

## Structural Analysis of a *Lotus japonicus* Genome. II. Sequence Features and Mapping of Sixty-five TAC Clones Which Cover the 6.5-Mb Regions of the Genome

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

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

(Received 21 March 2002)

### Abstract

Sixty-five TAC (transformation-competent artificial chromosomes) clones were selected from a genomic library of *Lotus japonicus* accession MG-20 based on the sequence information of expressed sequences tags (ESTs), cDNA and gene information, and their nucleotide sequences were determined. The average insert size of the TAC clone was approximately 100 kb, and the total length of the sequenced regions in this study is 6,556,100 bp. Together with the nucleotide sequences of 56 TAC clones previously reported, the regions sequenced so far total 12,029,295 bp. By comparison with the sequences in protein and EST databases and by analysis with computer programs for gene modeling, a total of 711 potential protein-encoding genes with known or predicted functions, 239 gene segments and 90 pseudogenes were identified in the newly sequenced regions. The average gene density assigned so far was 1 gene/9140 bp. The average length of the assigned genes was 2.6 kb, which is considerably larger than that assigned in the *Arabidopsis thaliana* genome (1.9 kb for 6451 genes). Introns were identified in approximately 73% of the potential genes, and the average number and length of the introns per gene were 3.4 and 377 bp, respectively. Simple sequence repeat length polymorphism (SSLP) or derived cleaved amplified polymorphic sequence (dCAPS) markers were generated based on the nucleotide sequences of the genomic clones obtained, and each clone was mapped onto the linkage map using the F2 mapping population derived from a cross of two accessions 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

Comparative genomics is one of the most promising approaches to study genetic systems that are both unique and common to individual living organisms. In higher plants, the information on the entire genome and gene structures of *Arabidopsis thaliana*,<sup>1</sup> together with information on gene functions accumulated in this plant, can be used as a standard for comparison. We initiated a genome analysis of the legume *Lotus japonicus*<sup>2</sup> to understand the genetic system of legume species and to facilitate isolation and characterization of genes responsible for legume-specific phenomena. In this project, high-density linkage maps of the genome comprising six chromosomes have been generated,<sup>3</sup> and a large number of expressed sequence tags (ESTs) have been established.<sup>4</sup> Furthermore, we have determined the nucleotide sequences of the 5.4 Mb gene-rich regions of the

genome covered by 56 TAC (transformation-competent artificial chromosomes) clones and revealed the structures of more than 600 potential protein-coding genes.<sup>5</sup>

In this paper, we newly determined the nucleotide sequences of 6.5 Mb of the genome by sequencing the additional 65 TAC clones which were genetically mapped onto the six chromosomes. Gene organization and structural and functional information of the presumptive genes were deduced by computer-aided analysis. We also report the preliminary result of comparison of the gene organization between the two plant species, *L. japonicus* and *A. thaliana*.

### 1. Isolation and Sequencing of TAC Genomic Clones

DNA sources and the method of clone selection were the same as described previously.<sup>5</sup> The TAC genomic libraries<sup>6</sup> were generated from *L. japonicus* accession

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

MG-20,<sup>7</sup> arrayed in ninety-three 384-well microtiter plates, and 48 DNA pools each containing 384 clones were constructed for PCR screening. TAC clones for sequence analysis were isolated as follows. Sixty-five oligonucleotide pairs were synthesized on the basis of nucleotide sequences of the public ESTs and cDNAs from *L. japonicus*; GAP2, NLP2, NDX2, and RAC2 genes (provided by J. Stougaard); TED2, TUB1, and CP1 genes (provided by A. Suzuki); and a symbiosis-induced lipase-like gene (provided by M. Parniske), as listed in Table 1. The TAC libraries were screened for each tag sequence by PCR with the designed primers using the three-dimensional DNA pools. The nucleotide sequences of the PCR products were determined and compared with those of the corresponding tag sequences to confirm the authenticity of screening.

The nucleotide sequence of each TAC insert was determined according to the bridging shotgun method described previously.<sup>8</sup> The accumulated random sequences, the total of which correspond to about 6 times the equivalent of an insert, were assembled using the Phred-Phrap program (Phil Green, University of Washington, Seattle, USA). A lower threshold of acceptability for the generation of consensus sequences was set at a Phred score 20 for each base. For the regions where the data quality does not fulfill the prerequisite for the generation of consensus sequences mostly due to the presence of heavily repetitive sequences and secondary structures, they were left as gaps; therefore, the sequence of each TAC insert was registered as those of multiple contigs. 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 6,556,100 bp. The average GC content was 36%.

## 2. Assignment of Potential Protein- and RNA-Encoding Regions

Assignment of potential protein-encoding regions and gene modeling were performed by a combination of similarity search and computer prediction, as described previously.<sup>5</sup> Similarity search against the non-redundant protein sequence database nr (compiled by NCBI) was carried out using the BLASTX program.<sup>9</sup> In parallel, the positions of potential protein-encoding regions were predicted with the Grail,<sup>10</sup> GENSCAN<sup>11</sup> and NetGene2<sup>12</sup> 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. All the results obtained were compiled, and assignment of the potential protein-coding genes was carried out by taking both similarity to known genes and computer prediction into consideration, leaving the possibility that additional genes may be discovered among the intergenic regions in the future. The assigned genes are 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 the TAC clones thus revealed is exemplified in Fig. 1. All the genes, pseudogenes and gene segments assigned in each TAC clone and their organization are presented in our *L. japonicus* genome database at <http://www.kazusa.or.jp/lotus/>. To sum up, the complete structures of 711 potential protein-encoding genes and 90 pseudogenes containing either frame shifts or termination codons in the original coding regions were deduced in 6,556,100 bp of the genome. In addition, partial structures of 239 gene segments were identified at the termini of the clones and the contigs.

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 program<sup>13</sup> was also taken into account. As a result, a total of 13 tRNA genes corresponding to 11 amino acid species, tRNA-Ala (TGC), Ala (TGC), Arg (CCT), Asp (GTC), Cys (GCA), Glu (CTC), Gly (CCC), His (GTG), His (GTG), Lys (TTT), Met (CAT), Phe (GAA), and Ser (AGA) were identified. They are denoted by numbers with the clone names followed by "r" and sequential numbers.

## 3. Structural Features and Functional Assignment of Potential Protein-Encoding Genes

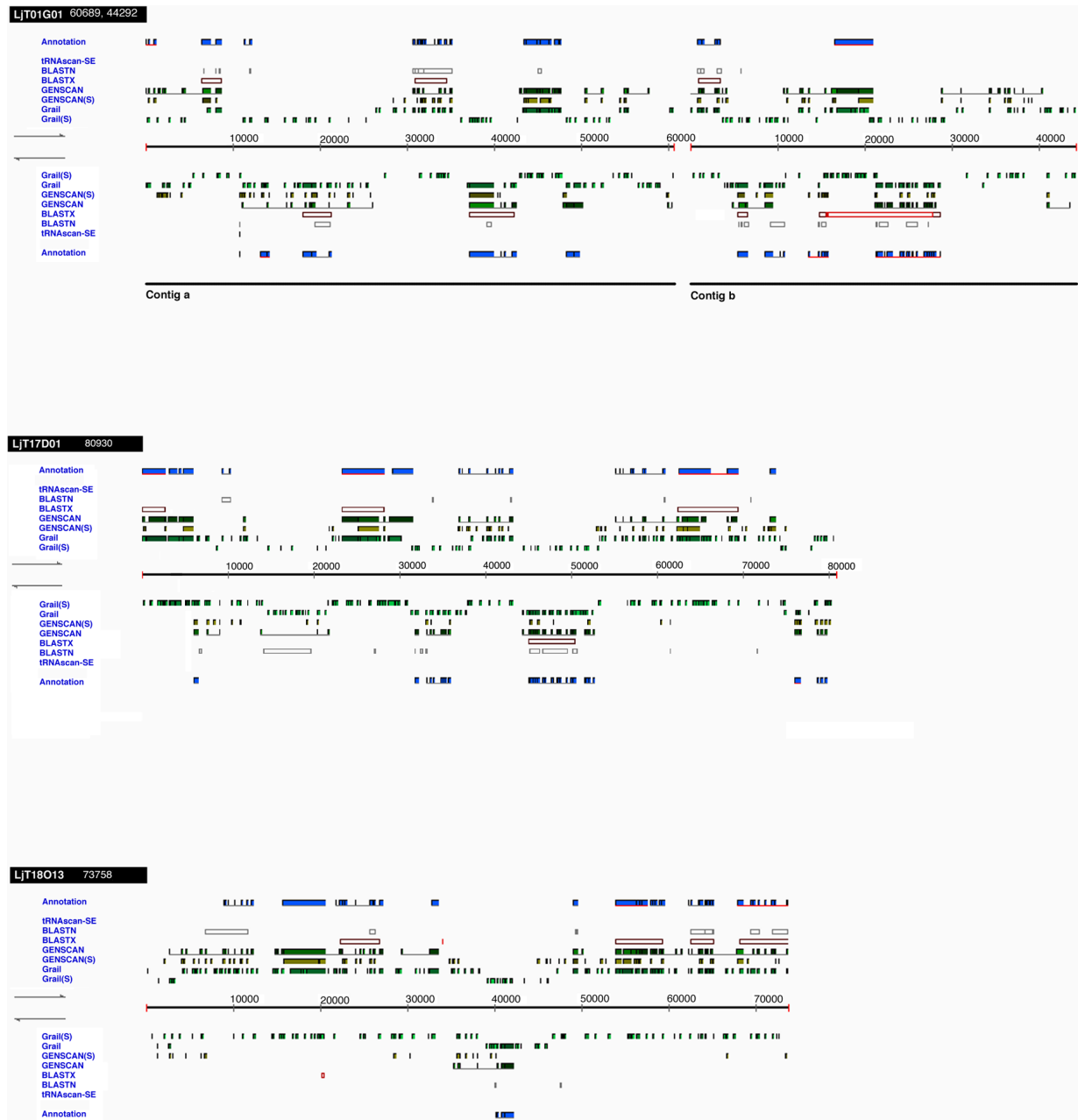
Table 2 shows the general features of the 1316 genes in *L. japonicus* which include those identified in the previous study, along with those of *A. thaliana*. The average length of the coding exons (297 bp) and the average number of introns per gene (3.4 introns) of *L. japonicus* were quite similar to those of *A. thaliana*. However, the average length of genes including introns (2581 bp) was longer in *L. japonicus* due to a longer average intron length. The average gene density of the sequenced regions of the *L. japonicus* genome was one gene in every 9140 bp, approximately half that of *A. thaliana*. This ratio, however, may be an over-estimate because the clones containing the expressed genes were selected according to the method taken in this study, therefore the genomic regions of higher gene contents were likely to be preferentially sequenced.

Similarity search of the 711 potential protein-encoding genes identified in this study against the nr databases indicated that 217 (31%) were homologous to genes of known function, 260 (37%) showed similarity to hypothetical genes (mostly of those in *A. thaliana*), and the remaining 234 (33%) showed no significant similarity to any registered genes. The 1316 potential protein-encoding genes including those previously reported were grouped into categories with respect to different biological roles according to the principle of

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. |
|--------|------------|---------------|-------------|-----|--|---------------|
| TM0024 | LjT17D01   | AP004894      | 80,930      | 5   | methylenetetrahydrofolate-CoA carboxylase biotin-binding chain | AV769566      |
| TM0039 | LjT02D01   | AP004895      | 91,268      | 1   | pectinesterase-like protein                                    | AV773018      |
| TM0040 | LjT16I07   | AP004896      | 92,281      | 5   | similar to early nodulins                                      | AV778942      |
| TM0047 | LjT01P23   | AP004897      | 16,512      | 3   | receptor protein kinase - like protein                         | AV779949      |
|        |            | AP004898      | 69,411      |     |  |               |
|        |            | AP004899      | 9,439       |     |  |               |
| TM0051 | LjT13M13   | AP004900      | 9,688       | 1   | porin  | AV770357      |
|        |            | AP004901      | 54,022      |     |  |               |
| TM0060 | LjT04G24   | AP004902      | 123,078     | 2   | unknown  | AU254056      |
| TM0061 | LjT01K08   | AP004903      | 94,052      | 4   | unknown  | AU254055      |
| TM0062 | LjT10C05   | AP004904      | 81,367      | 5   | unknown  | AU254058      |
| TM0064 | LjT06B17   | AP004905      | 14,870      | 1   | unknown  | AU254053      |
|        |            | AP004906      | 82,568      |     |  |               |
| TM0065 | LjT11D15   | AP004907      | 115,046     | 2   | hypothetical protein   | AU254052      |
| TM0066 | LjT12N11   | AP004908      | 79,585      | 6   | unknown  | AU254051      |
| TM0067 | LjT12O09   | AP004909      | 82,389      | 2   | putative protein   | AU254054      |
|        |            | AP004910      | 37,710      |     |  |               |
| TM0069 | LjT13O11   | AP004911      | 103,659     | 4   | unknown  | AV781180      |
| TM0070 | LjT11L19   | AP004912      | 128,164     | 3   | carboxypeptidase II  | AV775470      |
| TM0072 | LjT02A14   | AP004913      | 100,740     | 5   | unknown  | AV779935      |
|        |            | AP004914      | 28,924      |     |  |               |
| TM0073 | LjT16G15   | AP004915      | 90,077      | 4   | unknown  | AU254057      |
| TM0074 | LjT18E24   | AP004916      | 131,741     | 2   | unknown  | AV781012      |
| TM0075 | LjT16A13   | AP004917      | 83,881      | 4   | xyloglucan endo-transglycosylase-like protein                  | AV775434      |
| TM0076 | LjT13I21   | AP004918      | 71,794      | 2   | putative alanine aminotransferase                              | AV781241      |
|        |            | AP004919      | 4,524       |     |  |               |
|        |            | AP004920      | 67,145      |     |  |               |
| TM0078 | LjT18O13   | AP004923      | 73,758      | 1   | alcohol dehydrogenase  | AV774906      |
| TM0079 | LjT13P22   | AP004924      | 2,640       | 4   | similar to carboxylesterase family                             | AV775013      |
|        |            | AP004925      | 33,290      |     |  |               |
|        |            | AP004926      | 109,022     |     |  |               |
| TM0081 | LjT01G01   | AP004927      | 60,689      | 2   | unknown  | AV778972      |
|        |            | AP004928      | 44,292      |     |  |               |
| TM0083 | LjT06J16   | AP004929      | 120,016     | 3   | unknown  | AV781334      |
| TM0093 | LjT13G01   | AP004930      | 42,990      | 4   | MPB13.13   | AV406589      |
|        |            | AP004931      | 41,920      |     |  |               |
| TM0094 | LjT10J03   | AP004932      | 65,557      | 1   | MAC9.8   | AV766749      |
|        |            | AP004933      | 65,726      |     |  |               |
| TM0095 | LjT02L13   | AP004934      | 16,590      | 5   | unknown  | AU254043      |
|        |            | AP004935      | 6,348       |     |  |               |
|        |            | AP004936      | 63,750      |     |  |               |
| TM0096 | LjT08M07   | AP004937      | 24,390      | 5   | unknown  | AV781274      |
|        |            | AP004938      | 65,932      |     |  |               |
| TM0105 | LjT17M09   | AP004939      | 127,990     | 1   | Gap2 gene  | AF064788      |
| TM0106 | LjT48I11   | AP004940      | 103,636     | 3   | Nlp2 gene  |               |
| TM0109 | LjT25N10   | AP004941      | 75,234      | 1   | epoxide hydrolase  | AV770598      |
| TM0113 | LjT43P05   | AP004942      | 101,270     | 1   | late nodulin Nlj16   | AV769615      |
| TM0115 | LjT38E13   | AP004943      | 95,582      | 3   | lipoxigenase   | AV770419      |
| TM0117 | LjT43B20   | AP004944      | 33,359      | 1   | chitinase (EC 3.2.1.14) class I                                |               |
|        |            | AP004945      | 114,918     |     |  |               |
| TM0119 | LjT34D11   | AP004946      | 109,743     | 4   | naringenin,2-oxoglutarate 3-dioxygenase                        | AV774082      |
| TM0121 | LjT35I07   | AP004947      | 29,246      | 1   | plastocyanin precursor   | AV771845      |
|        |            | AP004948      | 56,086      |     |  |               |
| TM0122 | LjT45F11   | AP004949      | 90,864      | 1   | protein kinase MSK-3 (EC 2.7.1.-)                              | AV771922      |
| TM0123 | LjT42E10   | AP004950      | 79,601      | 1   | ubiquitin-conjugated enzyme E2                                 | AV773802      |
| TM0124 | LjT26I01   | AP004951      | 79,571      | 2   | auxin-responsive GH3-like protein                              | AV769747      |
| TM0125 | LjT19C08   | AP004952      | 89,551      | 1   | putative receptor-like protein kinase                          | AV781323      |
| TM0129 | LjT34A24   | AP004953      | 110,487     | 3   | MADS-box protein   | AV773694      |
| TM0131 | LjT21G09   | AP004954      | 93,645      | 4   | LIM15-like protein   | AV777574      |
| TM0132 | LjT19K21   | AP004955      | 59,303      | 1   | homeobox-leucine zipper protein HAT9                           | AV768245      |
| TM0133 | LjT34I04   | AP004956      | 106,411     | 1   | beta-ketoacyl-CoA synthase (FIDDLEHEAD)                        | AV767338      |
| TM0134 | LjT34H20   | AP004957      | 132,605     | 2   | sucrose transport protein SUT1                                 | AV779447      |
| TM0135 | LjT20B10   | AP004958      | 96,988      | 3   | MAP3K epsilon protein kinase                                   | AV781139      |
| TM0136 | LjT21I12   | AP004959      | 81,332      | 3   | arabinogalactan protein  | AV770846      |
| TM0139 | LjT19B18   | AP004960      | 106,089     | 6   | ARG10 gene   | AV773981      |
| TM0140 | LjT21P04   | AP004961      | 86,648      | 6   | unknown  | AU254049      |
| TM0141 | LjT28H14   | AP004962      | 100,810     | 1   | unknown  | AU254050      |
| TM0142 | LjT48D11   | AP004963      | 78,601      | 3   | unknown  | AU254048      |
| TM0143 | LjT20J05   | AP004964      | 7,894       | 1   | unknown  | AU254046      |
|        |            | AP004965      | 83,991      |     |  |               |
|        |            | AP004966      | 13,449      |     |  |               |
| TM0144 | LjT27L02   | AP004967      | 89,959      | 1   | unknown  | AU254047      |
| TM0145 | LjT43O24   | AP004968      | 134,976     | 1   | unknown  | AV781176      |
| TM0146 | LjT34P02   | AP004969      | 110,752     | 5   | SCUTL2   | AV775811      |
| TM0148 | LjT30P03   | AP004970      | 106,000     | 5   | remorin 1  | AV773761      |
| TM0151 | LjT45G21   | AP004971      | 109,264     | 5   | unknown  | AU254044      |
| TM0155 | LjT41A07   | AP004972      | 32,123      | 3   | unknown  | AV776151      |
|        |            | AP004973      | 52,067      |     |  |               |
|        |            | AP004974      | 33,822      |     |  |               |
| TM0158 | LjT21J12   | AP004975      | 83,377      | 5   | GM043  | AV423889      |
| TM0159 | LjT09L18   | AP004976      | 91,722      | 3   | Ndx2 homeobox gene   | AJ011829      |
| TM0160 | LjT12A10   | AP004977      | 74,313      | 3   | Rac small GTPase (Rac2) gene                                   | AF063867      |
| TM0162 | LjT44L05   | AP004978      | 92,619      | 4   | multifunctional transport intrinsic membrane protein 2         | AU254045      |
|        |            | AP004979      | 30,916      |     |  |               |
| TM0436 | LjT13N17   | AP004980      | 23,709      | 3   | TED2 gene  |               |
|        |            | AP004981      | 43,428      |     |  |               |
| TM0437 | LjT04L20   | AP004982      | 48,314      | 6   | TUB1 gene  |               |
|        |            | AP004983      | 51,960      |     |  |               |
| TM0438 | LjT44N06   | AP004984      | 107,764     | 1   | CP1 gene   |               |
| TM0476 | LjT33P04   | AP004985      | 36,654      | 1   | symbiosis induced lipase-like gene                             |               |
|        |            | AP004986      | 75,682      |     |  |               |

6,556,100



**Figure 1.** Gene organization in the TAC clones LjT01G01 (in two contigs a and b), LjT17D01 and LjT18O13. The positions of the predicted genes and gene segments in each clone are schematically presented by color-coded boxes above (rightward) and below (leftward) the solid line in the middle which represents the contig or the entire insert. The insert length is given in number together with the clone name in the top box. Gray arrows indicate the directions of the DNA strands (5' to 3'). Blue bars with gray and red underlines represent the positions of the identified potential protein encoding genes and partial/pseudo genes, respectively. Gray open boxes indicate the positions of the EST hits reported by BLASTN search. The regions which showed similarity to the sequences in the protein database are shown by light to dark red open boxes, each of which corresponds to BLASTX scores from low to high in gradation. The green bars indicate the positions of the potential exons and presumptive gene structures predicted by the GENSCAN and Grail programs. Color depth from yellow to dark green corresponds to increasing scores of the predicted exons. GENSCAN(S) and Grail(S) indicate GENSCAN suboptimal exons and Grail shadow exons, respectively.

Riley,<sup>14</sup> and the numbers of genes in each category are summarized in Table 3.

#### 4. Characteristic Features of Potential Protein-Encoding Genes

##### 4.1. Expression of potential protein-encoding genes

The number of the matched *L. japonicus* ESTs was counted to roughly monitor the transcriptional level of

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

| Features                           | <i>L. japonicus</i><br>1,316 genes <sup>a)</sup> | <i>A. thaliana</i><br>6,451 genes <sup>b)</sup> |
|------------------------------------|--|---|
| Gene length (bp) including introns | 168-21,418 (2,581)                               | 78-17,203 (1,918)                               |
| Product length (amino acids)       | 16-2,036 (433)                                   | 25-4,706 (427)                                  |
| Genes with introns                 | 960 (73%)  | 4,906 (76%)                                     |
| Number of intron/gene              | 0-37 (3.4)                                       | 0-48 (4.0)                                      |
| Exon length (bp)                   | 3-5,604 (297)                                    | 2-5,966 (256)                                   |
| Intron length (bp)                 | 30-6,718 (377)                                   | 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 1,316 genes are assigned in this and previous studies <sup>a)</sup> and the 6,451 genes<sup>b)</sup> previously assigned potential protein genes in our *A. thaliana* genome sequencing project. Average values are shown in parentheses.

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

|   | %            |              |
|---|--------------|--------------|
| Amino acid biosynthesis                                     | 8            | 0.6          |
| Biosynthesis of cofactors, prosthetic groups, and carriers  | 12           | 0.9          |
| Cell envelope   | 10           | 0.8          |
| Cellular processes  | 23           | 1.7          |
| Central intermediary metabolism                             | 8            | 0.6          |
| Energy metabolism   | 23           | 1.7          |
| Fatty acid, phospholipid and sterol metabolism              | 13           | 1.0          |
| Photosynthesis and respiration                              | 7            | 0.5          |
| Purines, pyrimidines, nucleosides, and nucleotides          | 4            | 0.3          |
| Regulatory functions  | 50           | 3.8          |
| DNA replication, recombination, and repair                  | 2            | 0.2          |
| Transcription   | 17           | 1.3          |
| Translation   | 16           | 1.2          |
| Transport and binding proteins                              | 26           | 2.0          |
| Other categories  | 185          | 14.1         |
| <b>Subtotal of genes similar to genes of known function</b> | <b>404</b>   | <b>30.7</b>  |
| Similar hypothetical protein                                | 460          | 35.0         |
| <b>Subtotal of genes similar to registered genes</b>        | <b>864</b>   | <b>65.7</b>  |
| No similarity   | 452          | 34.3         |
| <b>Total</b>  | <b>1,316</b> | <b>100.0</b> |

each of the potential protein-encoding genes assigned in this study. Similarity search was carried out against the *L. japonicus* ESTs both in-house (40,555 ESTs, manuscript in preparation) and in the public DNA databases (31,567 ESTs)<sup>4,15,16</sup> with a lower threshold of 95% identity for a region of 50 nucleotides. Of the 711 genes identified in this study, 227 (32%) carried EST sequences. Among the EST-matched genes, 11 (1.5%) 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 extracellular dermal glycoprotein (LjT19B18.5), 40S ribosomal protein S2 (LjT13O11.14), lipoxygenase (LjT38E13.11, LjT38E13.5), chitinase (LjT43B20b.1), outer plastidial membrane protein (LjT13M13b.2), extensin precu-

ror (LjT10C05.6), root nodule extensin (LjT10C05.7), *S*-adenosyl-L-methionine synthetase (LjT20J05a.1), alanine aminotransferase (LjT13I21a.7), arabinogalactan protein (LjT21I12.12).

#### 4.2. Retroelement-related genes

As reported in the previous paper,<sup>5</sup> significant portions of the genes in the *L. japonicus* genome are structurally related to those for retroelements. Actually, 217 genes out of the 1040 potential protein-encoding genes and gene segments (21%) assigned in this study were those for retroelements (*gag* and *pol*). It should also be noted that 176 of these 217 genes contained frame shifts or termination codons within the coding regions.

Table 4. List of SSR markers.

Table with columns: clone name, marker name, SSR pattern, motif, number of repeats, product size (bp) for MG-20, B-129, and heteroduplex, Fw primer (5 to 3), Rv primer (5 to 3), Chr, and variant(s) (bp).

\* markers on the region of translocation between MG-20 and B-129. a) mapped as a B-129 dominant marker.

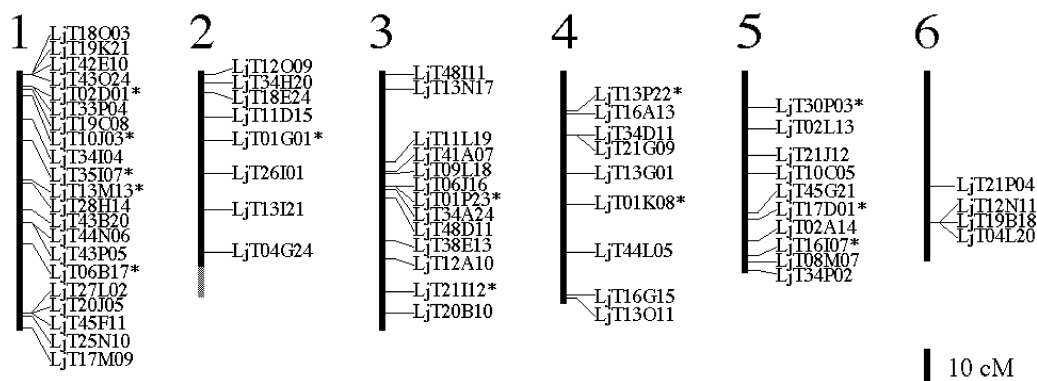
Table 5. List of dCAPS markers.

Table with columns: clone name, marker name, Enzyme, product size (bp) for MG-20 and B-129, Fw primer (5 to 3), Rv primer (5 to 3), and Chr.

4.3. Redundant genes

The presence of tandem arrays of repeated genes is one of the characteristic features of the higher plant genomes, as reported in A. thaliana1 and L. japonicus.5 Such gene arrays were often found in the L. japonicus genome sequenced in this study. These include genes for germin-like protein (LjT06B17b.3, LjT06B17b.5, LjT06B17b.10, LjT06B17b.12, LjT04L20a.3, LjT04L20a.4, LjT04L20a.5, LjT04L20a.8),

disease- resistance protein (LjT16G15.4, LjT16G15.8, LjT16G15.12), receptor serine/threonine protein kinase (LjT16G15.15, LjT16G15.16), 100-kDa coactivator protein (LjT10J03a.4, LjT10J03a.6, LjT10J03a.8), epoxide hydrolase (LjT25N10.11, LjT25N10.12), putative sec14 cytosolic factor (LjT43P05.10, LjT43P05.11), lipoxygenase (LjT38E13.5, LjT38E13.7, LjT38E13.11, LjT38E13.12), anther-specific proline-rich protein APG precursor (LjT02L13c.8, LjT02L13c.9, LjT34A24.2,



**Figure 2.** 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 previous report.<sup>3</sup> Asterisks indicate the dCAPS markers.

LjT34A24.3), arabinogalactan protein (LjT21I12.12, LjT21I12.13, LjT21I12.14), extracellular dermal glycoprotein (LjT19B18.5, LjT19B18.9, LjT19B18.10, LjT19B18.11, LjT19B18.12, LjT19B18.15), transcription factor Hap5a-like (LjT19B18.17, LjT19B18.18), putative hydrolase (LjT21J12.6, LjT21J12.7, LjT21J12.8), nodulin-26 (LjT44L05a.7, LjT44L05a.9, LjT44L05a.10), hypothetical protein F4I1.22 (LjT13N17a.3, LjT13N17a.4) and putative reticuline oxidase (LjT04L20b.8, LjT04L20b.11). Many of these genes including those for germin-like protein, disease-resistance protein, receptor serine/threonine protein kinase, epoxide hydrolase, putative sec14 cytosolic factor, anther-specific proline-rich protein APG precursor, extracellular dermal glycoprotein, transcription factor Hap5a-like, nodulin-26, and putative reticuline oxidase were also redundant and formed tandem arrays in the genome of *A. thaliana*.

## 5. Linkage Mapping of TAC Clones

To localize the sequenced clones onto the genetic linkage map of *L. japonicus*, two types of PCR-based DNA markers, SSLP and dCAPS, were generated using the sequence information of each clone, as described previously.<sup>5</sup> Linkage mapping was performed using the F<sub>2</sub> population of two accessions of *L. japonicus*, Miyakojima MG-20 and Gifu B-129. Primer sequences for PCR amplification of the markers, product sizes for both accessions, restriction enzymes for digestion and expected fragment sizes are listed in Tables 4 and 5. Of the 65 generated markers, 52 were SSLP and 13 were dCAPS, and all of these markers except one (TM0060) were co-dominant markers. The position of each marker was integrated onto the fine linkage map of the *L. japonicus* chromosomes previously reported (Fig. 2).<sup>3</sup> This map represents the linkages of MG-20 chromosomes, and strongly suggests the presence of a segmental translocation between chromosome 1 of Gifu B-129 and chromo-

some 2 of Miyakojima MG-20, indicated as a gray bar.<sup>3</sup>

The sequence data, gene information and mapping information are available through the World Wide Web at <http://www.kazusa.or.jp/lotus/>.

**Acknowledgements:** We thank S. Sasamoto, A. Watanabe, K. Kawashima, T. Kimura, Y. Kishida, M. Kohara, M. Matsumoto, A. Matsuno, A. Muraki, S. Nakayama, S. Shinpo, M. Sugimoto, C. Takeuchi, M. Yamada, and T. Wada for excellent technical assistance. Thanks are also due to Drs. M. Parniske, J. Stougaard, and A. Suzuki 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. Hayashi, M., Miyahara, A., Sato, S. et al. 2001, Construction of a genetic linkage map of the model legume *Lotus japonicus* using an intraspecific F<sub>2</sub> population, *DNA Res.*, **8**, 301–310.
4. 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.
5. Sato, S., Kaneko, T., Nakamura, Y. et al. 2001, Structural analysis of *Lotus japonicus* genome. I. Sequence features and mapping of fifty-six TAC clones which cover the 5.4 Mb regions of the genome, *DNA Res.*, **8**, 311–318.
6. 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.
7. Kawaguchi, M. 2000, *Lotus japonicus* “Miyakojima”

- MG-20: an early flowering accession suitable for indoor handling, *J. Plant Res.*, **113**, 507–509.
8. 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.
  9. 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.
  10. 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.
  11. Burge, C. and Karlin, S. 1997, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, **268**, 78–94.
  12. 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.
  13. 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.
  14. Riley, M. 1993, Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.*, **57**, 862–952.
  15. Szczyglowski, 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.
  16. 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.