DNA Research **15**, 227–239, (2008) doi:10.1093/dnares/dsn008

Genome Structure of the Legume, *Lotus japonicus*

Shusei Sato[1], Yasukazu Nakamura[1], Takakazu Kaneko[1], Erika Asamizu[1], Tomohiko Kato[1], Mitsuteru Nakao[1], Shigemi Sasamoto[1], Akiko Watanabe[1], Akiko Ono[1], Kumiko Kawashima[1], Tsunakazu Fujishiro[1], Midori Katoh[1], Mitsuyo Kohara[1], Yoshie Kishida[1], Chiharu Minami[1], Shinobu Nakayama[1], Naomi Nakazaki[1], Yoshimi Shimizu[1], Sayaka Shinpo[1], Chika Takahashi[1], Tsuyuko Wada[1], Manabu Yamada[1], Nobuko Ohmido[2], Makoto Hayashi[3], Kiichi Fukui[3], Tomoya Baba[4], Tomoko Nakamichi[5], Hirotada Mori[5], and Satoshi Tabata[1,*]

Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan[1]; Graduate School of Human Development and Environment, Kobe University, Kobe 657-8501, Japan[2]; Department of Biotechnology, Graduate School of Engineering, Osaka University 2-1 Yamadaoka, Suita 565-0871, Osaka, Japan[3]; Institute of Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan[4] and Graduate School of Biological Sciences, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan[5]

(Received 1 April 2008; accepted on 18 April 2008; published online 28 May 2008)

Abstract

The legume *Lotus japonicus* has been widely used as a model system to investigate the genetic background of legume-specific phenomena such as symbiotic nitrogen fixation. Here, we report structural features of the *L. japonicus* genome. The 315.1-Mb sequences determined in this and previous studies correspond to 67% of the genome (472 Mb), and are likely to cover 91.3% of the gene space. Linkage mapping anchored 130-Mb sequences onto the six linkage groups. A total of 10 951 complete and 19 848 partial structures of protein-encoding genes were assigned to the genome. Comparative analysis of these genes revealed the expansion of several functional domains and gene families that are characteristic of *L. japonicus*. Synteny analysis detected traces of whole-genome duplication and the presence of synteny blocks with other plant genomes to various degrees. This study provides the first opportunity to look into the complex and unique genetic system of legumes.

Key words: *Lotus japonicus*; genome structure; Fabaceae; comparative analysis

1. Introduction

Fabaceae is the third largest family of flowering plants, comprising 650 genera and 18 000 species with a variety of characteristics; many of which have long been targets of breeding because of their agronomic and industrial importance. Among them, a few species have been chosen as 'model legumes' for use in genetic and physiological studies. *Lotus japonicus* is a typical model legume with the characteristics of a short life cycle (2–3 months), self-fertility, and a relatively simple genome architecture of diploidy ($n = 6$), i.e. small in size, 472 Mb. Mutants in various biological phenomena specific to legumes such as symbiotic nitrogen fixation, and those common to flowering plants such as flower morphogenesis, have been characterized and the genes responsible have been isolated and further studied. The availability of the *Agrobacterium*-mediated DNA transformation system and genomic resources including a large number of expressed sequences tag (EST)/cDNA clones,[1] high-density genetic linkage maps,[2–4] and partial genome sequences[5–9] has played an essential role in this process.

The whole-genome sequences of two plant species, *Arabidopsis thaliana* (Cruciferae) and *Oryza sativa* (Poaceae), have drastically accelerated research into their genetic systems by providing investigators with both gene sequences and positional information.

Edited by Katsumi Isono

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

However, the process of high-accuracy sequencing is expensive and labor intensive since physical maps are created and a large number of individual BAC clones are used as templates. In contrast, draft sequencing according to the whole-genome shotgun approach in combination with BAC end sequencing generates information about the general genome structure at a fairly reasonable cost, but the resulting data are rather fragmentary and need an additional effort to be connected with physical/genetic maps. In this study, we aimed to analyze the genome structure of *L. japonicus* to reveal gene and genome features that are characteristic of legume plants. For this purpose, we adopted various established technologies including genomic library construction, DNA sequencing, fluorescent *in situ* hybridization (FISH), genetic mapping and bioinformatics in such a way that the advantages of each technology were combined in a cost-effective manner. Here, we report the first whole-genome structure of the legume, *L. japonicus*, its characteristic genome features, and a variety of information and material resources that were developed during this study.

2. Material and methods

2.1. Plant and DNA materials

Lotus japonicus accession Miyakojima MG-20 was provided by Masayoshi Kawaguchi, University of Tokyo.[10] Transformation-competent artificial chromosome (TAC) genomic libraries were constructed according to standard methods as described previously.[5] BAC genomic libraries were constructed using the genomic DNA of *L. japonicus* accession MG-20 partially digested with either *Sau*3AI or *Eco*RI and pBeloBAC as a cloning vector. The average insert size of these libraries was 104 kb for the *Sau*3AI library and 101 and 88 kb for two independent preparations of *Eco*RI libraries. Both libraries covered the haploid genome 8.4 times in total.

The seeds and genomic libraries can be obtained from LegumeBase supported by the National BioResource Project (http://www.legumebase.agr.miyazaki-u.ac.jp/index.jsp).

2.2. Genome sequencing and assembly

Two types of sequencing approaches were combined to sequence the *L. japonicus* genome: clone-by-clone sequencing and shotgun sequencing of selected regions of the genome.

TAC/BAC clones were selected from the genomic libraries as seed points using the sequence information from ESTs and cDNA markers from *L. japonicus* and other legumes. The nucleotide sequence of each clone was determined according to the shotgun strategy with three to five times redundancy. A total of 1909 TAC/BAC clones, those newly sequenced in this study and those that had been sequenced previously,[5–9] were assembled into 954 scaffolds using the Paracel Genome Assembler (PGA; version 2.6.2, Paracel Co., 2002), followed by manual TAC/BAC end-pair scaffolding, resulting in high-quality genomic sequence (HGS) contigs.

In parallel, shotgun sequencing of a selected TAC mixture (STM) enriched in gene spaces and a whole genomic DNA from which highly repetitive and organelle genomic sequences were subtracted (selected genomic regions, SGRs) was carried out. The TAC clones, neither end sequence of which hit repetitive or organelle genomic sequences in the *L. japonicus* genome, were selected from the libraries, pooled, and subjected to shotgun sequencing. For the SGRs, a genomic library with an average insert size of 2.5 kb was generated using pBluescript SK− as the cloning vector. For subtraction, polymerase chain reaction (PCR)-amplified fragments of LjTR1 were biotinylated using Biotin-High Prime (Roche, Basel, Switzerland) and used as a driver in subtractive hybridization with the WGS library. The WGS library was single-stranded prior to hybridization by combined action of gene II and exonuclease III. Hybrids were removed using Dynabeads M-280 Streptavidin (Invitrogen, Carlsbad, CA, USA) and the remaining single-stranded WGS library was double-stranded using Klenow fragments (Takara Bio, Japan) and transformed into host *E. coli* ElectroTen-Blue (Agilent Technologies, Santa Clara, CA, USA).

A total of 808 816 reads from STM generated from 4603 TAC inserts and 847 513 SGR reads were assembled into a set of 109 986 contigs, 147 805 446 bp in length (selected genome assembly, SGA) by the Arachne assembler, version 2.01.[11] The SGA sequences were then subjected to assemble with the HGS, and finally, a total of 110 940 supercontigs with a total coverage of 315 073 275 tentative genomic sequence (TGS) bases were obtained.

2.3. Linkage mapping

Two types of PCR-based DNA marker, SSLP and dCAPS, were generated using the sequence information in order to locate sequence contigs on the genetic linkage map, as described previously.[5] The 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, which were previously used for construction of the fine genetic linkage map.[2]

2.4. Fluorescent in situ hybridization

The 26S rDNA, 5S rDNA, and TAC clones were used as probes for FISH analysis. The 26S and the 5S rDNA

probes were produced by PCR using primer pairs that were designed based on the rRNA and 5S RNA gene sequences in the *L. japonicus* genome.

The FISH analysis using rRNA genes and TAC clones was performed on well-prepared chromosome spreads according to the method described previously.[12] The preparations were observed under a fluorescence microscope (OLYMPUS BX50) equipped with a sensitive cooled CCD camera (PXL1400), and the pro-metaphase chromosome spreads with clear patterns were photographed using blue or green light excitation and emission filters. Captured images were digitally stored in a computer and analyzed using CHIAS3 imaging software.[13]

2.5. Repetitive sequences

Repetitive elements in the TGS were identified by comparing all of the contig units, HGS and those produced from the STM and SGRs, each using BLASTN[14,15] and processing the outputs using the RECON program.[16] A total of 214 consensus sequences of repetitive elements that appeared at least 20 times were identified. The consensus sequences of these elements were subjected to a similarity search against known repeat elements in the RepBase (http://www.girinst.org/). For the consensus sequences with features of Class I or II transposable elements (TEs), full-length candidate sequences were identified by comparing 10 kb upstream and downstream of the corresponding genomic regions to find long terminal repeats or terminal inverted repeats. For unclassified consensus sequences, the longest representative sequences were selected by comparing the corresponding genome sequences using the CLUSTALW multiple alignment program. Full-length elements of TE and representative sequences of unclassified repeats were collected into a repeat sequence library along with the previously reported TE sequences[17] and used as references for RepeatMasker (www.repeatmasker.org) analysis to delineate the occurrence of these elements in the total TGS and TAC/BAC end sequences.

2.6. Assignment of RNA-encoding genes

In order to identify the potential RNA-encoding genes, a structural RNA sequence library was extracted from GenBank,[18] and searched for homologous sequences in the *L. japonicus* genome with the use of the BLASTN function in BLAST. Transfer RNA genes were predicted using tRNAscan-SE, version 1.4,[19] in a eukaryotic mode with default parameters. Genes for small nucleolar RNAs (snoRNAs) and 7S large nucleolar RNAs (LRNAs) were predicted using SnoScan[20] with a yeast model and srpSCAN,[21] respectively. A total of 93 genes for small nuclear RNAs including those for 21 U1, 19 U2, 7 U4, 23 U5, 21 U6, and 2 U12 were identified based on their similarity to known genes.

Candidates for miRNA precursors were identified by a search using the *Arabidopsis* mature miRNA sequences, which were retrieved from the miRNA Registry database (http://www.sanger.ac.uk/Software/Rfam/mirna/).[22] The *L. japonicus* genome and EST sequences were searched for sequence patterns that are characteristic of miRNA using the FUZZNUC program that is available at EMBOSS (http://emboss.sourceforge.net/) with the permission of a maximum of two nucleotide mismatches. A stem-loop structure was searched within a distance of 500 nucleotides from the coding region of the mature miRNA based on the finding that the longest *Arabidopsis thaliana* precursor sequences identified to date are 411 nucleotides long.[23] Target genes of the miRNA candidates were searched in the coding sequences (CDS) and 3′ untranslated regions (UTRs) of the presumptive genes assigned in the TGS.

2.7. Assignment of protein-encoding genes

Tentative genomic sequence was subjected to gene prediction and modeling by the Kazusa Annotation PipelinE for *Lotus japonicus* (KAPSEL).[5] The KAPSEL employs *ab initio* gene-finding software and similarity searches in order to generate the elements for gene model production. The *ab initio* gene-finding software used in the pipeline includes GeneMark.hmm,[24] Genscan[25] and Grail[26] using the *A. thaliana*-trained matrix. Splice-site candidates were deduced by NetGene2[27] and SplicePredictor.[28] The similarity searches to detect potential protein-coding exons were performed using the BLASTX function of BLAST against the UniProtKB database.[29] The assigned exon candidates were extracted from the original sequence library, then mapped on the TGS more precisely using the dps and nap programs in the program suite of the analysis and annotation tool (AAT) package.[30] Similarity searches of transcript sequences were performed by aligning the TGS against the Gene Indices[31] for legume species including *L. japonicus, M. truncatula and Glycine max*. The assigned transcript sequences were mapped on the TGS using the dds and gap2 programs in AAT to confirm working models of protein-encoding genes. As a result of the automated annotation process, a total of 19 848 partial and 10 951 complete models were assigned as protein-encoding genes in the TGS, except for those related to TEs. The 76.4-Mb sequences in the HGS were edited and annotated manually to ensure high-quality gene prediction.

The genes thus assigned were denoted by IDs with the clone (LjT**** for TACs and LjB**** for BACs) or contig (CM****) names followed by sequential numbers from one end to another. Of these, manually annotated

genes on the HGS were followed by ".nc", and others were followed by ".nd". The genes assigned on the SGA sequences were denoted by IDs with the assemble consensus names (LjSGA_****) followed by sequential numbers from one end to another in the insert.

A global alignment of the genome sequences and ESTs was performed using the NEEDLE program[32,33] that is provided at the EMBOSS site (http://emboss.sourceforge.net/). To identify a possible TATA box-like motif for recognition by RNA polymerase II, a search against the plant cis-acting regulatory DNA elements (PLACE) database[34] (http://www.dna.affrc.go.jp/PLACE/) was carried out.

2.8. Similarity and domain searches

Functional annotation for deduced *L. japonicus* proteins was performed by a similarity search against the genes of known function and a domain analysis. The similarity search was performed using the gapped BLASTP function of BLAST against the UniProtKB database and protein-encoding genes deduced in *A. thaliana*,[35] rice,[36] *Populus trichocarpa*,[37] and grapevine.[38] For the analysis of gene families and functional domains, the predicted proteome was searched against InterPro.[39] InterPro annotations were assigned to each functional motif and domain, and then the annotation information was translated into GO functional descriptions.[40] GO descriptions were grouped into the GOslim categories of molecular function and biological process. Subcellular localization of targeting signals and transmembrane helices of deduced protein-coding genes were predicted using the following programs: PSORT,[41] TargetP,[42] and SOSUI.[43]

2.9. Synteny analysis

Synteny was detected by identifying arrays of predicted protein-encoding genes between target genomic regions. Translated amino acid sequences of the products of genes assigned on the mapped TGS were compared with those in the reference genomes, and a BLASTP *E*-value of $<1e-20$ was considered to be significant. Synteny blocks were surveyed on the basis of physically linked sequence units such as contigs and singlet clones. A synteny block was defined as the region where three or more conserved homologs were located within a 100-kb DNA stretch in the two genomes.

3. Results and discussion

3.1. Genome sequencing

Preliminary random and TAC end sequencing revealed the presence of various types of repetitive sequences in the *L. japonicus* genome. The FISH analysis showed that some of these sequences were distributed along the entire genome with occasional condensation to different extents, strongly suggesting that the gene spaces are intermingled with repeated sequences. In order to efficiently obtain sequence information for the gene space, we combined two independent approaches: clone-by-clone sequencing from seed points of the genome and shotgun sequencing of both STM enriched in gene spaces and a whole-genomic DNA from which highly repetitive and organelle genome sequences were subtracted (SGRs).

A total of 1314 TAC and BAC clones were selected based on the sequences of ESTs, cDNA and gene information from *L. japonicus* and other legumes, and an additional 584 TAC and BAC clones were selected by overlaps. Their nucleotide sequences were determined according to the shotgun method with three to five times redundancy (Supplementary Table S1). Further assembly of the sequences of the 1898 clones produced 954 supercontigs with a total length of 167 267 829 bp. Among these sequences, 76 366 532 bp (46%) covered by 823 clones were Phase 3 sequences (HGS). In parallel, 4603 TAC clones, neither ends of which hit the repetitive sequences, were pooled and subjected to shotgun sequencing. Assembly of 808 816 STM random sequences and 847 513 SGR sequences, the sum of which was estimated to give 2.4 times the genome coverage, generated an additional 109 986 contigs that were 147 805 446 bp long in total (SGA). The efficiency of the STM and SGR approaches was indicated by the ratio of centrometic and heterochromatic repeat sequences in SGA sequences, which was about one-fifth of that in TAC/BAC end sequences (Supplementary Fig. S1). The total length of the determined sequences (TGS) amounted to 315 073 275 bp (Supplementary Table S2). Although the TGS corresponded to 67% of the reported *L. japonicus* genome (472 Mb),[44] it can be estimated that the TGS covers 91.3% of the gene space because 11 404 out of 12 485 collections of tentative consensus (TC) sequences of the *L. japonicus* Gene Index provided by the Gene Index Project (http://compbio.dfci.harvard.edu/tgi/plant.html) were located on the TGS. TGS was used as standard information for further analysis of gene structure and function in *L. japonicus* in this study.

3.2. Construction of the sequence-tagged genetic linkage map

In order to anchor the obtained genomic sequences to the genetic linkage map, DNA markers were developed for the sequenced TAC and BAC clones, and genetic mapping was carried out. As a result, a total

of 788 microsatellite and 80 derived cleaved amplified polymorphic sequences (dCAPS) markers were generated and mapped onto the six linkage groups, which resulted in anchoring of the HGS and the connecting TAC/BAC and SGR sequences onto the linkage map. In total, 594 supercontigs containing 130 251 279 bp, corresponding to 41% of the TGS, were anchored onto the genetic linkage map (Supplementary Table S1).

Authenticity of the above genetic linkage map was examined by FISH. TAC clones genetically mapped at distal and proximal positions to the linkage groups were used as probes for hybridization. As shown in Fig. 1, all of the clones examined were successfully located at the expected positions on the corresponding chromosomes. For chromosome 5, four clones were positioned inversely in the expected order on the genetic linkage map; therefore, we revised the marker order of the genetic map. Genetic and physical distances did not agree very well, as reported in *A. thaliana* and rice.[35,36]

The positions of centromeres were deduced by cytological features in both prometaphase and pachytene chromosomes and were further confirmed by FISH analysis using the pericentromere-specific retroelement LjRE2 (described later) as a probe (Fig. 1). TAC clones genetically mapped at distal positions of each linkage group were located on the telomeric regions of all the chromosomes by FISH with the exception of the bottom of chromosome 4 and the top of chromosome 6 (Fig. 1). However, none of these clones and the extending sequences contained the consensus telomeric repeat, CCCTAAA.[45] Small condensed structures were observed at most of the telomeric regions of the chromosomes.[44] The presence of such heterochromatic regions may have prevented the extension of the analyzed sequences to the telomeric repeat. Indeed, a short tandem repeat sequence, LjTR4, specific to the subtelomeric region was found on the contig (CM0105) at the bottom of chromosome 1, where no telomeric condentation was observed.[44]

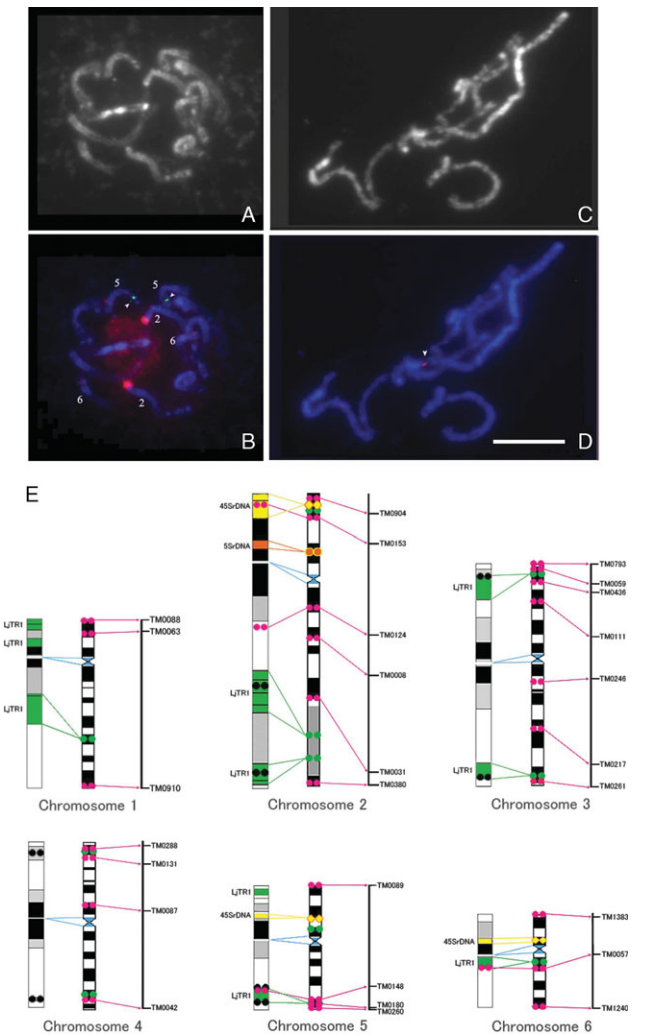

Figure 1. FISH detection and the integration map of *L. japonicus*. (**A** and **C**) Mitotic prometaphase chromosomes and meiotic pachytene chromosomes were stained with DAPI. (**B**) TAC clone, LjT30P03 (corresponding marker: TM0148) (green, arrowhead) was detected on the long arm of chromosome 5. 45S rDNA (red) was detected on the short arms of chromosomes 2, 5, and 6. (**D**) LjT30P03 (corresponding marker: TM0148) (green, arrowhead) was detected on chromosome 5 of the pachytene chromosome. The error bar represents 5 μm. (**E**) Integration among three maps; mitotic prometaphase chromosome map, meiotic pachytene chromosome map, linkage map of *L. japonicus*. Red circles show TAC clones, yellow and orange show ribosomal RNA genes, green shows tandem repeat LjTR1, blue shows retrotransposon LjRE2 representing the centromere. The length ratio among six chromosomes was adjusted to the ratio of pachytene chromosomes.

3.3. *Repetitive sequences*

A total of 33 730 di-, tri-, and tetra-nucleotide simple sequence repeats (SSRs) that were equal to or longer than 15 bp were identified in the TGS (Supplementary Table S3). Provided that the size of the TGS is 315 Mb, the frequency of occurrence of the above SSRs was estimated to be one SSR in every 9.3 kb. Di-, tri-, and tetra-nucleotide SSRs accounted for 48.6, 44.4, and 7.0% of the identified SSRs, respectively. The SSR patterns that appeared frequently were (AT)n, (AAG)n, and (AAAT)n, each representing 63% of di-, 28% of tri-, and 46% of tetra-nucleotide repeat units. The tri-nucleotide SSRs, particularly (GGT)n and (GGA)n, were preferentially found in exons. (AG)n was enriched in 5′ and 3′ UTRs, and (AC)n frequently occurred in 5′ and 3′ UTRs and introns (Supplementary Table S3).

A search using the repeat sequence finding program RECON[14] against the TGS unraveled the occurrence of a variety of repeat elements including Class I and Class

Table 1. Repetitive sequences in the *L. japonicus* TGS

	Copy number	Coverage (kb)	Fraction of genome (%)
Class I			
SINEs	252	31.4	0.01
LINEs	13 225	387.1	1.19
LTR: Ty1/copia	25 421	23 225.2	7.16
LTR: Ty3/gypsy	36 445	28 606.3	8.81
Other LTR	11 194	6660.4	2.05
Total class I	86 537	62 394.4	19.23
Class II			
Coding class II	11 786	3160.0	0.97
MITE	35 393	7568.4	2.33
Total class II	47 179	10 728.4	3.31
Short tandem repeats	5147	4259.5	1.31
Unclassified	109 361	33 851.3	10.43

II TE subfamilies and those that are difficult to classify into known subfamilies (Table 1). Over 35 000 copies of miniature-inverted repeat TEs (MITEs) were identified, which were classified into more than 60 types. A substantial portion of these MITEs was found in introns and UTRs. A short insert size Class I retroelement, a member of the short-interspersed nucleotide elements, was also preferentially observed in introns and the 3′ UTRs (Supplementary Table S4).[46]

As TGS were enriched in sequences from the gene-rich euchromatic regions, we used less biased TAC end sequences to survey the highly repetitive sequences in the entire genome of *L. japonicus*. By clustering 37 000 TAC end sequences, 34 types of highly repetitive sequences, each of which appeared in more than 50 end sequences, could be identified. These included five short tandem repeats with a repeat unit ranging from 172 to 333 bp (LjTR1 to 5), two retroelements (LjRE1 and 2) and nine unclassified repeats (LjUR1 to 9), one of which is the centromere-associated sequence Ljcen1 (GenBank accession no. AF390569) reported previously by Pedrosa et al.[45] The FISH analysis showed that LjRE2 was present exclusively in the pericentromeric heterochromatic regions, while LjRE1 was distributed along the entire chromosomes (Supplementary Fig. S2), and that LjTR1 was localized in highly condensed heterochromatic regions of the genome known as the chromosome knobs (Supplementary Fig. S2).

3.4. Genes encoding non-coding RNAs

A combination of prediction and similarity searches against the TGS identified 638 transfer RNA genes corresponding to 49 species of anticodons (Supplementary Table S5), which are sufficient for coding all of the amino acids in the universal codon table. In addition, two complete units of 18S–5.8S–26S ribosomal RNA genes and two or more copies of 5S ribosomal RNA genes with an interval of approximately 195 nucleotides were found in the TGS, strongly suggesting successful enrichment of non-repetitive regions of the genome. As to other structural RNA genes, 207 genes for snoRNAs, seven high-confidence predictions for 7S LRNAs, and 93 genes for small nuclear RNAs were deduced.

Genome-wide transcriptome analyses have identified a large number of non-coding RNA species being transcribed in the genome.[47,48] Using the previously accumulated information from *L. japoncius* transcripts by EST[1] and SAGE[49] analyses, we searched for transcripts derived from non-coding (intergenic) genomic regions. As a result, 1300 ESTs without similarity to known and predicted protein sequences were obtained, and 460 SAGE tag species solely identified in the SAGE analysis were identified as possible transcripts from intergenic regions in the genome. Since ESTs are partial sequences and SAGE tags originate from the 3′-most anchoring enzyme site of the transcripts, these 'intergenic' candidates may include those from UTRs of genes. Gene prediction of the genome sequence is done based on the presence of coding sequences of appropriate length. It is, therefore, possible that transcripts derived from unpredicted regions on the genome may encode a protein with a short open reading frame or that functions at the RNA level.

In order to identify the miRNA precursor candidate in the *L. japoncius* genome, we performed a search using the *A. thaliana* mature miRNA sequences deposited in the miRNA registry database in the Sanger Institute (http://microrna.sanger.ac.uk/sequences/). As a result, we found 53 precursor sequence candidates encoding 26 miRNA species (Supplementary Table S6). Among the miRNAs detected only in *A. thaliana* (eudicots), but not in other species like mosses, ferns, gymnosperms and monocots,[50] miRNAs 393, 395, 398, and 399 were found in the *L. japonicus* genome. In order to examine whether other leguminous plants possess these miRNAs, the TC sequences of *G. max* and *Medicago truncatula* were subjected to a search. As a result, miRNAs 395, 398, and 399 were found in the TC of at least one of the two legumes whereas miRNA 393 was detected in the *M. truncatula* BAC sequence (AC147434). We therefore speculate that these four miRNAs developed concurrently with the appearance of eudicots. We also focused on the copy number of miRNA genes (precursors). There are six genes encoding miR156 in *A. thaliana*, whereas only one was found in the *L. japonicus* genome. On the contrary, multiple genes in *L. japonicus* such as those encoding miRNAs

171, 399, 164, 166, 169, 172, and 319 were also found in multiple in *A. thaliana* (data not shown). An increase in the copy number of conserved miRNA species was not observed, indicating the possibility that conserved types of miRNAs had not been recruited in legume-specific gene regulatory mechanisms. We suppose that a significant part of gene regulation by miRNA in leguminous species is done by those that evolved specifically in the legume lineage. To identify species-specific miRNAs, a computational prediction was performed and 1312 miRNA candidates were found in the *L. japonicus* genome sequence, which can be grouped into 605 families (L. Schauser, personal communication).

We attempted the identification of antisense transcription in the *L. japonicus* genome by comparing coding sequences with overlapping ESTs on both sense and antisense DNA strands. To discriminate between ESTs of antisense transcripts and wrongly oriented clones, only 3′ ESTs with a poly(T) signal and 5′ ESTs without a poly(T) signal were selected for this analysis. As a result, we found 322 such presumptive protein-coding genes. We suspect that some of them were derived from unselected antisense transcription; however, we obtained several pieces of evidence that indicate selective antisense transcription. We checked for the sequences of the exon–intron splice junctions and TATA box-like motif proximal to the antisense transcripts, and as a result, (an) intron(s) could be predicted in antisense transcripts of 41 genes, where the donor and acceptor sequence of the exon–intron splice junction was GT-AG in 49 cases and GC-AG in one case. We found TATA box-like motifs proximal to the antisense transcripts in ~60% of the 322 protein-coding genes.

3.5. Characteristic features of protein-encoding genes

Tentative genomic sequence was subjected to the automatic assignment of protein-encoding genes, and 10 951 complete and 19 848 partial gene structures, except those for transposons/retrotranspons, could be identified. Of the 30 799 presumptive protein-encoding genes, 16 096 (52%) carried ESTs with sequence identity of over 95% for a stretch 50 nucleotides long. Structural features of the protein-encoding genes in *L. japonicus* were investigated in detail using 5248 genes manually annotated in the 76.4 Mb from HGS. As shown in Supplementary Table S7, the basic structure of protein-encoding genes in *L. japonicus* is similar to that of *A. thaliana*[35] except for the average lengths of genes and introns: 2917 versus 1918 bp and 395 versus 157 bp in *L. japonicus* and *A. thaliana*, respectively. The average gene density in *L. japonicus* was estimated to be one gene in every 10.2 kb (30 799 genes in 315 073 kb), which is longer than that in *A. thaliana*, one gene in every 4.5 kb. The total gene number in the *L. japonicus* genome could be deduced to be 34 245 by calculations based on the following assumptions: (i) the average density of the protein-encoding genes in the euchromatic regions is one gene in every 10.2 kb, (ii) the TGS is 315 073 kb long, and (iii) gene-space coverage of the TGS is estimated to be 90.2%.

A similarity search of translated amino acid sequences of the presumptive 30 799 protein-encoding genes was performed against the UniProt knowledgebase as a protein sequence library.[29] The result indicated that 19 405 (63%) genes had significant ($<$1e$-$20 cutoff of BLAST *E*-value) sequence similarity to genes of known function, 6790 (22%) were similar to hypothetical genes and the remaining 4604 (15%) did not show any similarity to the registered gene sequences. Eighty-five percent (24 212 genes out of 28 581), 51% (34 098 genes out of 66 710), 74% (31 456 genes out of 42 373), and 73% (22 328 genes out of 30 434) of the protein-encoding genes deduced from genomic sequences of *A. thaliana*,[35] rice,[36] *P. trichocarpa*,[37] and grapevine,[38] respectively, showed significant sequence similarity ($<$1e$-$20) to those in *L. japonicus*. The presumptive protein-encoding genes were classified into functional categories according to GO,[40] and the result was summarized in plant GOslim functional categories based on the InterPro domain, as shown in Fig. 2 and Supplementary Table S8.

3.5.1. Transcription factors Genes encoding transcription factors (TFs) were identified in the *L. japonicus* genome by assessing the TrEMBL and InterPro search results. The classification of TFs was performed according to the TF categories in the Arabidopsis Gene Regulatory Information Server at the Ohio State University (http://arabidopsis.med.ohio-state.edu/AtTFDB/). A total of 1481 genes were classified into 31 TF categories (Supplementary Table S9). EST matching provided evidence of transcription to 604 (41%) predicted TF genes (data not shown). TF genes classified into the C3H family were most frequently observed in the *L. japonicus* genome (183 genes), followed by 156 MYB and 142 AP2-EREBP family genes. While the distribution of TFs in *L. japonicus* and *A. thaliana* was generally similar (Supplementary Table S9), we found 75 ARF family genes in the *L. japonicus* genome, which is more than three times the number in *A. thaliana* (24 genes). ARF binds to auxin response elements in the promoter region of auxin-regulated genes.[51] We found a relatively large number of unique sequences showing similarity to ARF genes in the Gene Index Project, 87 (31 singletons and 56 TCs) and 33

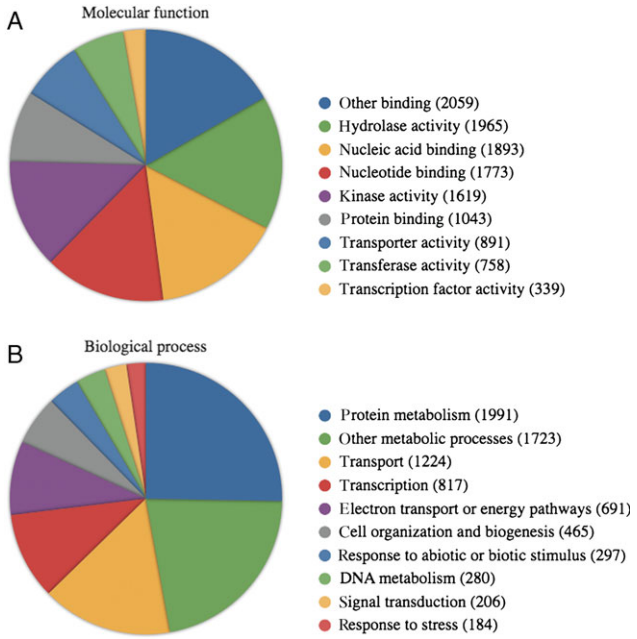

Figure 2. Plant GOslim category classifications. The results of functional categorization are shown for 10 representative GOslim classes of 'Molecular function' (**A**) and 'Biological process' (**B**). GO was automatically annotated on the *L. japonicus* protein-encoding genes by searching InterPro domains.

(13 singletons and 20 TCs), of *G. Max* and *M. truncatula*, respectively. An increased gene number (more than double) in the *L. japonicus* genome compared to *A. thaliana* was also observed for CAMTA and SBP family genes. CAMTA is a conserved TF among multicellular eukaryotes that is suggested to be involved in calcium signaling by direct binding of calmodulin.[52] SBP is a plant specific TF that is characterized by a conserved 76 amino acid protein domain.[53] The observed characteristics of the *L. japonicus* genome as well as those of *G. max* and *M. truncatula* suggest the possibility that some TF families have diverged to adapt to biological processes characteristic of leguminous plants such as nodulation and symbiotic nitrogen fixation with rhizobia.

3.5.2. Two-component systems A two-component system is a mechanism that is conserved in a wide variety of intracellular signal transduction pathways in response to certain external or internal stimuli in prokaryotes; plants have come to adopt this system as one of their multiple signal transduction mechanisms. This system is typically composed of histidine kinases (HKs), response regulators (RRs), and less typically some intermediators, all of which participate in the phosphor-transfer reaction. It has been shown that the two-component elements are involved in signaling of the plant hormones, ethylene, and cytokinin in *A. thaliana*,[54] and that these hormones play crucial roles in controlling nodule formation in legumes.[55–57]

In the *L. japoncius* genome, 23 genes for HKs, including five and nine genes for ethylene and cytokinin receptors, respectively, were identified by similarity and domain searches. *Lotus* HK 1 (LHK1) is a member of the cytokinin receptor family and plays a crucial role in initiating cortical cell divisions in response to rhizobial signaling;[55,56] this gene was located on chr 4 (chr4.CM0042.1660.nc), with a highly conserved paralog on chr 2 (chr2.LjB11M03.80.nc). A total of 20 genes for RRs, 11 of which have domains characteristic of DNA-binding transcriptional factors in their C-terminal portions, were identified in the *L. japonicus* genome. However, two of them (chr4.CM0042.1440.nc and chr4.CM0229.220.nc) had a replacement of the phospho-accepting Asp residue with Gly or Ala residues, suggesting that these genes might have lost the ability to participate in the His–Asp phosphor-relay pathway. It has been reported that a group of pseudo-response regulators (PRRs), in which the phospho-accepting Asp residue is replaced by Glu residue, are known to be distinct from the typical RR in that they are not involved in any phosphor-relay signaling. Five such genes including TOC1 in *A. thaliana* are known to be involved in the circadian rhythm.[58] In *L. japonicus*, seven genes with a pseudo-receiver residue (Glu) were identified for PRR. One of these genes, chr4.CM0087.600.nc, was most similar to TOC1 (APRR1), while the other two genes, chr3.CM0208.230.nc and chr3.TM0616.60.nd, are likely to be candidates for clock components because they share a CCT motif in the C-terminal portion with the APRR components.

3.5.3. Receptor-like kinases Receptor-like kinases (RLKs) play important roles in sensing diverse signals in plant growth, development, and defense responses. A typical RLK consists of a signal sequence, diverse extracellular domains, a transmembrane region, and an intracellular eukaryotic protein kinase domain. RLKs in the *A. thaliana* genome form the largest class of transmembrane sensors and are encoded by more than 600 genes.[59] The presumptive products of 1267 genes in the *L. japonicus* genome contained the protein kinase domain (InterProID, IPR000719), whereas those of 362 genes exhibited a domain composition that was similar to known RLKs. It has been reported that RLKs with leucine-rich repeats (LRRs: IPR001611) such as SYMRK (chr2.CM0177.340.nc)[60] and HAR1 (chr3.CM0216.560.nc)[61,62] were essential in the process of symbiotic nitrogen fixation in *L. japonicus*. The classification based on InterPro indicated that the presumptive products of 124 genes contained the LRR (IPR001611) in their extracellular region.

The cascade of legume plant responses leading to nodule formation is triggered by recognition of lipochitooligosaccharide Nod factors that are produced

by rhizobia by a family of receptor kinases containing the LysM domain (IPR002482) in their extracellular region.[63] Presumptive products of 12 RLK genes in *L. japonicus* contained the exracellular LysM domain including Nfr1 (chr2.CM0545.460.nc) and Nfr5 (chr2.CM0323.400.nd), which have been proposed to be Nod-factor receptors, while only five such genes were found in *A. thaliana*. Therefore, it can be speculated that some of the remaining nine LysM–RLKs may participate in signal transduction during the process of nodule formation and pathogen recognition, as reported in *A. thaliana* and rice.[64]

3.5.4. Transporters Transporters play indispensable roles in various cellular processes including the delivery of nutrients, elimination of waste products, and control of hormone efflux. As symbiotic nitrogen fixation requires the exchange of nutrients between host legume plant cells and intracellular rhizobia, a variety of transporters should be involved. A total of 1310 genes encoding possible transporters were deduced in the *L. japonicus* genome based on their sequence similarity to 995 *A. thaliana* genes registered in the TransportDB (http://www.membranetransport.org/). These genes were subsequently classified into 49 gene families. Some of these families were composed of a large number of paralogs (97–205 members), including the ATP-binding cassette (ABC) superfamily,[65] the P-type ATPase (P-ATPase) superfamily, the drug/metabolite transporter superfamily, and the major facilitator superfamily. They account for a significant portion of the large number of transporter genes in *L. japonicus*, as is the case in *A. thaliana*.

One of the transporter families that are characteristic to legumes encodes major intrinsic membrane proteins (MIPs), which are known to control water and solute homeostasis and constitute 10–15% of the total membrane proteins in the symbiosome. Eleven of the 37 genes for MIPs assigned in the *L. japonicus* genome were paralogous members of the LIMP2 gene (chr4.CM0046.520.nc), a multifunctional aquaglyceroporin, which exhibits specific expression in mature nitrogen fixing nodules.[66] The LIMP2 gene and seven of the paralogs were clustered on chr 4 and chr 6 in tandem arrays. Proteome analysis in the nodule of *L. japonicus* has identified several types of peribacteroid membrane proteins. LIMP1 is one such protein known to encode a water-specific aquaporin.[67] A gene for LIMP1 (chr1.CM0012.450.nc) as well as two paralogs (chr1.LjT36P09.70.nd and LjSGA_047026.1) were assigned in the *L. japonicus* genome.

3.5.5. Cytochrome P450 gene family One of the characteristic features of legume plants is the production of a large variety of secondary metabolites including legume-specific 5-deoxyisoflavonoid. Among the genes encoding enzymes involved in the biosynthesis of secondary metabolites, those for Cytochrome P450 (Cyt P450) play crucial roles. The genes for Cyt P450 constitute a relatively large gene family, which are further classified into sub-families on the basis of nomenclature files on the Cyt P450 Homepage (http://drnelson.utmem.edu/cytochromeP450.html). A total of 313 putative genes for Cyt P450 were assigned to the *L. japonicus* genome (Supplementary Table S10), whereas 272 and 455 genes for Cyt P450 have been reported in *A. thaliana* and rice, respectively.[68] Many of the Cyt P450 genes in *L. japonicus* formed tandem arrays, as in *A. thaliana* and rice, where those in a single array belong to the same gene family. The composition of some of the sub-families of Cyt P450 genes in *L. japonicus* was remarkably different from that of *A. thaliana*. Significant amplification of the members was observed in the sub-families CYP74A, CYP75B, CYC76C, CYP82C, CYP83B, CYP93C, and CYP716A, when compared with *A. thaliana*. Of these, the CYP75B and CYP93C sub-families have been reported to be involved in flavonoid synthesis;[68,69] thus, expansion of these sub-families may be involved in the complexity of flavonoids in *L. japonicus*.

3.5.6. Disease resistance genes Most of the disease resistance genes (R genes) encode intracellular proteins with a centrally located nucleotide binding site (NBS) and a C-terminal LRR region. This NBS–LRR family of R genes can be classified into two subgroups, toll and interleukin-1 receptor (TIR) and non-TIR, based on whether or not the putative gene product contains a region similar to the TIR domain at their amino termini. We identified 174 TIR NBS–LRR-type and 55 non-TIR NBS–LRR-type R genes and gene segments in the TGS of *L. japonicus* (Supplementary Table S11). The ratio of TIR versus non-TIR is similar to that found in *A. thaliana* (85 TIR NBS–LRR versus 36 non-TIR NBS–LRR), but differs from that in *P. trichocarpa* (64 TIR NBS–LRR versus 233 non-TIR NBS–LRR).[37] Fifty-five of the 120 (83 TIR NBS–LRR and 37 non-TIR NBS–LRR) presumptive R genes assigned in the contiguous sequences of single or connected TAC/BAC clones were found as singletons, while the remaining 65 formed gene clusters, each consists of two to seven members of the same subfamily. Eighty-seven presumptive R genes genetically located on the linkage map were distributed unevenly among chromosomes: 17 (14 TIR NBS–LRR and 3 non-TIR NBS–LRR) on chr 1, 33 (18 and 15) on chr 2, 20 (15 and 5) on chr 3, 7 (5 and 2) on chr 4, 5 (0 and 5) on chr 5, and 5 (2 and 3) on chr 6 (Supplementary Table S11).

3.5.7. Genes conserved in leguminous plants In order to identify genes conserved in legumes but

not in other plant species, predicted gene sequences of *L. japonicus* were subjected to comparison with predicted genes on the genomes of *A. thaliana*, rice, *P. trichocarpa* and grapevine as well as protein sequences in the TrEMBL protein database. Predicted gene sequences from the *M. truncatula* genome and *G. max* gene index were used as references for legume protein coding genes. Translated amino acid sequences of the predicted genes in *L. japonicus* were subjected to BLAST searches, and an *E*-value of $<1e-20$ was considered to be significant. As a result, we found 1190 genes (4%) that were conserved in leguminous species, but not in the currently available protein sequences of other plant species. InterPro annotations of these genes were surveyed to investigate the conserved motifs identified in these legume-specific genes, and 31 InterPro motifs were found to be conserved in five or more genes. Of these motifs, Cyclin-like F-box (IPR001810), LRR 2 (IPR013101), TIR (IPR000157) and trypsin-alpha amylase inhibitor (IPR003612) were found in 61, 26, 13 and 11 genes, respectively (Supplemental Table S12). Furthermore, a total of 2824 predicted genes in the *L. japonicus* genome were without a counterpart in the databases. Among them, complete gene structures were predicted for 599 genes; the most common InterPro motif was the RNA-binding region RNP-1 (IPR000504) (Supplementary Table S13). A similarity search against the *L. japonicus* ESTs indicated that at least 718 of these genes (25.4%) are transcribed. The entire lists of the legume- and *L. japonicus*-specific genes are provided in Supplementary Tables S14 and S15.

3.6. *Gene amplification and genome rearrangement*

In order to detect gene amplification at the genome-wide level, reciprocal comparisons among the sequences of the presumptive protein-encoding genes in the *L. japonicus* genome were carried out using the BLASTP program. As a result, 56% of the genes (17 380 out of 30 799) matched one or more genes (with the cut-off $<1e-20$): 14% had one matched gene, 7% had two genes, and the remaining 35% had three or more matches. It should be noted that these results may be underestimated because genes that are truncated at the ends of a large number of contigs and those assigned to sequences of low accuracy in the TGS are likely to be overlooked under the above conditions. Indeed, when the genes assigned to the 1909 TAC/BAC clones whose sequences had a stretch of $\sim$100 kb of high accuracy were subjected to search, 65% of the genes (10 151 out of 15 512) occurred as multiple copies; this is the same rate as that observed in *A. thaliana*. Further searches using the 76.4 Mb HGS, where the relative positions of most of the genes was reliable, showed that 772 genes formed 310 groups of tandem arrays of two or more family genes (Supplementary Table S16). This corresponds to 7% of the genes assigned in the HGS (11 380), which is lower than those observed in *A. thaliana* (17%), rice (14%), and *P. trichocarpa* (11%). One possible explanation could be that frequent insertion of retroelements in the *L. japonicus* genome have disrupted tandem array structures, which is consistent with our observation that 9% of the genes assigned in HGS had family members in a range of 100 kb .

During the course of the comparative analysis of the *L. japonicus* genome with those of other plants, we occasionally found that two distinct regions of the *L. japonicus* genome showed synteny with a single locus of other genomes. When these regions were plotted on the chromosomes, traces of large scale segmental duplications, possibly because of whole-genome duplication (WGD), became apparent (Fig. 3). These regions corresponded to 39% (231 out of 594) of the mapped TGS. Duplication of large blocks was found between chr 2/4, chr 4/6, chr 2/6, chr 1/3, chr 1/5, and chr 3/3. Only 13% (880 out of 6819) of the genes assigned in these regions were conserved in both of the duplicated regions, which is significantly lower than that between the syntenic regions of *L. japonicus* and *M. truncatula* where about 60% of the predicted genes were conserved. This observation supports the estimation that WGD occurred before the separation of *L. japonicus* and *M. truncatula* described in a previous report.[70]

3.7. *Comparative analysis with genomes of model plants and crop legume*

Microsynteny against the *M. truncatula* genome was detected for 70% (418 out of 594) of the mapped

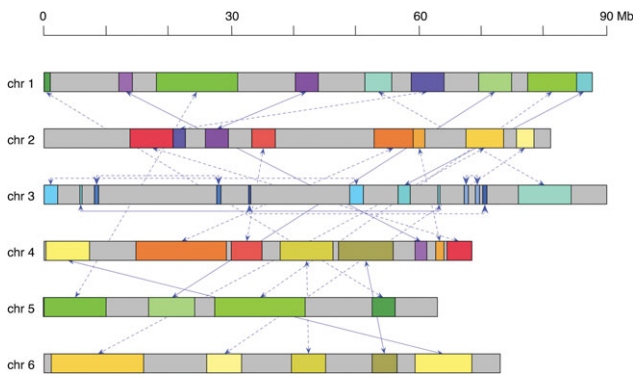

Figure 3. Segmentally duplicated regions in the *L. japonicus* genome. Individual chromosomes are depicted as horizontal bars. Colored boxes with the same color code connected by the blue line indicate each pair of duplicated segments. Duplicated segments in reversed orientation are connected with dotted lines. The scale is in megabases.

TGS in *L. japonicus*, although traces of local duplications, inversions, deletions, and insertions were also observed. By combining the syntenic relations assigned at the clone and contig level in the TGS with their positional information on the genetic linkage map, the microsynteny could be expanded to macrosynteny at the level of whole chromosomes, which confirmed and strengthened previous observations.[70] A syntenic relationship has also been detected between the genomes of *L. japonicus* and *A. thaliana* to a lesser degree. Conservation of relative positions of three or more genes has been observed in 69% (408 out of 594) of the mapped TGS of *L. japonicus.* This limited degree of syntenic relationship was proved to be useful for map-based cloning in *L. japonicus*.[71] A similar level of synteny was detected against *P. trichocarpa* where 69% (409 out of 594) of the mapped TGS of *L. japonicus* exhibited syntenic relations. The microsyntenic relationships between the above plant species may provide useful information for predicting gene organization in the ancestral genome of dicot plants.

3.8. Conclusions

To date, two types of sequencing strategy have been adopted in order to fully sequence plant genomes: clone-by-clone sequencing based on physical maps and whole-genome shotgun sequencing. Both strategies have their advantages and disadvantages in terms of labor, cost, speed, and data accuracy. In this study, we modified the two strategies into seed-based clone-by-clone and selected genome shotgun sequencing, combined them, and anchored the obtained sequences onto maps by linkage analysis and FISH. This strategy takes advantage of the strengths of the conventional methods, and provides information that meets the majority of requirements for a wide variety of uses in fundamental and applied biology. It can therefore be applied with appropriate modifications to the genomes of diverse plants species, especially large crop genomes that will be interesting to sequence in the future.

The nucleotide sequences and their positional information generated in this study revealed characteristic features of the legume genomes: the expansion of functional domains and gene families for TFs, elements of the two-component system, transporters, P450 proteins, proteins involved in disease resistance, and the presence of unique genes such as those for receptor kinases essential for symbiotic nitrogen fixation and those specific to legume plants. Synteny analysis among the sequenced plant genomes detected traces of WGD and subsequent genome evolution in *L. japonicus*, and also the presence of synteny blocks with genomes of model and crop plants to various degrees. Conservation of colinearity of the genes and synteny blocks on the chromosomes among plant species indicates that the genome information for *L. japonicus* could serve as a hub for transferring knowledge toward the improvement of crop legumes. The detailed and comprehensive analyses of the genome information reported in this study may lead to the identification of a variety of legume- and plant-specific genetic systems.

The genome sequence information is available through the international databases (DDBJ/Genbank/EMBL) under accession numbers DF093176–DF093536 (for genome clone contigs), AP009625–AP010255 (for genome clone sequences) and BABK01000001–BABK01109986 (for SGA sequences). The online database that provides nucleotide sequences of the pseudo-molecules with annotations describing the predicted genes as well as information on the DNA markers and the genetic linkage maps is available at http://kazusa.or.jp/lotus/.

Acknowledgements: We thank Drs. T. Aoki and L. Schauser for valuable discussions on the genes involved in secondly metabolism and the genes encoding miRNAs, respectively. We also acknowledge Drs. D. Holligan and S. R. Wessler for providing the sequence information on transposable elements.

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

Funding

This work was supported by the Kazusa DNA Research Institute Foundation. Funding for open access charge: The Kazusa DNA Research Institute Foundation.

References

1. Asamizu, E., Nakamura, Y., Sato, S. and Tabata, S. 2004, Characteristics of the *Lotus japonicus* gene repertoire deduced from large-scale expressed sequence tag (EST) analysis, *Plant Mol. Biol.*, **54**, 405–414.
2. Hayashi, 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.
3. Sandal, N., Krusell, L., Radutoiu, S., et al. 2002, A genetic linkage map of the model legume *Lotus japonicus* and strategies for fast mapping of new loci, *Genetics*, **161**, 1673–1683.
4. Sandal, N., Petersen, T. R., Murray, J., et al. 2006, Genetics of symbiosis in *Lotus japonicus*: recombinant inbred lines, comparative genetic maps, and map position of 35 symbiotic loci, *Mol. Plant Microbe Interact.*, **19**, 80–91.

5. Sato, S., Kaneko, T., Nakamura, Y., et al. 2001, 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, *DNA Res.*, **8**, 311–318.
6. Nakamura, Y., Kaneko, T., Asamizu, E., et al. 2002, 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, *DNA Res.*, **9**, 63–70.
7. Kaneko, T., Asamizu, E., Kato, T., et al. 2003, Structural analysis of a Lotus japonicus genome. III. Sequence features and mapping of sixty-two TAC clones which cover the 6.7 Mb regions of the genome, *DNA Res.*, **10**, 27–33.
8. Asamizu, E., Kato, T., Sato, S., et al. 2003, Structural analysis of a *Lotus japonicus* genome. IV. Sequence features and mapping of seventy-three TAC clones which cover the 7.5 Mb regions of the genome. *DNA Res.*, **10**, 115–122.
9. Kato, T., Sato, S., Nakamura, Y., et al. 2003, Structural analysis of a *Lotus japonicus* genome. V. Sequence features and mapping of sixty-four TAC clones which cover the 6.4 Mb regions of the genome, *DNA Res.*, **10**, 277–285.
10. Kawaguchi, M. 2000, *Lotus japonicus* 'Miyakojima' MG-20: an early-flowering accession suitable for indoor handling, *J. Plant Res.*, **113**, 507–509.
11. Jaffe, D. B., Butler, J., Gnerre, S., et al. 2003, Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96.
12. Sato, S., Isobe, S., Asamizu, E., et al. 2005, Comprehensive structural analysis of the genome of red clover (*Trifolium pratense* L.), *DNA Res.*, **12**, 301–364.
13. Kato, S., Ohmido, N. and Fukui, K. 2003, Development of a quantitative pachytene chromosome map in *Oryza sativa* by imaging methods, *Genes Genet. Syst.*, **78**, 155–161.
14. 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.
15. Altschul, S. F., Madden, T. L., Schäffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucl. Acids Res.*, **25**, 3389–3402.
16. Bao, Z. and Eddy, S. R. 2002, Automated de novo identification of repeat sequence families in sequenced genomes, *Genome Res.*, **12**, 1269–1276.
17. Holligan, D., Zhang, X., Jiang, N., Pritham, E. J. and Wessler, S. R. 2006, The transposable element landscape of the model legume *Lotus japonicus*, *Genetics*, **174**, 2215–2228.
18. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. 2007, GenBank, *Nucl. Acids Res.*, **32**, 21–25.
19. 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.
20. Lowe, T. M. and Eddy, S. R. 1999, A computational screen for methylation guide snoRNAs in yeast, *Science*, **283**, 1168–1171.
21. Regalia, M., Rosenblad, M. A. and Samuelsson, T. 2002, Prediction of signal recognition particle RNA genes, *Nucl. Acids Res.*, **30**, 3368–3377.
22. Griffiths-Jones, S. 2004, The microRNA registry, *Nucl. Acids Res.*, **32** (Database issue), D109–D111.
23. Jones-Rhoades, M. W. and Bartel, D. P. 2004, Computational identification of plant microRNAs and their targets, including a stress-induced miRNA, *Mol. Cell*, **14**, 787–799.
24. Lukashin, A. and Borodovsky, M. 1998, GeneMark.hmm: new solutions for gene finding, *Nucl. Acids Res.*, **26**, 1107–1115.
25. Burge, C. and Karlin, S. 1997, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, **268**, 78–94.
26. 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.
27. Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunak, S. 1996, Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information, *Nucl. Acids Res.*, **24**, 3439–3452.
28. Brendel, V. and Kleffe, J. 1998, Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA, *Nucl. Acids Res.*, **26**, 4748–4757.
29. UniProt Consortium, 2007, The universal protein resource (UniProt), *Nucl. Acids Res.*, **35**, 193–197.
30. Huang, X., Adams, M. D., Zhou, H. and Kerlavage, A. R. 1997, A tool for analyzing and annotating genomic sequences, *Genomics*, **46**, 37–45.
31. Lee, Y., Tsai, J., Sunkara, S., et al. 2005, The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes, *Nucl. Acids Res.*, **33**(Database issue), D71–D74.
32. Needleman, S. B. and Wunsch, C. D. 1970, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, **48**, 443–453.
33. Kruskal, J. B. 1983, An overview of sequence comparison, In: Sankoff, D. and Kruskal, J. B. (eds.), *Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison*, Boston: Addison Wesley, pp. 1–44.
34. Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. 1999, Plant cis-acting regulatory DNA elements (PLACE) database, *Nucl. Acids Res.*, **27**, 297–300.
35. Arabidopsis Genome Initiative, 2000, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature*, **408**, 796–815.
36. International Rice Genome Sequencing Project, 2005, The map-based sequence of the rice genome, *Nature*, **436**, 793–800.
37. Tuskan, G. A., Difazio, S., Jansson, S., et al. 2006, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science*, **313**, 1596–1604.
38. The French–Italian Public Consortium for Grapevine Genome Characterization, 2007, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature*, **449**, 463–467.

39. Mulder, N. J., Apweiler, R., Attwood, T. K., et al. 2007, New developments in the InterPro database, *Nucl. Acids Res.*, **35**(Database issue), D224–D228.
40. The Gene Ontology Consortium, 2000, Gene ontology: tool for the unification of biology, *Nat. Genet.*, **25**, 25–29.
41. Nakai, K. and Horton, P. 1999, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, *Trends Biochem. Sci.*, **24**, 34–36.
42. Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. 2000, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, *J. Mol. Biol.*, **300**, 1005–1016.
43. Hirokawa, T., Boon-Chieng, S. and Mitaku, S. 1998, SOSUI: classification and secondary structure prediction system for membrane proteins, *Bioinformatics*, **14**, 378–379.
44. Ito, M., Miyamoto, J., Mori, Y., et al. 2000, Genome and chromosome dimensions of *Lotus japonicus*, *J. Plant Res.*, **113**, 435–442.
45. Pedrosa, A., Sandal, N., Stougaard, J., Schweizer, D. and Bachmair, A. 2002, Chromosomal map of the model legume *Lotus japonicus*, *Genetics*, **161**, 1661–1672.
46. Fawcett, J. A., Kawahara, T., Watanabe, H. and Yasui, Y. 2006, A SINE family widely distributed in the plant kingdom and its evolutionary history, *Plant Mol. Biol.*, **61**, 505–514.
47. Yamada, K., Lim, J., Dale, J. M., et al. 2003, Empirical analysis of transcriptional activity in the *Arabidopsis* genome, *Science*, **302**, 842–846.
48. Yelin, R., Dahary, D., Sorek, R., et al. 2003, Widespread occurrence of antisense transcription in the human genome, *Nat. Biotechnol.*, **21**, 379–386.
49. Asamizu, E., Nakamura, Y., Sato, S. and Tabata, S. 2005, Comparison of the transcript profiles from the root and the nodulating root of the model legume *Lotus japonicus* by serial analysis of gene expression, *Mol. Plant Microbe Interact.*, **18**, 487–498.
50. Axtell, M. J. and Bartel, D. P. 2005, Antiquity of microRNAs and their targets in land plants, *Plant Cell*, **17**, 1658–1673.
51. Guilfoyle, T. J., Ulmasov, T. and Hagen, G. 1998, The ARF family of transcription factors and their role in plant hormone-responsive transcription, *Cell Mol. Life Sci.*, **54**, 619–627.
52. Bouche, N., Scharlat, A., Snedden, W., Bouchez, D. and Fromm, H. 2002, A novel family of calmodulin-binding transcription activators in multicellular organisms, *J. Biol. Chem.*, **277**, 21851–21861.
53. Cardon, G., Höhmann, S., Klein, J., Nettesheim, K., Saedler, H. and Huijser, P. 1999, Molecular characterisation of the *Arabidopsis* SBP-box genes, *Gene*, **237**, 91–104.
54. Hwang, I., Chen, H. C. and Sheen, J. 2002, Two-component signal transduction pathways in *Arabidopsis*, *Plant Physiol.*, **129**, 500–515.
55. Murray, J. D., Karas, B. J., Sato, S., Tabata, S., Amyot, L. and Szczyglowski, K. 2007, A cytokinin perception mutant colonized by Rhizobium in the absence of nodule organogenesis, *Science*, **315**, 101–104.
56. Tirichine, L., Sandal, N., Madsen, L. H., et al. 2007, A gain-of-function mutation in a cytokinin receptor triggers spontaneous root nodule organogenesis, *Science*, **315**, 104–107.
57. Nukui, N., Ezura, H. and Minamisawa, K. 2004, Transgenic *Lotus japonicus* with an ethylene receptor gene Cm-ERS1/H70A enhances formation of infection threads and nodule primordia, *Plant Cell Physiol.*, **45**, 427–435.
58. Mizuno, T. and Nakamichi, N. 2005, Pseudo-response regulators (PRRs) or true oscillator components (TOCs), *Plant Cell Physiol.*, **46**, 677–685.
59. Shiu, S. H., Karlowski, W. M., Pan, R., Tzeng, Y. H., Mayer, K. F. and Li, W. H. 2004, Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice, *Plant Cell*, **16**,1220–1234.
60. Stracke, S., Kistner, C., Yoshida, S., et al. 2002, A plant receptor-like kinase required for both bacterial and fungal symbiosis, *Nature*, **417**, 959–962.
61. Krusell, L., Madsen, L. H., Sato, S., et al. 2002, Shoot control of root development and nodulation is mediated by a receptor-like kinase, *Nature*, **420**, 422–426.
62. Nishimura, R., Hayashi, M., Wu, G. J., et al. 2002, HAR1 mediates systemic regulation of symbiotic organ development, *Nature*, **420**, 426–429.
63. Radutoiu, S., Madsen, L. H., Madsen, E. B., et al. 2003, Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases, *Nature*, **425**, 585–592.
64. Zhang, X. C., Wu, X., Findley, S., et al. 2007, Molecular evolution of lysin motif-type receptor-like kinases in plants, *Plant Physiol.*, **144**, 623–636.
65. Sugiyama, A., Shitan, N., Sato, S., Nakamura, Y., Tabata, S. and Yazaki, K. 2006, Genome-wide analysis of ATP-binding cassette (ABC) proteins in a model legume plant, *Lotus japonicus*: comparison with *Arabidopsis* ABC protein family, *DNA Res.*, **13**, 205–228.
66. Guenther, J. F. and Roberts, D. M. 2000, Water-selective and multifunctional aquaporins from *Lotus japonicus* nodules, (Abstract), *Planta*, **210**, 741–748.
67. Wienkoop, S. and Saalbach, G. 2003, Proteome analysis. Novel proteins identified at the peribacteroid membrane from *Lotus japonicus* root nodules, *Plant Physiol.*, **131**, 1080–1090.
68. Nelson, D. R., Schuler, M. A., Paquette, S. M., Werck-Reichhart, D. and Bak, S. 2004, Comparative genomics of rice and *Arabidopsis*. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot, *Plant Physiol.*, **135**, 756–772.
69. Shimada, N., Sato, S., Akashi, T., et al. 2007, Genome-wide analyses of the structural gene families involved in the legume-specific 5-deoxyisoflavonoid biosynthesis of *Lotus japonicus*, *DNA Res.*, **14**, 25–36.
70. Cannon, S. B., Sterck, L., Rombauts, S., et al. 2006, Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes, *Proc. Natl. Acad. Sci. USA*, **103**, 14959–14964.
71. Stracke, S., Sato, S., Sandal, N., et al. 2004, Exploitation of colinear relationships between the genomes of *Lotus japonicus*, *Pisum sativum* and *Arabidopsis thaliana*, for positional cloning of a legume symbiosis gene, *Theor. Appl. Genet.*, **108**, 442–449.