number of matched *L. japonicus* ESTs in-house (40,555 ESTs, manuscript in preparation) and in the public DNA databases (31,567 ESTs).¹¹⁻¹³ Of the 605 genes identified in this study, 213 (35%) carried EST sequences. Among the EST-matched genes,

16 (2.6%) were hit by 20 or more EST files, suggesting that they are a class of highly transcribed genes. The putative products of such genes include those showing sequence similarity to chlorophyll A-B binding protein AB80 precursor (LjT10J15.11),

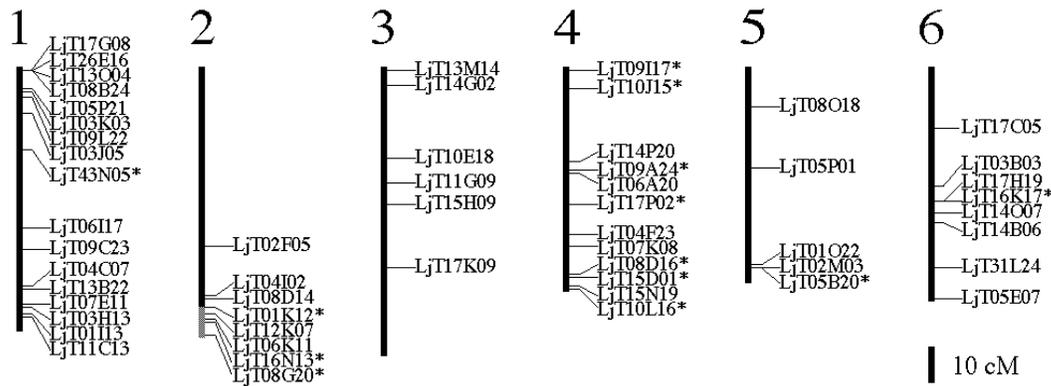


Figure 1. Relative positions of the sequenced TAC clones on the genetic linkage map. The TAC clones sequenced were mapped onto the linkage map of *L. japonicus* accession MG-20 generated in the accompanying paper.¹⁴ Asterisks indicate the dCAPS markers.

tonoplast intrinsic protein, gamma (LjT03H13b.2), 14-3-3-like protein (LjT13O04.1), glycine dehydrogenase mitochondrial precursor (LjT05B20a.2), transcriptional factor 17 (LjT03J05.4), aquaporin (LjT06I17.8), metallothionein-like protein 2 (LjT11G09.2), 36.4 Kda proline-rich protein (LjT06A20.11), calmodulin-related protein 2 (LjT10L16.8), abscisic stress ripening protein 1 (LjT08O18b.4), fructose-bisphosphate aldolase 1, chloroplast precursor (LjT14P20.2), sucrose synthase (LjT14O07.1), isocitrate dehydrogenase (LjT17G08.2), and triosephosphate isomerase, and chloroplast precursor (LjT03H13b.1).

3.3.3. Retroelement-related genes

Of the 846 potential protein-encoding genes and gene segments, 161 (19%) were related to retroelements (*gag* and *pol*). The average density of such retroelement-related genes was one in every 34 kb. It is noteworthy that 126 of these 161 genes were pseudogenes which contain frameshifts or termination codons within the coding regions.

3.3.4. Redundant genes

Tandemly repeated gene arrays were often found in the sequenced regions of the *L. japonicus* genome. These include genes for phenylalanine ammonia-lyase 1 (LjT07E11b.4, LjT07E11b.6, LjT07E11b.7, LjT07E11b.9), minor extracellular protease VPR precursor (LjT14G02.2, LjT14G02.3, LjT14G02.5), alpha-L-arabinofuranosidase A precursor (LjT04C07a.5, LjT04C07a.6 [partial], LjT04C07b.4, LjT04C07b.5 [partial]), and eukaryotic initiation factor (ISO)4F subunit P82-34 (LjT16K17a.8, LjT16K17a.9).

The genes showing sequence similarity to that for anther-specific proline-rich protein APG also formed the tandem arrays: (LjT09C23.6, LjT09C23.7, LjT09C23.9), (LjT07K08b.8, LjT07K08b.9), and (LjT31L24a.8,

LjT31L24a.9, LjT31L24a.10; LjT31L24b.1, LjT31L24b.2; LjT31L24c.1, LjT31L24c.2, LjT31L24c.3, LjT31L24c.4). A characteristic feature of these genes is the presence of the “GDSL” family lipase motif, which is found both in prokaryotes and eukaryotes, and is thought to be involved in the regulation of development and morphogenesis in plants.²⁶ Genes with this motif appeared much more frequently in *L. japonicus* than in *A. thaliana*, where only 23 out of 25,754 genes contain this motif.

3.4. Linkage mapping of TAC clones

The sequenced clones were genetically localized onto the genetic linkage map of *L. japonicus*, as described in Materials and Methods. Two types of PCR-based DNA markers, SSLP and dCAPS, were generated utilizing the sequence information of each clone, and mapping was performed using the F₂ population of two accessions of *L. japonicus*, Miyakojima MG-20 and Gifu B-129. Primer sequences for PCR, product sizes for both accessions, restriction enzymes for digestion, and expected fragment sizes are listed in Tables 4 and 5. Out of 56 generated markers, 43 were SSLP and the remaining 13 were dCAPS, and all of these markers except one (TM0011) were co-dominant markers.

The mapping results were integrated onto the linkage map in the accompanying paper,¹⁴ as shown in Fig. 1. Segmental translocation occurs between chromosome 1 of Gifu B-129 and chromosome 2 of Miyakojima MG-20. The map in Fig. 1 was illustrated according to the map of MG-20 and the translocated region is indicated by a gray bar. The information on the DNA markers and on the surrounding DNA sequences generated in this study will be extremely useful for mapping and isolation of target genes in *L. japonicus*. We will continue collecting both sequence and marker information according to the procedures used in this study.

Acknowledgements: We thank S. Sasamoto, A.

Watanabe, K. Kawashima, T. Kimura, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, A. Muraki, S. Nakayama, N. Nakazaki, S. Shinpo, M. Sugimoto, C. Takeuchi, M. Yamada, and T. Wada for excellent technical assistance. Thanks are also due to Drs. K. Harada, M. Parniske, and J. Stougaard for providing the EST and gene information for screening.

This work was supported by the Kazusa DNA Research Institute Foundation.

References

1. The Arabidopsis-Genome Initiative, 2000, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature*, **408**, 796–815.
2. Handberg, K. and Stougaard, J. 1992, *Lotus japonicus*, an autogamous, diploid legume species for classical and molecular genetics, *Plant J.*, **2**, 487–496.
3. Kawaguchi, M. 2000, *Lotus japonicus* “Miyakojima” MG-20: an early flowering accession suitable for indoor handling, *J. Plant Res.*, **113**, 507–509.
4. Ito, M., Miyamoto, J., Mori, Y. et al. 2000, Genome and chromosome dimensions of *Lotus japonicus*, *J. Plant Res.*, **113**, 435–442.
5. Imaizumi-Anraku, H., Kawaguchi, M., Koiwa, H., Akao, S., and Syono, K. 1997, Two ineffective-nodulating mutants of *Lotus japonicus*—Different phenotypes caused by the blockage of endocytotic bacterial release and nodule mutation, *Plant Cell Physiol.*, **38**, 871–881.
6. Szczygłowski, K., Shaw, R. S., Wopereis, J. et al. 1998, Nodule organogenesis and symbiotic mutants of the model legume *Lotus japonicus*, *Mol Plant Microbe Interact.*, **11**, 684–697.
7. Schauser, L., Handberg, K., Sandal, N. et al. 1998, Symbiotic mutants deficient in nodule establishment identified after T-DNA transformation of *Lotus japonicus*, *Mol. Gen. Genet.*, **4**, 414–423.
8. Schauser, L., Roussis, A., Stiller, J., and Stougaard, J. 1999, A plant regulator controlling development of symbiotic root nodules, *Nature*, **402**, 191–195.
9. Martirani, L., Stiller, J., Mirabella, R. et al. 1999, T-DNA tagging of nodulation- and root-related genes in *Lotus japonicus*: Expression patterns and potential for promoter trapping and insertional mutagenesis, *Mol. Plant Microbe Interact.*, **12**, 275–284.
10. Webb, K. J., Skot, L., Nicholson, M. N. et al. 2000, Mesorhizobium loti increases root-specific expression of a calcium-binding protein homologue identified by promoter tagging in *Lotus japonicus*, *Mol. Plant Microbe Interact.*, **13**, 606–616.
11. Szczygłowski, K., Hamburger, D., Kapranov, P., and de Bruijn, F. J. 1997, Construction of a *Lotus japonicus* late nodulin expressed sequence tag library and identification of novel nodule-specific genes, *Plant Physiol.*, **114**, 1335–1346.
12. Asamizu, E., Nakamura, Y., Sato, S., and Tabata, S. 2000, Generation of 7137 non-redundant expressed sequence tags from a legume, *Lotus japonicus*, *DNA Res.*, **7**, 127–130.
13. Endo, M., Kokubun, T., Takahata, Y., Higashitani, A., Tabata, S., and Watanabe, M. 2000, Analysis of expressed sequence tags of flower buds in *Lotus japonicus*, *DNA Res.*, **7**, 213–216.
14. Hyashi, M., Miyahara, A., Sato, S. et al. 2001, Construction of a genetic linkage map of the model legume *Lotus japonicus* using an intraspecific F₂ population, *DNA Res.*, **8**, 301–310.
15. Zhang, H.-B., Zhao, X., Ding, X., Paterson, A. H., and Wing, R. A. 1995, Preparation of megabase-sized DNA from plant nuclei, *Plant J.*, **7**, 175–184.
16. Liu, Y.-G., Shirano, Y., Fukaki, H. et al. 1999, Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning, *Proc. Natl. Acad. Sci. USA*, **96**, 6535–6540.
17. Sato, S., Kotani, H., Nakamura, Y. et al. 1997, Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones, *DNA Res.*, **4**, 215–230.
18. 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.
19. Uberbacher, E. C. and Mural, R. J. 1991, Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach, *Proc. Natl. Acad. Sci. USA*, **88**, 11261–11265.
20. Burge, C. and Karlin, S. 1997, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, **268**, 78–94.
21. Hebsgaard, S. M., Korning, P. G., Tolstrup, N. et al. 1996, Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information, *Nucl. Acids Res.*, **24**, 3439–3452.
22. Lowe, T. M. and Eddy, S. R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucl. Acids Res.*, **25**, 955–964.
23. Gordon, D., Abajian, C., and Green, P., 1998, Consed: a graphical tool for sequence finishing, *Genome Res.*, **8**, 195–202.
24. Neff, M. M., Neff, J. D., Chory, J., and Pepper, A. E. 1998, dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics, *Plant J.*, **114**, 387–392.
25. Riley, M. 1993, Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.*, **57**, 862–952.
26. Brick, D. J., Brumlik, M. J., Buckley, T. et al. 1995, A new family of lipolytic plant enzymes with members in rice, *Arabidopsis* and maize, *FEBS Lett.*, **377**, 475–480.