Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes

Steven B. Cannon[a,b], Lieven Sterck[c], Stephane Rombauts[c], Shusei Sato[d], Foo Cheung[e], Jérôme Gouzy[f], Xiaohong Wang[a], Joann Mudge[a], Jayprakash Vasdewani[a], Thomas Schiex[g], Manuel Spannagl[h], Erin Monaghan[e], Christine Nicholson[i], Sean J. Humphray[i], Heiko Schoof[j], Klaus F. X. Mayer[h], Jane Rogers[i], Francis Quétier[k], Giles E. Oldroyd[l], Frédéric Debellé[f], Douglas R. Cook[m], Ernest F. Retzel[n], Bruce A. Roe[o], Christopher D. Town[e], Satoshi Tabata[d], Yves Van de Peer[c], and Nevin D. Young[a,p]

[a]Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108; [b]U.S. Department of Agriculture–Agricultural Research Service and Department of Agronomy, Iowa State University, Ames, IA 50010; [c]Department of Plant Systems Biology (VIB), Ghent University, B-9052 Ghent, Belgium; [d]Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan; [e]Institute for Genomic Research, Rockville, MD 20850; [f]Laboratoire des Interactions Plantes–Microorganismes, Institut National de la Recherche Agronomique–Centre National de la Recherche Scientifique, 31326 Castanet-Tolosan, France; [g]Unité de Biométrie et Intelligence Artificielle, B.P. 52627, Institut National de la Recherche Agronomique, 31326 Castanet-Tolosan, France; [h]Munich Information Center for Protein Sequences Institute for Bioinformatics, Gesellschaft für Strahlung und Umweltforschung, Research Center for Environment and Health, 85764 Neuherberg, Germany; [i]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; [j]Max Planck Institute for Plant Breeding Research, 50829 Köln, Germany; [k]Centre Nationale de Séquençage, 91057 Evry, France; [l]John Innes Centre, Norwich NR4 7UH, United Kingdom; [m]Department of Plant Pathology, University of California, One Shields Avenue, Davis, CA 95616; [n]Center for Computational Genomics and Bioinformatics, Minneapolis, MN 55455; and [o]Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK 73019

Edited by Steven D. Tanksley, Cornell University, Ithaca, NY, and approved August 4, 2006 (received for review April 20, 2006)

Genome sequencing of the model legumes, *Medicago truncatula* and *Lotus japonicus*, provides an opportunity for large-scale sequence-based comparison of two genomes in the same plant family. Here we report synteny comparisons between these species, including details about chromosome relationships, large-scale synteny blocks, microsynteny within blocks, and genome regions lacking clear correspondence. The *Lotus* and *Medicago* genomes share a minimum of 10 large-scale synteny blocks, each with substantial collinearity and frequently extending the length of whole chromosome arms. The proportion of genes syntenic and collinear within each synteny block is relatively homogeneous. *Medicago–Lotus* comparisons also indicate similar and largely homogeneous gene densities, although gene-containing regions in *Mt* occupy 20–30% more space than *Lj* counterparts, primarily because of larger numbers of *Mt* retrotransposons. Because the interpretation of genome comparisons is complicated by large-scale genome duplications, we describe synteny, synonymous substitutions and phylogenetic analyses to identify and date a probable whole-genome duplication event. There is no direct evidence for any recent large-scale genome duplication in either *Medicago* or *Lotus* but instead a duplication predating speciation. Phylogenetic comparisons place this duplication within the Rosid I clade, clearly after the split between legumes and Salicaceae (poplar).

polyploidy | synteny | genome duplication

Legumes (Fabaceae), the third-largest family of flowering plants, are vitally important to agriculture and the environment. Because of their capacity for symbiotic nitrogen fixation, legumes provide a substantial fraction of all nutritional protein and reduce the need for agricultural chemicals. Legumes are an old and diverse plant family comprising nearly 18,000 species in essentially all terrestrial habitats (1). Legumes have been the subject of numerous studies aimed at understanding their genome organization and evolution (reviewed in refs. 2 and 3). Of particular interest have been questions of synteny and large-scale duplications, although our knowledge in these areas remains surprisingly limited. In contrast to the Gramineae, where fine-scale collinearity among rice, wheat, barley, and corn has been examined in detail (4, 5), details about macro- and microsynteny in legumes are more fragmentary. Previous studies have demonstrated broad-scale conservation of legume genes and gene order (6, 7), and a few studies have examined specific cases of microsynteny (8–11). Choi *et al.* (7), for example, used gene-based markers to create an integrated map and infer genome-wide synteny among five species in the Papilionoideae subfamily. Other studies have come to similar conclusions, focusing on different sets of species (6, 12–14). Consequently, it is now clear that synteny, often disrupted, is widespread among cultivated legumes, with some regions exhibiting very high levels of microsynteny. Nonetheless, comparisons of sufficient scale and resolution to reveal the details and extent of legume genome conservation are still required.

At the same time, it is important to learn more about the role of polyploidy and whole (large-scale)-genome duplication (WGD) in shaping legume genomes. A profound realization from plant genome sequencing has been the high frequency of large-scale duplications in plants (15, 16). Evidence for one or two WGD is found in the rice genome (16), whereas two or three rounds of WGD are apparent in both poplar and *Arabidopsis*, with the most recent events occurring independently (17, 18). In legumes, the timing of hypothesized WGD events remains in dispute. Using substitutions per synonymous site (Ks) analysis of EST data, Blanc and Wolfe proposed polyploidy near the time of separation between *Medicago truncatula* (*Mt*) and soybean, possibly after their split, concluding that "a complex set of events occurred in the legume lineage at around the time of the soybean/*Medicago* divergence" (17). On the basis of phylogenetic and Ks analysis in 39 gene families, Pfeil *et al.* (19) proposed a more ancient round of polyploidy that probably occurred in the common ancestor of *Medicago* and soybean, although Ks peaks were diffuse and the timing of the duplication event uncertain.

Recently, two legume models were chosen for large-scale genome sequencing, *Mt* and *Lotus japonicus* (*Lj*) (20). These

Author contributions: S.B.C., L.S., and S.R. contributed equally to this work; N.D.Y. designed research; S.S., J.G., T.S., M.S., C.N., S.J.H., H.S., K.F.X.M., J.R., F.Q., G.E.O., F.D., D.R.C., B.A.R., C.D.T., and S.T. contributed new reagents/analytic tools; S.B.C., L.S., S.R., F.C., X.W., J.M., J.V., E.M., and Y.V.d.P. analyzed data; and S.B.C., Y.V.d.P., and N.D.Y. wrote the paper.

The authors declare no conflict of interest.

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviations: *Mt*, *Medicago truncatula*; *Lj*, *Lotus japonicus*; WGD, whole-genome duplication; TC, tentative consensus; TAC, transformation-competent artificial chromosome; Ks, substitutions per synonymous site.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (GenBank accession nos. available at www.medicago.org/genome and www.kazusa.or.jp/lotus).

[p]To whom correspondence should be addressed. E-mail: neviny@umn.edu.

PLANT BIOLOGY

species were selected because they were already useful models for studying nitrogen fixation and symbiosis, and because both were found to have relatively small genomes with genes concentrated in euchromatic chromosomes arms. *Mt* and *Lj* had a common ancestor ≈40 Mya (21), near the time of radiation of most agriculturally important tribes in the Papilionoideae. Here we combine large-scale sequence comparisons between and within *Mt* and *Lj* together with Ks and phylogenetic analyses to describe syntenic relationships and duplication histories. In the process, we lay the foundation for reconstructing the ancestral genome of the *Mt*/*Lj* progenitor, a goal that will eventually establish a robust model for legume genomics and node for studies of angiosperm evolution.

Results and Discussion

Genome Sequence Coverage. At the time of the analysis, nonredundant map-anchored genome sequence in the *Mt* and *Lj* builds was 149 and 121 Mbp, respectively. An additional 15 and 27 Mbp had not yet been anchored to genetic maps and were not used for synteny analysis. Genome sizes for both *Mt* and *Lj* have been estimated at ≈475 Mbp (www.rbgkew.org.uk/cval). Extensive FISH data indicate both genomes are organized into gene-rich euchromatic chromosome arms and distinct gene-poor centromeric/pericentromeric regions (22–25). Thus, both projects have targeted the euchromatin by walking from gene-rich seed BACs or, in the case of *Lj*, transformation-competent artificial chromosomes (TACs) (26).

Coverage of gene-rich euchromatin was estimated by representation in two different datasets, EST sequences and low-copy genes. Both methods point to ≈50% coverage of euchromatin in the datasets analyzed with ≈40% of each genome both sequenced and anchored. Based on BLASTP and BLASTX comparisons of 995 *Arabidopsis* conserved single-copy genes against the available sequences from *Mt* and *Lj*, the proportions of recovered hits under stringent criteria (see *Methods*) were 48.5% for *Lj* and 51.3% for *Mt*. These results were largely consistent with the proportion of ESTs showing matches to genomic DNA. In *Mt*, for example, where greater numbers of EST sequence data are available than in *Lj*, 55% of high-quality tentative consensuses (TCs; see *Methods*) matched genome sequence at the time of the analysis. After adjusting for BACs and TACs not anchored to the genetic map, the proportion of gene-rich euchromatin both sequenced and anchored was ≈46% for *Mt* and ≈37% for *Lj*.

Gene Prediction and Relative Gene Densities. Gene prediction using a version of EuGene (27) specifically trained for *Medicago* (International *Medicago* Genome Annotation Group) returned 18,844 gene models (plus 1,826 on unanchored sequences) for *Mt* and 20,800 (plus 5,088 on unanchored sequences) genes for *Lj* after masking for transposons and other repeats. We expect the number of genes in *Lj* to be slightly overestimated, because the program had not been trained specifically for this genome and sometimes tended to split genes in the *Lj* sequence. The average gene density in the assemblies was 12.6 genes per 100 kb for *Mt* (SD = 2.9, n = 1,644 bins of 100 kb) and 17.4 genes per 100 kb for *Lj* (SD = 3.5, n = 1,490). Thus, the *Lj*/*Mt* ratio of gene densities was 1.38, with the small standard deviations indicating relatively homogeneous density distributions in the portions of the genomes sequenced so far.

This estimate of the *Lj*/*Mt* gene density ratio could be sensitive to differences in gene-calling accuracy between the two genomes, so we also carried out a complementary approach in which the lengths of *Mt* and *Lj* segments within syntenic blocks (see below) were compared and the ratio used to estimate relative gene densities. This approach is less sensitive to the accuracy of individual gene calls. Distances between synteny block endpoints in *Mt* and *Lj* have an R^2 of 0.71 with a d*Mt*/d*Lj* slope of 1.20. Based on these results, we conclude that gene-containing regions in *Mt* occupy 20–30% more space than their *Lj* counterparts. Most of this difference can be explained by additional repetitive sequences, primarily retrotransposons, in *Mt* based on the observation that the proportion of masked (repetitive) sequence in *Mt* was 38% compared with just 19% in *Lj*. Other explanations, including additional types of repeat elements, introns, and tandem duplications, could also contribute to the difference but were not examined in detail.

Synteny Between *Medicago* and *Lotus*. To detect syntenic regions in *Mt* and *Lj*, we used DiagHunter (28) and i-ADHoRe (29). Predictions by the two programs were similar, with i-ADHoRe predicting higher numbers of small- and lower-scoring synteny blocks (i.e., greater sensitivity but also greater likelihood of false positives). Because we were primarily interested in relatively recent synteny (≈40 Mya) and genome duplication remnants (<100 Mya), we report results from DiagHunter unless otherwise indicated. Full results from both programs are available in supporting information, which is published on the PNAS web site. The DiagHunter and i-ADHoRe programs, respectively, detected 32–43% of sequenced and anchored *Lj* sequence within synteny blocks in *Mt* and 26–40% of sequenced and anchored *Mt* sequence within synteny blocks in *Lj*. In both cases, multiple blocks overlapping the same region were counted. Of course, just 37% of *Lj* and 46% of *Mt* euchromatin were represented in these comparisons, so the maximum amount of synteny expected based on random coverage (and assuming no lineage-specific genome duplication) would be 46% of the *Lj* sequence (vs. 32–43% synteny observed) and 37% of the *Mt* sequence (vs. 26–40% observed). Consequently, these results can be interpreted to indicate essentially complete synteny coverage between the genomes, or, more likely, the presence of secondary blocks resulting from genome duplication (see below).

Synteny results were examined at three different levels of organization, whole genome, large-scale synteny blocks, and microsynteny <1 Mbp (Fig. 1). For many *Mt*/*Lj* chromosome comparisons, the level of synteny is striking (Fig. 1 *A* and *C*). Although both genome builds were incomplete with some misordered contigs at the time of analysis, *Mt*1 and *Lj*5 show synteny along their entire lengths, and *Lj*1 is almost entirely syntenic with *Mt*3 and *Mt*7. Correspondences prominent in Fig. 1 *A* and *C* and supporting information are *Mt*1-*Lj*5, *Mt*2-*Lj*4/*Lj*6, *Mt*3/*Lj*1, *Mt*4/8-*Lj*3/4, and *Mt*5/6-*Lj*2, with secondary duplications probably corresponding to older duplication remnants. These results substantially extend observations of *Mt*/*Lj* synteny in earlier publications, where many of the same relationships had been observed (7, 20).

Chromosome-sized relationships between *Mt* and *Lj* correspond for the most part to 10 large-scale synteny blocks that generally extend the length of whole chromosome arms (Fig. 1*A*). Genome wide, these large-scale synteny blocks account for ≈67% of the *Mt* genome and ≈64% of *Lj*. Within individual synteny blocks, conservation of gene content and order (synteny quality) is substantial. In the genome region illustrated in Fig. 1*B*, for example, 61% (58/94) of genes exhibit corresponding homologs within these regions. Averaged across all predicted orthologous blocks, synteny quality is 54% ± 14%, where "synteny quality" is calculated as twice the number of matches divided by the total number of genes in both segments after excluding transposable elements and collapsing tandem duplications. Values reported here are probably underestimates and are, in fact, lower than those reported in ref. 7. At the time of our analysis, a significant fraction of BAC and TAC clone sequences were still unfinished, with some errors in order and orientation. Our analysis was also based on preliminary and automated gene calls. Together, these sources of error probably obscured some real cases of synteny.

Also noteworthy is the absence of synteny in some genome

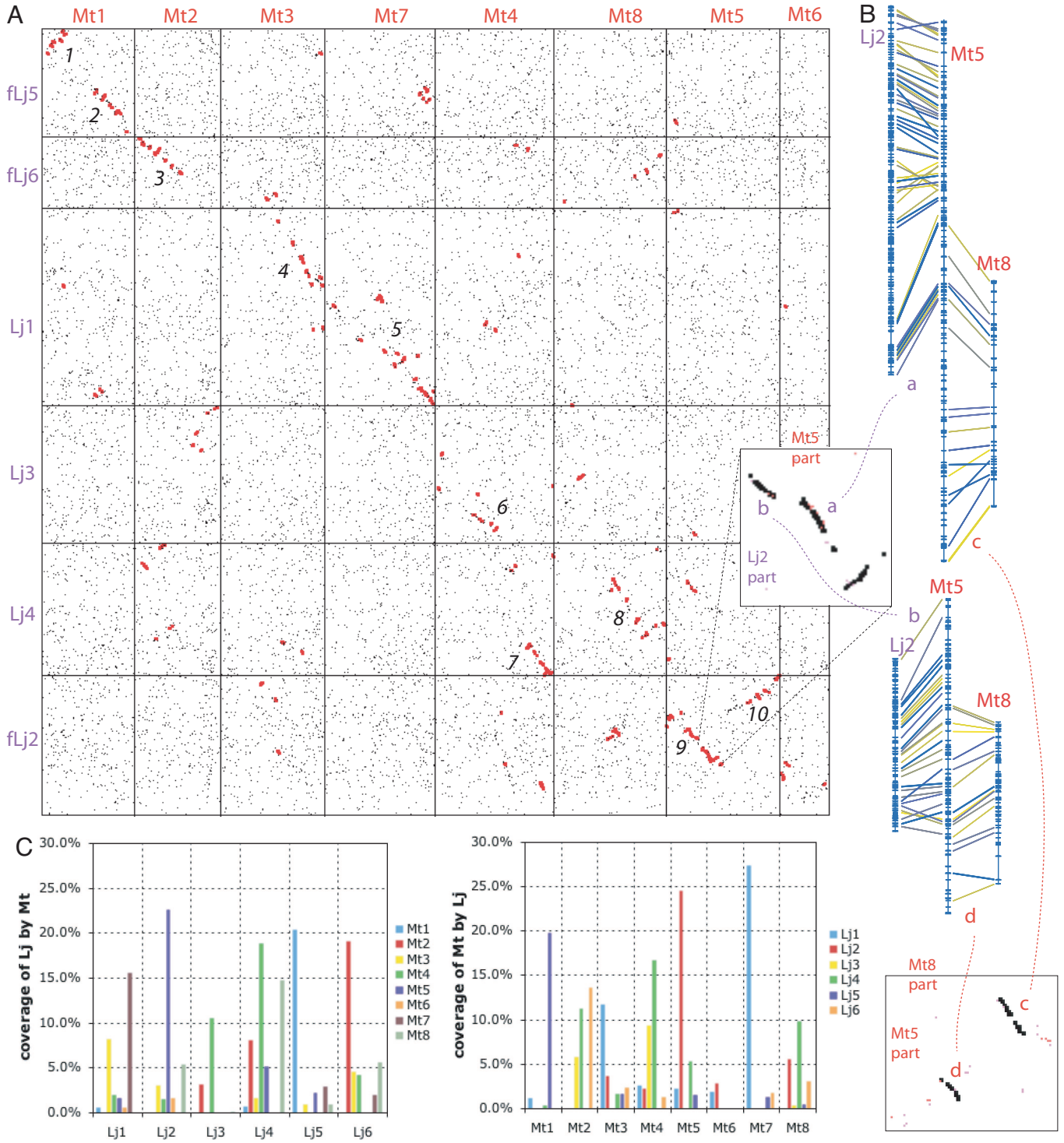

Fig. 1. Dot plots, synteny closeup, and chromosome correspondences. (*A*) In the dot plot, each dot represents the reciprocal best BLASTP match between gene pairs. Red dots show regions of synteny as identified by DiagHunter. Some *Lj* chromosome orientations have been flipped (''fLj5, fLj6, fLj2'') to visually correspond to *Mt* orientations. Both *Mt* and *Lj* have been scaled to occupy the same lengths. See supporting information for all dot plots and related results. (*B*) Closeup views of synteny. Lines in shades of blue or yellow indicate BLASTP E-values, with strongest correspondence blue (0.0) to weakest yellow (0.01). Only single strongest reciprocal hits are shown. Comparisons a and b (upper callout) show synteny in *Lj*2 × *Mt*5; comparisons c and d (lower callout) show synteny in *Mt*5 × *Mt*8 (see text). (*C*) Graphs showing percentages of individual pseudochromosomes with synteny between *Mt* and *Lj*. Coverage is calculated as the sum of block sizes (a block has dimensions of the end gene minus the start gene). Multiple synteny blocks over the same region are counted, so altogether coverage can exceed 100%.

regions, including much of *Mt*6 and smaller sections of *Mt*3, *Mt*8, *Lj*3, *Lj*4, and *Lj*6. Coverage of *Mt*6 by the entire *Lj* genome is just 4.8% (primarily 1.9% coverage by *Lj*1 and 2.8% by *Lj*2) compared with 27.4% coverage of *Mt*7 by *Lj*1. Chromosome *Mt*6 is unusual in being much shorter and transposon-dense than any other *Mt* chromosome (23, 24). It is also home to novel repeats not found

elsewhere in the *Mt* genome (data not shown) as well as a significant fraction of all *Mt* NBS-LRR genes (30). Further genome sequencing will help to resolve the nature of this missing synteny, including the opportunity to determine whether scattered genome regions correspond to regions lacking large-scale synteny and, if so, the genome distribution of such scattered regions.

Duplication History of *Medicago* and *Lotus*. ***Evidence from synteny analysis.*** There is substantially less genome sequence within internal duplications in either *Mt* or *Lj*, measured as within-genome synteny blocks, than in synteny blocks between the two genomes. Using the same parameters as in the *Mt*/*Lj* comparisons above, DiagHunter detected just 9.7% and 6.8% of the *Mt* and *Lj* genomes, respectively, as internally duplicated (Fig. 1*A* and supporting information). Because 46% of the *Mt* sequence and 37% of *Lj* were available at the time of the analysis (representing the upper limits for the amount of internal duplication that could be observed), these levels of internal synteny are just one-fifth that expected from a single recent WGD. Observed internal synteny within both *Mt* and *Lj* is also three to four times lower than the intergenomic comparison between *Mt* and *Lj*. The smaller amount of internal synteny presumably reflects an ancient WGD followed by gene loss and rearrangement. This ancient WGD helps to explain the high level of synteny observed in the *Mt*/*Lj* comparison relative to the expectation of no prior WGD.

The nature of synteny between *Mt* and *Lj* provides further evidence that a WGD predated speciation. Syntenic blocks tend to be extensive between *Mt* and *Lj* but more degraded internally in either *Mt* or *Lj*. In the illustration of microsynteny in Fig. 1*B*, the *Lj*/*Mt* comparison exhibits 62% conservation, whereas the *Mt*/*Mt* region retains just 36% and the *Lj*/*Lj* region, 30% (supporting information). Overall, synteny quality is 31% in *Mt*/*Mt* duplications and 25% in *Lj*/*Lj* duplications, substantially lower than the 54% observed in *Mt*/*Lj* synteny blocks.

Assuming genome duplication preceded speciation, a dot plot between *Mt* and *Lj* should exhibit cases of paired synteny blocks, each corresponding to descendants of the ancient duplication event, and each exhibiting comparable levels of synteny between the two species. We see this 2-fold coverage in several regions of the *Mt*/*Lj* dot plot comparison (Fig. 1*A*). For example, the *Lj*2/*Mt*5 synteny in Fig. 1*B* is matched by *Mt*5/*Mt*8 synteny, a region that also contains *Lj*2/*Lj*4 synteny (supporting information). Overall, the proportion of *Lj*/*Mt* synteny blocks showing this pattern of overlaid duplications is 28.6% for *Mt* and 23.1% for *Lj*. These genome regions may provide useful material for studying the evolution of large duplicated regions in plants, including the question of whether duplicated regions have been relatively stable since the *Mt*/*Lj* separation or whether differential duplicated gene loss has been ongoing.

Evidence from synonymous substitution analysis. If homologous segments were generated by a single large-scale duplication event, they should all have been created at the same time. By plotting the number of duplication blocks against the average Ks value of the blocks (see *Methods*), we obtain a distribution reflecting the approximate age of the duplication (31). As seen in Fig. 2, the age distribution of *Mt* has a clear peak at 0.8 (median 0.80) synonymous substitutions per site, and *Lj* has a broader peak or set of peaks at 0.7–0.9 (median 0.73). Thus, it seems likely that a large-scale duplication event occurred around the time corresponding to Ks 0.7–0.9. On the same figure, the age distribution for synteny blocks between *Mt* and *Lj* is also plotted, showing a peak at a Ks of 0.6 (median 0.64), corresponding to speciation between *Mt* and *Lj*. Differences between the distributions are highly significant: $P = 2.3\text{E-}11$ for *Mt*/*Mt* vs. *Mt*/*Lj*, and $P = 1.1\text{E-}23$ for *Lj*/*Lj* vs. *Mt*/*Lj*. These results suggest a large-scale duplication preceding speciation, although other details remain uncertain, including the relative timing of older peaks in the range Ks ≈1.5–2.0, event(s) presumably shared with other angiosperms.

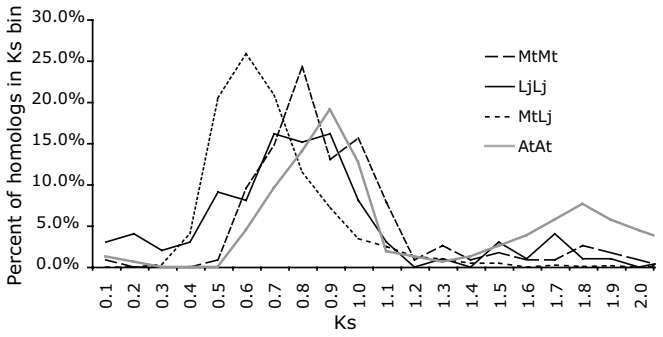

Fig. 2. Ks dating of duplication blocks and *Mt*/*Lj* synteny blocks. Age distributions of duplicated (*Mt*, *Lj*, and *Arabidopsis*) and collinear segments (*Mt*/*Lj*). The vertical axis indicates percent of Ks values, and the horizontal axis denotes Ks (one bin corresponds with a Ks value of 0.1). Ks values are averages of three adjacent homologs within a collinear segment, as described in *Methods*. The *Arabidopsis* distribution is taken from Simillion *et al.* (49).

Evidence from phylogenetic analysis. To further establish the timing of WGD relative to speciation, we carried out a high-throughput genome-wide phylogenetic analysis of duplicated genes in *Arabidopsis*, poplar, *Mt*, and *Lj* (32). Trees were constructed to determine whether a majority of gene duplications in the two legumes occurred separately in each lineage (after speciation) or shared between both species (before speciation).

Among 413 informative trees, several nontandem terminal clade duplications were apparent in *Mt* and *Lj* (65 and 30, respectively). These numbers must be interpreted with caution. If the phylogenetic pattern (M,M) represents paralogous *Mt* genes, (L,L) paralogous *Lj* genes, and (L,M) orthologous *Lj* and *Mt* genes, then in the absence of gene loss, greater numbers of orthologous duplications (L,M) compared with paralogous nontandem duplications (M,M and L,L) suggest WGD before speciation. We call this the "WGD-early" model. Greater numbers of paralogous duplications suggest WGD after speciation, the "WGD-late" model. Because gene loss and nonrecovery of existing genes can lead to multiple outcomes, only a complete pattern provides clear evidence. Less clear-cut interpretations can be made from patterns with a single gene loss.

Among 413 trees, there were 11 exhibiting completely informative patterns for *Mt* and *Lj* plus 116 with a single gene loss. Nine completely informative trees supported WGD-early, whereas just two supported WGD-late. The numbers were 81 WGD-early vs. 35 WGD-late among partially informative trees (Table 1).

These phylogenetic trees can be further analyzed to ask whether WGD occurred before or after the separation between Rosid 1 and the Salicaceae (poplar). In this analysis, we required at least three legume sequences in each legume clade and at least one poplar and one *Arabidopsis* sequence in correct phylogenetic

Table 1. Tree patterns and timing of WGD

	Complete		Partial	
WGD timing	Without bootstrap	With bootstrap	Without bootstrap	With bootstrap
WGD-early	9	5	81	65
WGD-late	2	1	35	30

Values reflect counts of tree patterns that support different timings of WGD. Complete, trees that exhibit patterns of no gene loss; partial, trees with only a single gene loss. Trees exhibiting the pattern [(M,L),(M,L)] are counted as WGD-early; those with the pattern [(M,M),(L,L)] are counted as WGD-late.

positions. Among 65 WGD-early trees with at least 70% bootstrap support in the diagnostic legume clade, 23 matched the "WGD within Rosid I" model, and none matched "WGD before Rosid I." Thus, it is virtually certain that a WGD occurred within Rosid I, after the split between poplar and legumes.

Conclusions

Comparisons between *Mt* and *Lj* genomic sequences, even while the sequencing projects are still underway, demonstrate extensive synteny plus the existence and timing of one or more large-scale genome duplications early in legume evolution. Genome comparisons also indicate relatively homogeneous gene densities throughout the euchromatin of both genomes, with the *Mt* gene space occupying 20–30% greater size than *Lj* due to larger numbers of *Mt* transposons.

Our results substantially extend previous observations of synteny between *Mt* and *Lj* (7, 20), defining more precisely the end points of large-scale synteny blocks and illustrating the fine-scale details of similarities within syntenic regions. Even accounting for incomplete sequence data and internal duplications, the scale of synteny at both the micro- and macroscale is impressive. As these sequencing projects move forward, the availability of two highly syntenic legume genomes with nearly complete sequence coverage of their euchromatin will enable reconstruction of ancestral chromosome sequences. The 10 large-scale synteny blocks described here provide the basis for this reconstruction process, although more complete genome sequence is still needed. In addition to their utility in legume genomics, these virtual chromosomes will provide a well-placed node for evolutionary comparisons with other angiosperms.

Shorter and more degraded synteny is observed in comparisons of each genome to itself, evident as secondary synteny blocks in comparisons between *Mt* and *Lj*. This pattern of strong extended synteny between *Mt* and *Lj* together with shorter, weaker synteny in self comparisons demonstrates that any WGD significantly predated the common ancestor of *Mt* and *Lj*. Moreover, Ks frequencies indicate a strong peak in *Mt*/*Lj* synteny blocks that is significantly more recent than diffuse Ks peaks in either *Mt*/*Mt* or *Lj*/*Lj* synteny blocks. Finally, a genome-wide phylogenetic analysis of duplicated genes uncovered five times as many trees with patterns supporting a WGD predating speciation compared with patterns supporting independent WGD events after speciation. Similar tests unambiguously place the WGD event within the Rosid I clade, after the separation of the Salicaceae and Fabaceae. Together, these results support earlier suggestions of an old legume WGD predating the ≈50 Mya *Mt*–soybean separation (10–12, 19, 33). Although the literature also contains Ks-based results suggesting this "old" WGD might have occurred relatively recently, and possibly independently in soybean and *Mt* (17), our results clearly support a more ancient duplication event.

Methods

Genome Sequencing Strategies. Both sequencing projects used a clone-by-clone approach (20), with the *Medicago* project sequencing from BACs and the *Lotus* project, from TACs (26). To anchor clone sequences, both projects used a combination of sequence-based genetic mapping, chromosome walking to extend sequence contigs, and fingerprint contig data (34). The *Lotus* project is also using whole-genome shotgun sequence data and low-density clone sequencing to extend genome coverage and fill gaps (35).

Chromosome Sequence Assembly. Construction of chromosome assemblies began with the creation of sequence contigs (sequence composed of more than one BAC sequence) based on overlapping sequence, then ordering on the basis of genetic markers and fingerprint contig data (34). For *Medicago*, BAC sequences were assembled into larger sequence contigs by first comparing all BAC sequences against one another using Mummer (36) and assembling overlapping BACs into contigs using the Paracel Genome Assembler (Paracel, Pasadena, CA). Similar procedures were used in *Lotus* but with overlap coordinates determined through manual evaluation of individual BAC overlaps and supporting information including PCR verification, marker location, and paired BAC end matches. For both genomes, locations of BAC singletons and sequence contigs were determined primarily by using genetic marker locations, with fingerprint contigs and paired BAC ends providing additional information about local BAC and contig orderings.

Repeat Identification, Masking, and Gene Calling. Pseudochromosome sequences were masked by using RepeatMasker (www.repeatmasker.org) by using a dataset that combined Repbase (37) and a specialized database of *Medicago* and *Lotus* transposable elements. Sequences in the database of *Medicago* and *Lotus* transposable elements were identified by iteratively finding PFAM (38) hits to transposon-related domains on predicted proteins. Afterward, a region of 10 kb upstream and 10 kb downstream around the selected loci was extracted followed by Reputer (39) to look for LTR or terminal inverted repeats within. Once potential borders were defined, putative transposons were checked with BLAST (40) against Uniprot (41) to ensure there were no better hits to nontransposon sequences. Gene models were predicted by using EuGene Ver. 3.2 (27) with the parameters trained for *Medicago* by the International Medicago Genome Annotation Group (www.medicago.org/genome/IMGAG).

Estimates of Genome Sequencing Coverage. We estimated sequence coverage of the euchromatic regions in two ways: by determining proportions of ESTs with strong matches and by calculating proportions of "low-copy conserved genes" with strong matches. For EST comparisons, we used only EST contigs with at least five ESTs in a contig (increasing the likelihood of high-quality query sequences). These TCs (42) were considered to have a genomic match if at least 90% of the TC matched at least 95% identity in a BLAT search (43). For comparison with "low-copy, conserved genes," we began with a list of 995 genes that are single copy in *Arabidopsis*, poplar, and rice. These were considered to have a genomic match in *Medicago* or *Lotus* if at least 50% of the gene matched in a TBLASTN search with at least 50% positive residues in the alignable region. Stringent criteria were defined as >50% of alignable region identical or similar amino acid content and E-value < E-10, whereas lenient criteria required E-value < E-10 only.

Genome Comparisons and Synteny Identification. Chromosome-scale synteny comparisons were made with two methods, DiagHunter (28) and i-ADHoRe (29). Both methods identify runs of collinear predicted proteins between genomic regions. Protein sets were identified as described above. Parameters for DiagHunter were as follows; only top reciprocal BLASTP matches per chromosome pair were considered at E-values less than E-20. The hit-matrix "compression factor" was 2,500. Gene orientation was considered, and four genes with the same respective orientations in both genomes were required to establish a synteny block. Insertions, deletions, and inversions were accommodated as described in Cannon *et al.* (28). Parameters for i-ADHoRe were as follows. Homologous relations between genes, which serve as input for the i-ADHoRe algorithm, were determined by implementing the Li-Rost criterion (44) on all-vs.-all BLASTP result of predicted proteins. The following parameters were used in the i-ADHoRe analysis: gap size of 60 genes, Q value of 0.9, a minimum of four homologs to define a block, and the higher-level multiplicon detection disabled (level

2 only). For Ks analysis, synteny blocks were further required to have at least four homologs in each genome, with the condition that each homolog fell within 1–10 kb of another in the block. Synteny comparisons on a scale of hundreds (rather than millions) of kb (e.g., Fig. 1*B*) were also made by using unpublished PERL scripts that visualize protein homologies based on identifications with BLASTP. Large-scale synteny blocks were inferred by visual inspection of dot-plot blocks on proximity and collinearity of runs identified by DiagHunter and i-ADHoRe.

Ks Analysis of Homologous Segments. The "age" of duplication or divergence of homologous segments was estimated by computing the number of synonymous substitutions per synonymous site (Ks) between homologous genes. To determine the Ks value of the pairs we used CODEML (45) from the PAML package (46). Because the program can become trapped in local optima, we ran the program five times for each gene pair and took the Ks estimation with the highest likelihood. Initially, Ks values were computed for individual homologs within a homologous segment. Outliers (strongly deviating Ks values) were eliminated by Grubbs outlier detection. Before determining Ks distributions, we calculated the mean Ks value for adjacent Ks triplets within the segment. Ks distributions were based on these locally averaged Ks values.

Phylogenetic Analysis. On the basis of an all-versus-all BLASTP search of all *Arabidopsis*, poplar, rice, *Lj*, and *Mt* proteins, we identified pairs of homologous genes using the Li-Rost criterion (44). These pairs were clustered into gene families by using a single linkage clustering method. A filtering step was performed based on the following criteria: all five species should be represented in the family with at least two protein members in both *Mt* or *Lj* and a maximum threshold of nine members for any one species. This resulted in a subset of the initial gene families for use in phylogenetic analysis. A multiple sequence alignment was constructed with CLUSTALW for each gene family (47), after which the alignments were stripped by removal of noninformative and gap-containing columns from the alignment. Neighbor-joining trees (100 bootstraps) were constructed with PHYLIP (48) and the resulting trees analyzed with a custom-made PERL script that evaluated duplication events in each tree and determined the relative dating for observed duplication events. For this final step in the analysis, only nodes supported with a bootstrap value higher than 70 were considered.

We thank the many participants in the international *Medicago* sequencing consortium (www.medicago.org/genome/people.php). Collaborators meriting special note include R. Denny, B. Chacko, E. Cannon, A. Baumgarten, W. Odland, and R. Geurts. Research at University of Minnesota, University of Oklahoma, The Institute for Genomic Research, and University of California at Davis came from the National Science Foundation (Projects 01-10206 and 03-21460). Additional support at University of Oklahoma came from the Samual Roberts Noble Foundation. Research at Ghent University, Genoscope, Wellcome Trust Sanger Institute Centre, John Innes Centre, Institut National de la Recherche Agronomique–Toulouse, and Munich Information Center for Protein Sequence came from European Community Project FOOD-CT-2004-506223-GRAIN LEGUMES. Additional support at Genoscope came from le Ministère de la Recherche, France, and at Sanger Centre and John Innes Centre from the Biotechnology and Biological Sciences Research Council. Research support for the Kasuza Center came from the Kasuza DNA Research Institute Foundation.

1. Doyle JJ, Luckow MA (2003) *Plant Physiol* 131:900–910.
2. Young ND, Mudge J, Ellis TH (2003) *Curr Opin Plant Biol* 6:199–204.
3. Zhu H, Choi HK, Cook DR, Shoemaker RC (2005) *Plant Physiol* 137:1189–1196.
4. Bowers JE, Chapman BA, Rong J, Paterson AH (2003) *Nature* 422:433–438.
5. Rice Chromosome 3 Sequencing Consortium (2005) *Genome Res* 15:1284–1291.
6. Boutin SR, Young ND, Olson TC, Yu ZH, Shoemaker RC, Vallejos CE (1995) *Genome* 38:928–937.
7. Choi HK, Mun JH, Kim DJ, Zhu H, Baek JM, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB, *et al.* (2004) *Proc Natl Acad Sci USA* 101:15289–15294.
8. Cannon SB, McCombie WR, Sato S, Tabata S, Denny R, Palmer L, Katari M, Young ND, Stacey G (2003) *Mol Genet Genomics* 270:347–361.
9. Gualtieri G, Kulikova O, Limpens E, Kim DJ, Cook DR, Bisselin T, Geurts R (2002) *Plant Mol Biol* 50:225–235.
10. Mudge J, Cannon SB, Kalo P, Oldroyd GE, Roe BA, Town CD, Young ND (2005) *BMC Plant Biol* 5:15.
11. Yan HH, Mudge J, Kim DJ, Shoemaker RC, Cook DR, Young ND (2004) *Genome* 47:141–155.
12. Lee JM, Bush A, Specht JE, Shoemaker RC (1999) *Genome* 42:829–836.
13. Menancio-Hautea D, Fatokun CA, Kumar L, Danesh D, Young ND (1993) *Theor Appl Genet* 86:797–810.
14. Simon CJ, Muehlbauer FJ (1997) *J Hered* 88:115–119.
15. *Arabidopsis* Genome Initiative (2000) *Nature* 408:796–815.
16. Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng, *et al.* (2005) *PLoS Biol* 3:e38.
17. Blanc G, Wolfe KH (2004) *Plant Cell* 16:1667–1678.
18. Sterck L, Rombauts S, Jansson S, Sterky F, Rouzé P, Van de Peer Y (2005) *New Phytol* 167:165–170.
19. Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ (2005) *Syst Biol* 54:441–454.
20. Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, Roe BA, Tabata S (2005) *Plant Physiol* 137:1174–1181.
21. Lavin M, Herendeen PS, Wojciechowski MF (2005) *Syst Biol* 54:575–594.
22. Hayashi M, Miyahara A, Sato S, Kato T, Yoshikawa M, Taketa M, Pedrosa A, Onda R, Imaizumi-Anraku H, Bachmair A, *et al.* (2001) *DNA Res* 8:301–310.
23. Kulikova O, Geurts R, Lamine M, Kim DJ, Cook DR, Leunissen J, de Jong H, Roe BA, Bisseling T (2004) *Chromosoma* 113:276–283.
24. Kulikova O, Gualtieri G, Geurts R, Kim DJ, Cook D, Huguet T, de Jong JH, Fransz PF, Bisseling T (2001) *Plant J* 27:49–58.
25. Pedrosa A, Sandal N, Stougaard J, Schweizer D, Bachmair A (2002) *Genetics* 161:1661–1672.
26. Liu Y-G, Shirano Y, Fukaki H, Yanai Y, Tasaka M, Tabata S, Shibata D (1999) *Proc Natl Acad Sci USA* 96:6535–6540.
27. Foissac S, Bardou P, Moisan A, Cros MJ, Schiex T (2003) *Nucleic Acids Res* 31:3742–3745.
28. Cannon SB, Kozik A, Chan B, Michelmore R, Young ND (2003) *Genome Biol* 4:R68.
29. Simillion C, Vandepoele K, Saeys Y, Van de Peer Y (2004) *Genome Res* 14:1095–1106.
30. Zhu H, Cannon SB, Young ND, Cook DR (2002) *Mol Plant–Microbe Interact* 15:529–539.
31. Van de Peer Y (2004) *Nat Rev Genet* 5:752–763.
32. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) *Proc Natl Acad Sci USA* 102:5454–5459.
33. Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) *Genome* 47:868–876.
34. Engler FW, Hatfield J, Nelson W, Soderlund CA (2003) *Genome Res* 13:2152–2163.
35. Sato S, Tabata S (2005) *Curr Opin Plant Biol* 9:128–132.
36. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) *Nucleic Acids Res* 30:2478–2483.
37. Jurka J, Kapitonov VV, Pavlicek A, Kolonowski P, Kohany O, Walichiewicz J (2005) *Cytogenet Genome Res* 110:462–467.
38. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL (2000) *Nucleic Acids Res* 28:263–266.
39. Kurtz S, Schleiermacher C (1999) *Bioinformatics* 15:426–427.
40. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) *Nucleic Acids Res* 25:3389–3402.
41. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al* (2005) *Nucleic Acids Res* 33:D154–D159.
42. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) *Nucleic Acids Res* 33:D71–D74.
43. Kent WJ (2002) *Genome Res* 12:656–664.
44. Li WH, Gu Z, Wang H, Nekrutenko A (2001) *Nature* 409:847–849.
45. Goldman N, Yang Z (1994) *Mol Biol Evol* 11:725–736.
46. Yang Z (1997) *Comput Appl Biosci* 5:555–556.
47. Thompson JD, Higgins DG, Gibson TJ (1994) *Nucleic Acids Res* 22:4673–4680.
48. Felsenstein J (1989) *Cladistics* 5:164–166.
49. Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y (2002) *Proc Natl Acad Sci USA* 99:13627–13632.