Chapter 3
Legume Comparative Genomics

Steven Cannon

Introduction

The ability to make comparisons between genome sequences will be crucial for leveraging and exchanging knowledge learned in these model systems, and applying that knowledge to a wide range of agronomically important species. Sequence comparisons are also a key tool for the evolutionary trajectories giving rise to new plant functions, structures, chemistries, and physiologies.

At the time of writing, four plant genome sequences are almost completely determined: *Arabidopsis thaliana* (At), two rice cultivars (*Oryza sativa*; Os), and *Populus trichocarpa* (Pt; black cottonwood or western balsam poplar). Genome sequencing is well underway for three legumes genomes: the model forage legumes *Medicago truncatula* (Mt) and *Lotus japonicus* (Lj), and *Glycine max* (Gm; soybean). Numerous other plant genome sequencing projects are also underway or planned, including tomato, corn, *Mimulus guttatus* (monkey flower), *Physcomitrella patens* (a moss), *Miscanthus* (switchgrass), *Citrus sinensis* (orange), Sorghum, cotton, cassava, *Brachypodium distachyon* (a model grass), and *Aquilegia formosa* (columbine).

The extent to which knowledge can be extrapolated between genomes depends in large part on this fundamental question: how do genomes change, and do they all change the same way and at roughly similar rates? This very broad question can be divided and made more specific: what are (1) the organization of genes and non-genes; (2) the mechanisms of large-scale genome change; (3) the pace of synteny loss?

Restating and elaborating these questions, (1) How similar are various genomes in organization of their component small parts: genes, regulatory regions, repetitive DNA, low-copy intergenic material, centromeric repeats, pericentromeric and telomeric sequences, etc? Do these elements behave essentially the same in all plant genomes? (2) Do all genomes change via similar mechanisms – acting in

S. Cannon
USDA-ARS, Department of Agronomy, Iowa State University, Ames, IA 50011, USA
e-mail: scannon@iastate.edu

G. Stacey (ed.), *Genetics and Genomics of Soybean*,

similar proportion? That is, do genomes in all lineages change primarily by the same mechanisms, such as polyploidy, breakages, fusions, inversions, translocations, and transposon insertions? (3) How similar are various genomes in degree of gene-order conservation (synteny) across taxa? That is, at what pace is synteny lost?

These questions are only partly orthogonal. Gene order might be generally retained between two genomes (3), with gene density differing greatly between genomes, or even in different parts of a single genome (1). The degree to which synteny is retained across various lineages (3) depends on the rates and mechanisms of genome change in these lineages (2) – though, conceivably, different mechanisms could produce similar changes in gene distribution of synteny.

This chapter will briefly review some of the key literature on these fundamental questions, with organization generally consisting of brief summaries of comparative genomic findings from other plant families, followed by a summary of results of comparisons among the legumes. As much of the early comparative genomic work was first carried out in the grasses, these will be a featured "comparison" for this paper on legume comparative genomics. The remainder of the Introduction briefly describes background information on legume sequencing projects and legume systematics.

One certainty in plant comparative genomics is that there will be exceptions and surprises as we examine more genomes. Nevertheless, a theme that emerges from comparisons so far is surprising commonality and similarity among plant genomes, even at great evolutionary distances.

Legume Genome Sequencing Strategies and Status

The international *Medicago truncatuala* (Mt) genome sequencing consortium, initiated by early funding from the Samuel Roberts Noble Foundation, and now funded by the National Science Foundation and the European Union, is scheduled to complete the euchromatic genome regions (16 chromosome arms) by the end of 2008. This project is using a clone-by-clone approach, in which bacterial artifical chromosomes (BACs), with average insert size of approximately 120 kb, are sequenced and used to extend BAC-contig tiling paths to produce increasingly large sequence contigs. Contigs are anchored and oriented using genetic markers developed from a large proportion of the BAC sequences. As of early 2007, BAC contigs and sequences covered approximately 60% of the major euchromatic regions of the Mt genome (Cannon et al. 2006 and unpublished data).

The *Lotus japonicus* (Lj) genome sequencing project is being carried out by the Kazusa DNA Research Institute in Japan. This project is also primarily using a clone-by-clone approach, sequencing transformation-competent artificial bacterial chromosomes (TACs), with average insert size of approximately 100 kb. The clone-by-clone sequence is also being augmented by a combination of whole genome shotgun (WGS) and low-coverage TAC sequencing. As of last published reports in 2006, Lj sequence coverage was also approximately 60% of the euchromatic regions of the Lj genome (Young et al. 2005; Cannon et al. 2006).

The *Glycine max* (Gm) genome is being sequenced primarily with a WGS approach, with sequence coming from a combination of random reads, paired fosmid ends, and paired BAC end sequences. Additionally, approximately 500 BACs will be sequenced to high coverage; these will be a mix of BACs selected for biological interest by members of the soybean research community, and to span gaps where necessary.

Legume Systematics and Consequences for Comparative Studies

The legume family is extremely diverse, with around 20,000 species and 700 genera, found in every terrestrial and some aquatic environments (Doyle and Luckow 2003). The majority of species are in the papilionoid subfamily, with 476 genera and about 14,000 species (Lewis et al. 2003). The Mimosoideae subfamily contains 77 genera and around 3,000 species. The remainder fall in the casalpinoideae subfamily – something of a grab-bag, of 162 genera and around 3,000 species, including diverse early-diverging legume taxa (Fig. 3.1).

This papilionoid subfamily includes the crop legumes and the major model legume species and, thus, is the taxonomic space across which much of legume comparative genomics and "translational genomics" will take place. Most papilionoid species of agronomic interest fall within one of two large clades: first, the Hologalegina clade, containing most of the temperate herbaceous legumes (thus, the colloquial shorthand "temperate herbaceous legumes"), including clovers, vetches, pea, lentil, *Medicago*, and *Lotus*; and second, the Millettioid clade, mostly consisting of tropical and subtropical species, and including common beans, soybean, and cowpea (Maddison and Schulz 1996–2006; Doyle and Luckow 2003; Doyle et al. 1997; Hu et al. 2000). Some commonly encountered genera in Hologalegina are *Vicia, Medicago, Pisum, Trifolium, Cicer, Lens, Astragalus, Wisteria, Lotus, Robinia*, and *Sesbania*. Some commonly encountered genera in the Millettioid clade are *Glycine, Phaseolus, Vigna, Erythrina* (coral bean), and *Apios americana* (groundnut), as well as some earlier-diverging clades, one with the eponymous genus *Milletia*, and the other with *Indigofera* (containing the shrub that was used to produce indigo dye). Beyond of these large clades, basal genera in the papilionoid subfamily include the "dalbergioid" clade, including numerous tropical trees (e.g. rosewood) and *Arachis* (peanut), and the "genistoid" clade, including *Lupinus* (lupine).

Fossil and molecular *dating* methods indicate that most morphological and species diversity in the legumes originated during a burst of speciation early in the Tertiary, ~ 60-50 mya (Lavin et al. 2005; Cronk et al. 2006). This is after the Cretaceous and the major extinction event that ended the "age of the dinosaurs." This early radiation means that, perhaps surprisingly, many early-diverging genera – including those in the caesalpnoideae and mimosoideae – did not originate a great deal earlier than early-diverging lineages in the papilionoidae. Lavin et al. (2005) date the genistoid clade (*Lupinus*) at ~56 mya; the dalbergioid clade (*Arachis*) at ~55 mya; the millettoid clade (*Glycine*) at ~45 mya; and the Hologalegina clade (*Medicago,*

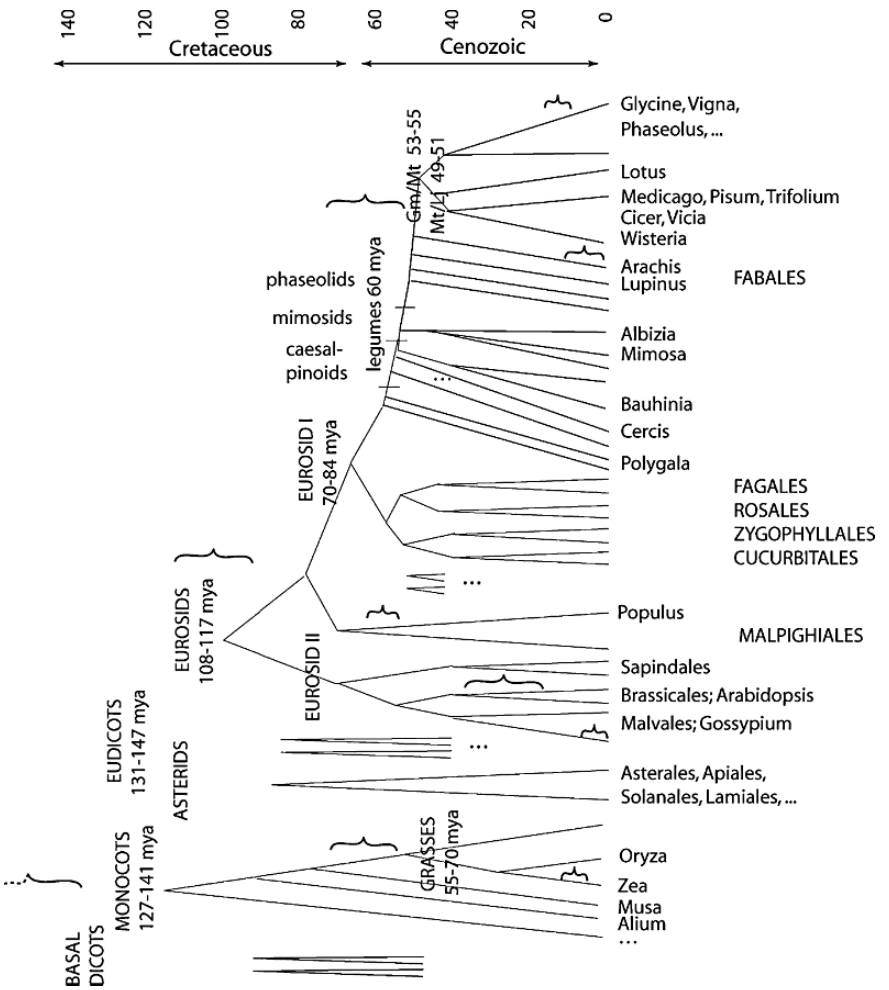

Fig. 3.1 Legume phylogeny, outgroups, and estimated WGD timeframes

Lotus) at ~51 mya. The *Glycine-Medicago* split occurred ~54 mya. And *Medicago* and *Lotus* separated early in Hologalegina, so they diverged at ~51 mya.

These dates and the likely early burst of legume speciation have important implications for comparisons between the model legumes (*Glycine, Medicago, Lotus, Phaseolus*, pea) and other agronomic species. Comparisons between soybean and *Medicago*, or between *Medicago* and *Lotus*, actually require traversing substantial evolutionary time (~50–55 million years to common ancestors). Additionally, evolutionary events that may have occurred "early" in the legumes (most prominently, nodulation or polyploidy) may actually have occurred within a relatively short evolutionary timeframe – of, say, ~10 million years in the early Cenozoic.

Polyploidy and Consequences for Genome Comparisons

Definitions and History

The terms "polyploidy" "paleopolyploidy," and "whole genome duplication" (WGD) all point to the same process of doubling of chromosomal number, and are essentially interchangeable. Over time, with rearrangements and loss of genes and chromosomal segments, the genome "diploidizes," losing most evidence of the original duplication. The terms paleopolyploidy or WGD are perhaps used more frequently to describe ancient events, describing genomes that are a long way towards a diploid state. For particularly old events, duplication remnants may be difficult to distinguish from aneuploid (partial) duplications or other causes of duplication of multiple genomic segments. In these cases, WGD is an inferred, hypothetical event.

Plant genome comparisons established that polyploidy occurred early in angiosperm evolution, and occurred numerous times independently in subsequent plant lineages. Thus, most if not all angiosperms retain remnants of several rounds of WGD. (Masterson 1994; Bowers et al. 2003; De Bodt et al. 2005; Cui et al. 2006).

Ranges for genome duplications (WGD) are shown with braces. References for WGD and clade timings are given in Table 3.1 (following page). Note early radiation in the legumes, indicated by long branch terminal lengths for many lineages (from

Table 3.1 References for estimated dates of clades (A) and genome duplications (B) References and notes in right-hand column are: (1) Tuskan et al. 2006; (2) Sanderson et al. 2004; (3) Wikstrom et al. 2001; (4) Wikstrom et al. 2003; (5) Davies et al. 2004; (6) Lavin et al. 2005; (7) Kellogg 2001; (8) Blanc et al. 2003; (9) Bowers et al. 2003; (10) Paterson et al. 2004; (11) Schlueter et al. 2004; (12) Rauscher et al. 2004; (13) extrapolated from Lavin et al. between papilionid crown and poplar/legume split; also 44–58 mya in Schlueter et al. (2004)

A. Speciations	Example	Date (mya)	Ref.
Rosid I/Rosid II	soybean/Arabidopsis	100–120	1,2
Fabaceae/Salicaceae	soybean/poplar	70–84	3,4,5
Hologalegina/Millettoid	*Medicago*/soybean	54.3 ± 0.6	6
Medicago/Lotus	*Medicago/Lotus*	50.6 ± 0.8	6
dicots/monocots	soybean/rice	131–147	2,3,4
monocot crown age	*Joinvillea* (outgroup)	127–141	3,4,5,7
eudicot crown age	*Ranunculus* (buttercup)	125–147	3,4,5
eurosid I crown age	cucumber	70–84	3,4,5
rosid crown age	*Cercis* (redbud)	108–117	3,4,5
B. Duplications		Date (mya)	
Brassicaceae		24–40	8,9
Salicaceae		60–65	1
Fabaceae		55–80	13
grasses		~70	10
corn		11	10
Glycine max		14.5	11
Glycine tomentella		< 50 kya	12

Lavin et al. 2005). A small number of other lineages are included for reference, including Malpighiales (with poplar), Brassicales (with *Arabidopsis*).

Polyploidy expands allelic variation and phenotypic diversity, it opens the door to functional divergence and innovation in entire metabolic and developmental pathways, and it creates a reproductive barrier and evolutionary bottleneck. Polyploidy also has effects similar to heterosis, with transient increases in measures such as stature, total dry weight, and seed size (e.g. Guo et al. 1996; Bretagnolle and Thompson 2001; Birchler et al. 2003). This effect may be due in part to gene dosage effects: every gene is immediately present in at least two copies. Polyploidy is also of interest because it complicates gene positional comparisons between related species, whereas species with a shared polyploidy history are more likely to have simple chromosomal relationships. All of these characteristics make it important to determine the history of polyploidy events in the legumes.

Prominent Angiosperm Genome Duplications

As it will frequently be helpful to make comparisons between legume and non-legume model genomes, it is important to identify plant lineages in which polyploidy has occurred. Some important WGD events are shown in Fig. 3.1, with timing estimates indicated with blue brackets. Approximate timings of speciation and WGD events are inferred from literature shown in Table 3.1.

In the monocots, all members of the grasses have undergone at least one round of polyploidy, due to an event that affected a progenitor of the grasses at ~70 mya (Paterson et al. 2004; Yu et al. 2005), before radiation of the grasses at ~55–70 mya (Kellogg 2001). Maize has undergone an additional round of polyploidy, as have some other grass lineages such as wheat. Timing of WGD in maize depends on the silent-site/time rate constants used, but the range is ~4.8–11 mya (Swigonova et al. 2004; Paterson et al. 2004).

In the Arabidopsis genome, remnants of three probable WGD are visible: the most recent around 24–40 mya (Blanc et al. 2003; Bowers et al. 2003).

In the poplar genome, WGD occurred near the origin of the Salicaceae, at around 60–65 mya (Tuskan et al. 2006). Interestingly, this genome is changing much more slowly than many plant lineages, likely because gametes from millenia-old clonal populations are regularly pumped into the gene pool in this genus of wind-pollinated trees (Tuskan et al. 2006). As a consequence, many internal duplications ("synteny blocks") are larger and clearer than those in the Arabidopsis genome, even though the WGD episode is estimated to have occurred significantly earlier than the WGD in Arabidopsis.

Remnants of two additional earlier WGD appear to be evident in the poplar and Arabidopsis genomes. The timings of these events is unclear, but recent studies place the middle duplication near the origin of the eurosids, around the split between Eurosid I (with legumes and poplar) and Eurosid II (with Arabidopsis and cotton) (Bowers et al. 2003; Chapman et al. 2006; Sanderson et al. 2004; De Bodt et al.

2005; Tuskan et al. 2006). If the event didn't predate this speciation, then independent events after the speciation are required.

Timing of oldest angiosperm WGD event(s) is still obscure, but likely predates the monocot-dicot split (Bowers et al. 2003; Simillion et al. 2002; Blanc et al. 2003).

Genome Duplications in the Legumes

In the legumes, most evidence points to one round of WGD very early in or shortly preceding the origin of the family. Studies in the 1990s of chromosomal correspondences within the soybean genome, using genetic marker comparisons, suggested that the soybean genome contained at least some regions present in more than two copies (Shoemaker et al. 1996; Lee et al. 2001; Yan et al. 2003).

Self-comparisons of large ESTs data sets from soybean or *Medicago* show a clear recent duplication in soybean, and somewhat weaker evidence for older duplications in soybean and *Medicago* (Schlueter et al. 2004; Blanc and Wolfe 2004). The basis for these studies is that silent-site mutations in homologous gene pairs give a distribution of changes per silent site (often called a "Ks" measurement). The older Ks peaks in soybean and *Medicago* were dated to ~44–64 mya. Schlueter et al. (2004) places a duplication event in *Medicago* at ~58 mya, consistent with ~54 mya estimated by Lavin et al. (2005). Interestingly, Schlueter et al. (2004) also place the early duplication in soybean at ~44 mya. This is significantly earlier than the (likely same) event in *Medicago*, but this might be explained by variance in the Ks peaks and/or different rates of silent-site change in the two lineages. The rate used for these calculations (and in the next paragraph) is 6.1×10^{-9} substitutions per synonymous site per year (Lynch and Conery 2000).

Using similar dating of gene pairs, but taking gene pairs from internal synteny blocks, Mudge et al. (2005) estimated the duplication in *Medicago* occurred at ~64 mya (0.79 synonymous substitutions per site), compared with and *Glycine/Medicago* speciation at 48–50 mya. Similarly, Cannon et al. (2006) found a peak at 0.80 synonymous substitutions per site, corresponding to ~65 mya, using a whole-genome comparison of *Medicago* to itself. This is significantly before the split with *Lotus*, estimated in the same study at ~51 mya (0.64 synonymous substitutions per site) – consistent with the range 50.6 ± 0.9 mya in Lavin et al. (2005). A duplication date of ~65 mya would place the duplication before the ~60 mya origin of the legumes proposed by Lavin et al. (2005) – though all of these molecular rate conversions need to be treated cautiously, as silent-site variation may not always be entirely silent, and rates should not be assumed to be the same in different lineages.

Analyses of the relative phylogenetic positions of genes from several species from a gene family can also be used to determine the relative timings of speciations and duplications. Pfeil et al. (2005) used this approach to establish that an early legume duplication predated the *Medicago/Glycine* split. Using this approach and a comparison of synteny in and between *Medicago* and *Lotus*, Cannon et al. (2006)

confirm that the early legume duplication occurred after the split with poplar and well before the split between *Medicago* and *Lotus*. This brackets the early legume duplication between about 55 and 84 mya. The upper limit (55 mya) is bounded by the estimated divergence time between soybean and *Medicago* (Lavin et al. 2005), and the lower limit (84 mya) is bounded by the split between Fabales (legumes) and Malpighiales (including poplar) (Sanderson et al. 2004).

It should be said that some aspects of the nature and timing of the "early legume duplication" remain unclear. Timing of the Ks frequency peaks is uncertain because the peaks are broad, and we lack molecular clock calibrations for most lineages. In the synteny comparisons of Mt and Lj, the genome self-comparisons (Mt × Mt or Lj × Lj) show extensive fragmentation and loss (absence of synteny) (Cannon et al. 2006 and unpublished results). Only one region of synteny in either of the self-comparisons extends beyond about a megabase; this is between Mt chromosomes 5 and 8, and between corresponding Lj chromosomes 4 and 2. The remaining synteny blocks are scattered and small, occurring on 31 of 36 possible Mt × Mt chromosome pairings (Cannon et al. 2006 and unpublished results). Some of the absence of synteny is undoubtedly explained by the incomplete state of both genome sequences, but the synteny fragmentation is more extensive than is seen in genome self-comparisons of poplar, Arabidopsis, or rice in simulations with similar amounts of loss from these genomes (Wang and Young, unpublished results).

Much of the fragmentation evident in genome self-comparisons may also be due to the large amount of time that has elapsed since this event. If the duplication did occur ~65 mya, as implied by Mudge et al. (2005) and Cannon et al. (2006), this would place it before the origin of the legumes, making it possibly twice as ancient as the most recent duplication in Arabidopsis (at 24–40 mya; see references above).

Taken together, the Ks, phylogenetic, and synteny evaluations do appear to point to an event that affected the majority of the genome, early in or preceding the origin of the legume family. However, a clearer picture of the nature and timing of the early duplication and its aftermath will require completion of the genome sequences, and probably additional partial sequencing from earlier-diverging taxa to better pinpoint the timing of the event.

In addition to the early legume genomic duplication, polyploidy has occurred in several legume lineages. Polyploidy has occurred several times in *Arachis* (peanut, possibly in the course of domestication by early agriculturists (Kochert et al. 1996; Moretzsohn et al. 2004). Polyploidy has also occurred early in the *Glycine* genus (Schlueter et al. 2004; Shoemaker et al. 1996; Lee et al. 1999, 2001; Yan et al. 2003). Additionally, *Glycine tomentella*, a perennial Australian complex of several diploid and allotetraploid "races", has repeatedly undergone polyploidy, ongoing in historical time (Rauscher et al. 2004; Doyle et al. 2004).

As with the grass genomes, polyploidy adds a complication to comparative studies. In *Glycine*, after two rounds of duplication, as many as four homoeologous genomic segments should be expected relative to a corresponding genomic segment from outside the legumes. For comparisons to other plants that have undergone their own independent duplications, such as poplar (Tuskan et al. 2006) or

Arabidopsis (The The Arabidopsis Genome Initiative 2000), the relationships are further complicated: four soybean homoeologs could correspond equally to two poplar homoeologs (and to two more distant poplar homoeologs originating from an earlier WGD). Similarly, four soybean homoeologs could correspond equally to at least four Arabidopsis homoeologs, since Arabidopsis has undergone two rounds of WGD since separation from the Rosid I clade (the clade containing the legumes and poplar) (Tuskan et al. 2006 and unpublished data).

The kinds of many-to-many relationships predicted by two rounds of duplications in the legumes and independent duplications in ancestors of poplar and Arabidopsis are, in fact, seen in such comparisons. Several of these are described below in the context of microsynteny studies.

Synteny

Definitions and History

The term "synteny" was coined to describe genes on the same chromosome (regardless of whether they show linkage in classical tests for recombination). More frequently now, the term is used to indicate conserved gene order between chromosomal regions (either between species or within a duplicated region of one genome).

The paper by Gale and Devos (1998) describing conservation among nine genomes in the grasses provided a conceptual model of the grasses as a coherent genetic system (also reviewed in Freeling 2001; Devos 2005). The key observation – drawing on the work of several groups over the preceding decade – was that large chromosomal blocks have been conserved across most the diverse grass species in the study. Most of the DNA across these genomes could be described in terms of 25 "rice linkage blocks," or portions of the rice genome that retained homologs in the comparison genomes. Some of these blocks are large, corresponding to whole chromosomes in several of the species. In the paper's "circle diagram," showing chromosomal correspondences across seven species, most of rice chromosome 1, for example, is homologous to millet V, sugar cane II and III, Sorghum LG G, parts of maize 3 and 8, and most of oat LG C.At greater evolutionary distances, numerous small and degraded synteny blocks can still be observed, for example, between tomato and Arabidopsis (Ku et al. 2000).

Macrosynteny in the Legumes

Through the 1990s, it was unclear whether synteny was as extensive within the legumes as in the grasses. With the summary by Choi et al. (2004) of more than a decade worth of synteny and comparative marker studies in the legumes, it became clear that synteny extended across broad swaths of diverse species in at least the

Papilionoid subfamily of the legumes (the subfamily containing the majority of legume species, and nearly all of the agronomically important species).

Comparisons between draft sequence from the Mt and Lj genome sequencing projects shows conservation of regions nearly to the length of entire chromosomes in some cases – for example, between Mt 1 and Lj 5 or Mt 2 and Lj 6 (Cannon et al. 2006). However, interestingly, this study finds very little synteny between Mt 6 and any Lj chromosome. The Mt 6 also has an unusually high transposon density, with several density peaks across the chromosome, perhaps indicating knobs such as seen on Arabidopsis chromosome 4. The Mt chromosome 6 also includes an unusually large cluster of genes from the TIR-NBS-LRR disease resistance family. The cause of these unusual patterns is not yet known – whether, for example, the chromosome is new, or is not removing transposons because of suppressed recombination, or is hypermethylated, as is the knob on Arabidopsis chromosome 4 (Gendrel et al. 2002).

The Mt-Lj comparison also confirms the lack of WGD in either lineage following their split at ~40 mya, but does help confirm an earlier WGD event (described further below).

In both the grasses and the legumes, despite extensive conservation for many regions, there are exceptions to these simple mappings. In particular species, some genomic regions have undergone extensive and rapid rearrangement (e.g. Mt 6 or knob on Arabidopsis chromosome 4); polyploidy appears to have occurred early in the legumes; some taxa experienced at least on additional round of polyploidy; and the genomes of some species expanded many-fold through transposon activity.

Effect of Polyploidy on Synteny Comparisons

One factor significantly complicating synteny studies between genomes that have undergone WGD is that polyploidy generates multiple corresponding regions, and appears to spur rearrangement and segmental chromosomal losses (Song et al. 1995; Pontes et al. 2004; Adams and Wendel 2005; Comai et al. 2003).

This pattern of gene loss from duplicated segments is one that Mike Freeling described in terms of "fractionation and consolidation" (Freeling 2001; Langham et al. 2004). Following duplication, selection is initially relaxed for any given gene in a pair. With stochastic loss of one or another of the duplicated genes or regulatory regions, a part ("fraction") of the original complement of genes remains on one homoeolog, and others remain on the other homoeolog. The approximate original complement can be inferred by "consolidating" the genes from the two homoeologous regions.

The problem presented by fractionation following polyploidy is that over time, it would be possible to lose all homology from two regions that nevertheless together contain all the genes from the ancestral region. This high level of fractionation occurred for the homoeologous regions around the loci for maize liguleless2 (*lg2*) and its genomic duplicate, liguleless related sequence1 (*lrs1*) (Langham et al. 2004). Together, these regions contain 13 genes, with only *lg2* and *lrs1* remaining as

duplicated genes. However, 12 of the 13 genes can be found in a corresponding region in rice (Langham et al. 2004). Only by constructing the ancestral gene state is it possible to see the synteny in between the homoeologous regions. A method for inferring such ancestral states is described by Odland et al. (2006). Odland et al. compare rice homoeologous regions (separated ~70 mya), "collapsing" them into simulated ancestral chromosome blocks, and then use these ancestral blocks to make more robust comparisons to collinear maize sequence-based genetic markers.

Synteny studies can provide a means of testing hypothesized genome duplication histories. A study of a 10-cM region from soybean linkage group G homoeologous regions (on linkage group D2 and at least one other linkage group) finds correspondences to six regions in the Arabidopsis genome (Foster-Hartnett et al. 2002). The regions of correspondence probably are short, from several genes up to ~2 Mbp in Arabidopsis (the paper compared sampled sequences from soybean, so has lower resolution than complete sequence will provide). The two longest corresponding regions in Arabidopsis come from the largest contiguous internal Arabidopsis duplication fragment, between Arabidopsis chromosomes 2 and 3. The most likely explanation for this pattern of correspondences is one of at least two independent rounds of polyploidy in the Arabidopsis and soybean lineages, followed by selective gene loss from all resulting regions. A comparison of ~1 Mbp of genomic sequence from two regions in soybean to *Medicago* and Arabidopsis showed similar patterns: the soybean region(s) matching one to two *Medicago* regions and two to four Arabidopsis regions (Mudge et al. 2005). Another study shows correspondences between *Lotus, Medicago*, poplar, and Arabidopsis in regions containing the *Lotus* SYMRK and *Medicago* NORK receptor kinase genes (Kevei et al. 2005). In this region, Kevei et al. find synteny with four regions in Arabidopsis and three in poplar. As with the regions described by Foster-Hartnett et al. (2002) and Mudge et al. (2005), synteny is interrupted between any two regions by interspersed gene losses or local duplications.

Not all classes of genes respond to polyploidy in the same way. In some gene families, such as transcription factors, most genes are retained, whereas other gene families (such as those involved in defense recognition) undergo rapid turnover (Cannon et al. 2004). In a more comprehensive analysis of Arabidopsis genes, Maere et al. (2005) argue that three whole-genome duplications in that genome were directly responsible for >90% of the increase in transcription factors, signal transducers, and developmental genes in the last 350 million years. Chapman et al. (2006) propose that genes retained after polyploidy may buffer critical functions, and further, that gradual loss of this buffering capacity of duplicated genes may contribute to the cyclicality of genome duplication over time.

Microsynteny in the Legumes

Substantial microsynteny is seen in comparisons between *Medicago* and *Lotus* (Choi et al. 2004, 2006; Zhu et al. 2006; Kevei et al. 2005; Cannon et al. 2006),

between *Medicago* and *Glycine* (Mudge et al. 2005), and between *Lotus* and *Glycine* (Hwang et al. 2006).

Quantifying synteny is complicated by tandem duplications and by gene-calling parameters and accuracy. For example, inclusion of coding sequences from transposons would decrease apparent synteny, as would counting of differential tandem expansions in one region vs. the other.

Defining "synteny quality" as twice the number of gene matches divided by the total number of genes in both segments (after excluding transposable elements and collapsing tandem duplications), Cannon et al. (2006) report that "synteny quality" for Mt × Lj is 62% for an extended syntenic block (58/94 of genes exhibit corresponding homologs within these regions). This region is shown in Fig. 3.2. This figure also shows homoeologous segments from Mt and Lj self-comparisons. Synteny quality in the Mt × Mt comparison is just 36%, and is 30% in the Lj × Lj region. The synteny in the self-comparisons of either the Mt or Lj is highly degraded, consistent with a history of very early polyploidy. The synteny seen within *G. max* was variable in the regions examined by Schlueter et al. (2007), but at the high ends, was far less degraded than between any duplications within Mt or Lj. Again, this would be consistent with polyploidy in *G. max* much more recent than in the ancestral legume duplication.

Mudge et al. (2005) made a comparison of several corresponding regions in *Medicago*, soybean, and Arabidopsis. The comparison illustrates several important points. In one synteny comparison spanning ~400 kb from each of two homoeologous regions in *Medicago* and a corresponding soybean region, phylogenetic analysis of each gene in the region shows one of the *Medicago* homoeologs is more closely related to the soybean region (in other words, is separated by speciation and so is orthologous); and the other *Medicago* region separated much earlier and is paralogous to both *Medicago* and soybean regions. This clearly fits a model of an early legume polyploidy, significantly predating the soybean-*Medicago* split. Extent of microsynteny is consistent with this model. Synteny quality between the soybean and *Medicago* orthologous regions is 60% but between the soybean and *Medicago* paralogous regions is 27%. And between the *Medicago* homoeologs, the synteny

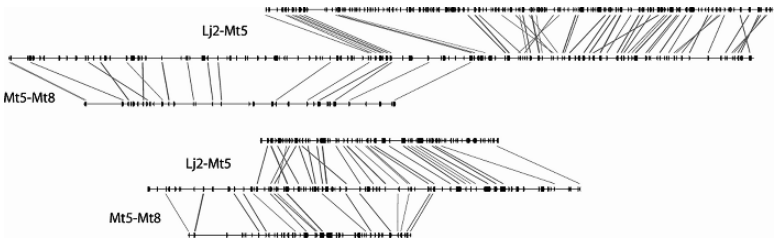

Fig. 3.2 Synteny in selected chromosomal regions between *Medicago* (Mt) and *Lotus* (Lj), and within a duplication in Mt compared with itself. *Top pair*: Lj 2 × Mt 5; *second pair*: Mt 5 × Mt 8; *third pair*: Lj 2 × Mt 5 (another region, within 1 Mbp of first regions); *last pair*: Mt 5 × Mt 8 (also within 1 Mbp of first regions). Note higher densities of collinear genes in the Mt × Lj comparison than in the internal genome duplication. Figure is adapted from Cannon et al. (2006)

quality is only 18% (only nine genes shared of approximately 50 in either *Medicago* homoeolog).

It may be significant in the Mudge et al. study that the synteny quality is lower in the *Medicago* paralogous regions than in the *Medicago-Glycine* homoeolog. Within a single genome, selection pressure should be lowered for duplicated genes, so rate of loss of either gene should be higher than in the stochastic, independent losses between two different genomes. An important conclusion is that internal synteny, remaining after polyploidy, may be more difficult to detect than synteny between two species at separated by a similar amount of time. For example, we might expect much clearer synteny between *Glycine* and *Phaseolus* than between two *Glycine* homoeologs. Relatedly, synteny may turn out to be cleaner between *Medicago* and *Lotus* than between *Medicago* and *Glycine* (which has undergone polyploidy), even though all three species diverged in similar time frames (∼40–50 mya).

Genome Sizes, Gene–Space Organization, and Consequences

An important question for genome comparisons – and for genome studies generally – is the nature of the enormous variation in genome sizes and, relatedly, the nature of gene organization within genomes.

It is now generally accepted that most genome size variation (besides the effect of polyploidy) is explained by expansions of transposons (general: Bennetzen et al. 2005; corn: Du et al. 2006; wheat: Devos et al. 2005; Vitte and Bennetzen 2006; pea: Jing et al. 2005; Vicia: Neumann et al. 2006). Both the grasses and the legumes contain species differing more than 40-fold in genome size. Legume genomes range from *Leucaena macrophylla* (299 Mbp) to *Vicia faba* (13,059 Mbp); and grass genomes range from *Oropetium thomaeum* (245 Mbp) to *Triticum aestivum* (∼16,979 Mbp) (Kew C-values database, Bennett and Leitch 2004). The nature of gene distribution in large genomes will strongly affect the ability to sequence large genomes, and to exchange information between them. This matters in the legumes, as the genomes of several agronomically important species are large. Soybean is ∼1,103 Mbp, pea is ∼4,778 Mbp, and *Vicia faba* is 13,059 Mbp (Bennett and Leitch 2004).

Less well understood is how genes, repetitive sequences, and other DNA are organized genome-wide. What is the range of variation in gene organization in euchromatic regions? What is the range of variation in organization of centromeres, pericentromeres, telomeres, and euchromatin?

In terms of organization in euchromatic regions, accumulating evidence suggests that gene density may be relatively homogeneous across very large regions in most plant genomes (or at least in the diverse genomes under active study) – though can vary enormously between genomes. This contrasts with a "gene islands" model described in several influential papers at the end of the 1990s, which suggested that in large genomes, genes might nevertheless be located in a relatively small "gene-space" – for example as "gene islands in [a] great sea of maize repetitive

DNAs" (SanMiguel et al. 1998; Bennetzen et al. 1998). A concise statement of this model is that "complex cereal genomes are largely composed of small gene-rich regions intermixed with 5–200 kb blocks of repetitive DNA" (Yuan et al. 2002).

In maize, the comparison of the largest contiguous sequence available to-date suggests that the "island" model was, in fact, not an apt description. In a comparison of 7.8 Mbp and 6.6 Mbp from corresponding (homoeologous) regions of the corn genome and 4.9 Mbp from the corresponding region of rice, Bruggmann et al. (2006) report relatively homogeneous gene density within each region, but large differences between homoeologs. They report "Analysis of these two large regions does not reveal evidence of large gene islands separated by retrotransposon blocks. As previously reported, most gene islands are small (one to two genes; Bennetzen et al. 2005) and vary between the different homoeologous regions. A picture is emerging in which different chromosomal regions evolve into a mosaic of syntenic blocks with differential expansion caused by the contraction of genic and intergenic space in combination with the addition of different combinations of repeat elements." (Bruggmann et al. 2006). While one of the maize homoeologs is 95% the size of the rice segment, the other maize homoeolog has expanded over most of its length, to 299% the size of the rice segment. The expansion is due to expansion by factors of 1.2–1.4 in genic regions (UTRs, exons, introns), and 2.3–2.7 for the remaining space (nongenic and repetitive sequence) (Bruggmann et al. 2006). In wheat, with a genome six times larger than corn's, early indications also point to genes widely dispersed rather than in gene-rich "islands." In four sequenced wheat BACs, corresponding to gene-rich regions in rice, each wheat BAC contained one-two genes, with a gene density of ~1 gene per 75 kb.

What is the range of variation in organization of centromeres, pericentromeres, telomeres, and euchromatin? In sequenced genomes to-date, there is a gradual transition on nearly all chromosomes from gene-rich euchromatic regions to multi-megabase, gene-poor, transposn-rich "pericentromeric" regions, and finally to tandem arrays of many "satellite repeats" of ~180 bp. However, the sizes of these regions (both pericentromeric and satellite repeats) vary greatly (reviewed in Ma et al. 2007; Hall et al. 2004; Lam et al. 2004).

Sizes of the satellite repeat regions ranges from 0.4 to 1.4 Mbp in Arabidopsis chromosomes, and 60 kb–1.9 Mbp in rice (Ma et al. 2007). Sizes of these regions vary even within ecotypes in a species, as shown in Arabidopsis, maize, rice, and *Medicago* (reviewed in Ma et al. 2007). In Mt, measurable genome size differences in accessions Jemalong A17 and R108-1 are due, at least in part, to much shorter satellite repeat regions in R108-1 (Kulikova et al. 2004). In Mt Jemalong A17 (the variety being sequenced), three types of satellite repeats were identified: MtR1 and MtR2, which are found in pericentromeric regions of all chromosomes, and MtR3 which is suspected to be the functional centromere domain, varying in size from ~450 bp to more than 1 Mbp (Kulikova et al. 2004). Together, these three types of satellite repeats are estimated to comprise ~6.5–8% of the Mt Jemalong A17 genome. Although these core satellite repeat regions are generally expected to contain few or no genes, there may be exceptions, as in the ~45 transcribed genes in the core centromeric region of rice chromosome 8 (Nagaki et al. 2004).

The sizes of pericentromeric regions also vary significantly between species and between chromosomes within a genome. In Arabidopsis, roughly 93% of the genome was characterized as euchromatic (Koornneef et al. 2003). However, pericentromere boundaries are not clear, and are probably better characterized as "pericentromeric gradients," with gene densities declining and transposon densities increasing (particularly class I LTR transposons) over spans of ~2–5 Mbp on approaches to centromeres (The Arabidopsis Genome Initiative 2000). Thus, depending on definition of pericentromeric border, roughly 25 Mbp (20%) of the Arabidopsis genome could be considered pericentromeric. In poplar, approximately 70% of the genome is primarily euchromatic, with most of the remainder being in pericentromeric gradients (Tuskan et al. 2006). Similarly, in rice, regions of ~2–10 Mbp are in pericentromeric gradients (Yu et al. 2005, e.g. Fig. S5).

The nature of the pericentromere will be of critical importance for every plant genome sequencing project and for comparative genomic work. Sequencing and genome assembly are made difficult by the high repeat content in pericentromeric heterochromatin, and gene content is low. Similarly, genome comparisons in these regions are complicated by large numbers of transposon sequences and high rates of turnover in transposons. There are a number of interesting unanswered questions. How frequently do centromere locations change? What genes tolerate location in pericentromeric heterochromatin? What range of gene densities exist in pericentromeric regions? Further sequencing in *Medicago, Lotus*, tomato, soybean and numerous other plant genomes will help answer these and many other unanswered questions.

Conclusions

Comparative genomics will be crucial for translating knowledge between model species, and between models and crop species. The transfer won't just be a matter of research efficiency, but will be important for making best use of the enormously inventive germplasm and phenotypic variation across the legumes. There is a great deal of benefit in considering the legumes as a coherent, broad genetic system. This concept was stated succinctly in a report on the 2004 meeting on the "Legume Crops Genome Initiative" (LCGI): "Cross-legume genomics seeks to advance: (1) knowledge about the legume family as a whole; (2) understanding about the evolutionary origin of legume-characteristic features such as rhizobial symbiosis, flower and fruit development, and its nitrogen economy; and (3) pooling of genomic resources across legume species to address issues of scientific, agronomic, environmental, and societal importance." (Gepts et al. 2005). What is the "model" will depend on the trait being studied. For example, it will make more sense to study oil biosynthesis in soybean, cold-tolerance in alfalfa, and phosphate uptake in lupin. Lessons learned in any of these "models" will be applicable, however, in the other species.

Comparative genomic techniques will not be useful solely as a means of positional cloning or gene-finding in related species. The techniques have the capacity to

elucidate how traits have evolved and continue to evolve. For example, it is only by comparing nodulation in diverse species that we will learn how this important trait originated, whether once or several times, using what existing molecular machinery; etc. Similarly, comparisons will show how the diversity – and capacity for change – in defense response mechanisms. The same can be said of a very large number of traits: flavanoid biosynthesis, perenniality, nutrient uptake; etc.

Although we have a great deal to learn about comparative structural genomics, some tentative general conclusions can be drawn about the three questions posed in the introduction. First: what is the organization of genes and non-genes? In genomes sequenced to-date, gene organization has been essentially similar: locally relatively homogeneous gene densities across most parts of most euchromatic arms, and declining gradually on approach to the centromere. However, some regions break this pattern – for example, knobs in Arabidopsis, or transposon-dense chromosome 6 in *Medicago*. Second, what are the mechanisms of large-scale genome change? Undoubtedly, this question will be answered negatively in interesting, particular ways. Broadly, though, it appears that all lineages do change primarily by the same mechanisms, such as polyploidy, breakages, fusions, inversions, translocations, and transposon insertions. However, not all lineages experience these effects in similar dose. Some lineages experience multi-fold expansion due to transposon activation, some undergo more episodes of polyploidy, and some have experienced greater levels of rearrangement. Third, what is the pace of synteny loss? Clearly, rates of rearrangement differ in various lineages (e.g. slower in poplar, more rapidly in Arabidopsis), but detectable macrosynteny remains over the range of 100 million years. However, microsynteny is strongly affected by fractionation following large-scale genomic duplications. Within a duplicated region, many single gene pairs may be degraded.

Comparisons of genomes sequenced to-date show both remarkable conservation and change. It is amazing to realize that through phylogenetic comparisons, or structural-genomic comparisons, it is possible to see the trace of events that occurred in early angiosperm evolultion, or earlier – in effect, "genome archaeology." At the same time, it is clear that genomes are dynamic, creative, rapidly-changing environments in particular ways. By understanding processes such as polyploidy, local gene duplications and losses, rearrangements, and rapid turnover of repetitive elements, we have the opportunity to see plant biology more clearly, both distant-past and present.

References

Adams, K.L., and Wendel, J.F. (2005) Polyploidy and genome evolution in plants: Genome studies and molecular genetics. Curr. Opin. Plant Biol. 8:135–141.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814):796–815.

Bennett, M.D. and Leitch I.J. (2004) Angiosperm DNA C-values database (release 5.0, Dec. 2004) http://www.rbgkew.org.uk/cval/homepage.html

Bennetzen, J.L., SanMiguel P., Chen, M., Tikhonov, A., Francki, M., Avramova, Z. (1998) Grass genomes. Proc. Natl. Acad. Sci. USA 95:1975–1978

Bennetzen, J.L., Ma, J., Devos, K.M. (2005) Mechanisms of recent genome size variation in flowering plants. An. Bot. 95:127–132.

Birchler, J.A., Auger, D.L., and Riddle, N.C. (2003) In search of a molecular basis of heterosis. Plant Cell 15: 2236–2239.

Blanc, G., Hokamp, K.H., and Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. 13,137–144.

Blanc, G., Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678.

Bowers J.E., Chapman B.A., Rong J., and Paterson A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422: 433–436.

Bretagnolle, F. and Thompson, J.D. (2001) Phenotypic plasticity in sympatric diploid and autotetraploid *Dactylis glomerata*. Int. J. Plant Sci. 162: 309–316.

Bruggmann, R., Bharti, A.K., Gundlach, H., Lai, J., Young, S., Pontaroli, A.C., Wei, F., Haberer, G., Galina, F., Du, C., Raymond, C., Estep, M.C., Liu, R., Bennetzen, J.L., Chan, A.P., Rabinowicz, P.D., Quackenbush, J., Barbazuk, W.B., Wing, R.A., Birren, B., Nusbaum, C., Rounsley, S., Mayere, K.F.X., Messing, J. (2006) Uneven chromosome contraction and expansion in the maize genome. Genome Res. 16:1241–1251.

Cannon, S.B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J.P., Wang, X., Mudge, J., Vasdewani, J., Scheix, T., Spannagl, M., Nicholson, C., Humphray, S.J., Schoof, H., Mayer, K.F.X., Rogers, J., Quetier, F., Oldroyd, G.E., Debelle, F., Cook, D.R., Retzel, E.F., Roe, B.A., Town, C.D., Tabata, S., Van de Peer, Y., and Young, N.D. (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. Proc. Natl. Acad. Sci. USA 103(40):14959–64.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D., May, G. (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. BMC Plant Biol. 4:10

Chapman B.A., Bowers J.E., Feltus F.A., and Paterson A.H. (2006) Buffering crucial functions by paleologous duplicated genes may contribute to cyclicality to angiosperm genome duplication. Proc. Natl. Acad. Sci. USA. 103(8):2730–2735.

Choi, H-K., Mun, J-H., Kim, D-J., Zhu, H., Baek, J-M., Mudge, J., Roe, B.A., Ellis, N., Doyle, J., Kiss, G.B., Young, N.D., Cook, D.R. (2004) Estimating genome conserevation between crop and model legume species. Proc. Natl. Acad. Sci. USA. 101(45):15289–15294.

Choi, H-K., Luckow, M.A., Doyle, J.J., and Cook, D.R. (2006) Development of nuclear gene-derived markers linked to legume genetic maps. Mol. Genet. Genom. 276(1):56–70.

Comai, L., Madlung, A., Josefsson, C., and Tyagi, A. (2003) Do the different parental 'heteromes' cause genomic shock in newly formed allopolyploids? Philos. Trans Royal Soc. 358: 1149–55.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., and dePamphilis, C.W. (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res. 16: 738–749.

Cronk Q., Ojeda I., and Pennington R.T. (2006) Legume comparative genomics: progress in phylogenetics and phylogenomics. Curr. Opin. Plant Biol. 9(2):99–103.

Davies, T.J., Barraclough, T.G., Chase, M.W., Soltis, P.S., Soltis, D.E., Savolainen, V. (2004) Darwin's abominable mystery: insights from a supertree of the angiosperms. Proc. Natl. Acad. Sci. USA 101: 1904–1909.

De Bodt S., Maere S., and Van de Peer Y. (2005) Genome duplication and the origin of angiosperms. Trends Ecol. Evol. 20(11):592–597.

Devos, K.M. (2005) Updating the 'crop circle'. Curr. Opin. Plant Biol. Apr;8(2):155–62.

Devos, K.M., Ma J., Pontaroli A.C., Pratt L.H., and Bennetzen J.L. (2005) Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. Proc. Natl. Acad. Sci. U S A. 2005 102(52):19243–8

Doyle, J.J. and Luckow, M.A. (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiology 131:900–910.

Doyle, J.J., Doyle, J.L., Rauscher, J.T., and Brown, A.H.D. 2004. Evolution of the perennial soybean polyploid complex (*Glycine* subgenus *Glycine*): A study of contrasts. Biol. Journal Linnean Soc. 82:583–597.

Doyle, J.J., Doyle, J.L., Ballenger, J.A., Dickson, E.E., Kajita, T., and Ohashi, H. (1997) A phylogeny of the chloroplast gene *rbcL* in the Leguminosae: taxonomic correlations and insights into the evolution of nodulation. Am. J. Bot. 84: 541–554.

Du, G., Swigonova, Z., and Messing, J. (2006) Retrotranspositions in orthologous regions of closely related grass species. BMC Evol Biol 6:62

Foster-Hartnett, D., Mudge, J., Larsen, D., Danesh, D., Yan, H., Denny, R., Penuela, S., and Young, N.D. (2002) Comparative genomic analysis of sequences sampled from a small region on soybean (*Glycine* max) molecular linkage group G. Genome. 45(4):634–45.

Freeling, M., (2001) Grasses as a single Genetic System: Reassessment 2001. Plant Physiol. 125:1191–1197.

Gale, M.D., Devos, K.M. (1998) Comparative genetics in the grasses. Proc. Natl. Acad. Sci. USA 95:1971–1974.

Gendrel, A-V., Lippman, Z.Z., Yordan, C.C., Colot, V., and Martienssen, R. (2002) Heterochromatic histone H3 methylation patterns depend on the Arabidopsis gene *DDM1*. Science 297: 1871–1873.

Gepts, P., Beavis, W.D., Brummer, E.C., Shoemaker, R.C., Stalker, H.T., Weeden, N.F., Young, N.D. (2005) Legumes as a model plant family. Genomics for food and feed report of the cross-legume dvances through genomics conference. Plant Physiol. 137:1228–1235.

Guo, M., Davis, D., and Birchler, J.A. (1996) Dosage Effects on Gene Expression in a Maize Ploidy Series. Genetics 142:1349–1355.

Hall, A.E., Keith, K.C., Hall, S.E., Copenhaver, G.P., Preuss, D. (2004) The rapidly evolving field of plant centromeres. Curr Opin Plant Biol 7:108–114.

Hu, J.-M., Lavin, M., Wojciechowski, M. and Sanderson, M.J. (2000) Phylogenetic systematics of the tribe Millettieae (Leguminosae) based on trnK/matK sequences, and its implications for the evolutionary patterns in Papilionoideae. American Journal of Botany 87(3): 418–430.

Hwang, T-Y., Moon, J-K., Yu, S., Yang, K., Mohankumar, S., Yu, Y.H., Lee, Y.H., Kim, H.S., Kim, H.M., Maroof, M.A.S., Jeong, S-C. (2006) Application of comparative genomics in developing molecular markers tightly linked to the virus resistance gene Rsv4 in soybean. Genome 49:380–388.

Jing, R., Knox, M.R., Lee, J.M., Vershinin, A.V., Ambrose, M., Ellis, N.E., and Flavell, A.J. (2005) Insertional polymorphism and antiquity of PDR1 retrotransposon insertions in Pisum species. Genetics 171:741–752.

Kellogg E.A. (2001) Evolutionary history of the grasses. Plant Physiol 125:1198–1205.

Kevei, Z., Seres, A., Kereszt, A., Kalo, P., Kiss, P., Toth, G., Endre, G., Kiss, G.B. (2005) Significant microsynteny with new evolutionary highlights is detected between Arabidopsis and legume model plants despite the lack of macrosynteny. Mol Gen Genomics (2005) 274: 644–657.

Kochert, G., Stalker, H.T., Gimenes, M., Galgaro, L., Lopes, C.R., and Moore, K. (1996) RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, Arachis hypogaea (Leguminosae). Am J Bot 83:1282–1291.

Koornneef, M., Fransz, P., de Jong, H. (2003) Cytogenetic tools for Arabidopsis thaliana. Chromosome Res. 11(3)183–194.

Ku, H.M., Vision, T., Liu J.P., and Tanksley, S.D. (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. U.S.A. 97: 9121–9126.

Kulikova, O., Geurts, R., Lamine, M., Kim, D-J., Cook, D.R., Leunissen, J., de Jong, H., Roe, B.A., Bisseling, T. (2004) Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago* truncatula. Chromosoma 113:276–283.

Lam, E., Kato, N., Watanabe, K. (2004) Visualizing chromosome structure/organization. Annu. Rev. Plant Biol. 55:537–54.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A., and Freeling, M. (2004). Genomic duplication, fractionation and the origin of regulatory novelty. Genetics,

Lavin, M., Herendeen, P.S., and Wojciechowski, M.F. (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. Systematic Biology 54: 530–549.

Lee, J.M., Bush, A., Specht, J.E., and Shoemaker, R. (1999). Mapping duplicate genes in soybean. Genome, 42: 829–836.

Lee, J.M., Grant, D., Vallejos, C.E., and Shoemaker, R. (2001) Genome organization in dicots. II. Arabidopsis as a 'bridging species' to resolve genome evolution events among legumes. Theor. Appl. Genet. 103: 765–773.

Lewis, G.P., Schrire, B.D., Mackinder, B.A., Lock, J.M., ed. (2003) Legumes of the World. Royal Botanic Gardens, Kew, UK

Lynch M. and Conery J.S. (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

Ma, J., Wing, R.A., Bennetzen, J.L., Jackson, S.A. (2007) Plant centromere organization: a dynamic structure with conserved functions. Trends in Genetics 23(3):134–139.

Maddison, D.R. and K.-S. Schulz (eds.) 1996–2006. The Tree of Life Web Project. Internet address: http://tolweb.org

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Montagu, M.V., Kuiper, M., Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 102(15):5454–5459.

Masterson, J. (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science (Washington, D.C.), 264: 421–424.

Moretzsohn M.C., Hopkins M.S., Mitchell S.E., Kresovich S, Valls J.F.M., and Ferreira M.E. (2004) Genetic diversity of peanut (Arachis hypogaea L.) and its wild relatives based on the analysis of hypervariable regions of the genome. BMC Plant Biology 2004, 4:11.

Mudge J., Cannon, S.B., Kalo, P., Oldroyd, G.E.D., Roe, B.A., Town, C.D., Young, N.D. (2005) Highly syntenic regions in the genomes of soybean, *Medicago* truncatula, and Arabidopsis thaliana. BMC Plant Biology 2005, 5:15.

Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P.B., Kim, M., Jones, K.M., Henikoff, S., Buell, R.C., Jiang, J. (2004) Sequencing of a rice centromere uncoveres active genes. Nature Genetics 36(2):139–145.

Neumann, P., Koblizkova, A., Navratilova, A., Macas, J. (2006) Significant Expansion of Vicia pannonica Genome Size Mediated by Amplification of a Single Type of Giant Retroelement. Genetics 173:1047–1056.

Odland W., Baumgarten A., and Phillips R. (2006) Ancestral rice blocks define multiple related regions in the maize genome. Crop Science 46:41–48.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc. Natl. Acad. Sci. USA 101:9903–9908.

Pfeil, B.E., Schlueter, J.A. Shoemaker, R.C., and Doyle, J.J. (2005) Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Systematic Biology 54:441–454.

Pontes, O., Neves N., Silva M., Lewis M.S., Madlung A., Comai L., Viegas W., and Pikaard C.S. (2004) Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid Arabidopsis suecica genome. Proc. Natl. Acad. Sci. USA 101: 18240–18245.

Rauscher J.T., Doyle J.J., and Brown A.H. (2004) Multiple origins and nrDNA internal transcribed spacer homeologue evolution in the *Glycine* tomentella (Leguminosae) allopolyploid complex. Genetics 166(2):987–98.

Sanderson M.J., Thorne J.L., Wikstrom N, and Bremer K. (2004) Molecular evidence on plant divergence times. Am. J. Bot. 91(10):1656–1665.

SanMiguel P., Gaut B.S., Tikhonov A., Nakajima Y., and Bennetzen J.L. (1998) The paleontology of intergene retrotransposons of maize. Nat Genet. 20(1):43–5.

Schlueter, J.A., Dixon P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C. (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome 47: 868–876.

Schlueter, J.A., Lin, J-Y, Schlueter, S.D., Vasylenko-Sanders, I.F., Deshpande, S., Yi, J., O'Bleness, M., Roe, B.A., Nelson, R.T., Scheffler, B.E., Jackson, S.A., Shoemaker, R.C. (2007) Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. BMC Genomics 2007, 8:330.

Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N.D., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G., and Boerma, H.R. (1996) Genome duplication in soybean (*Glycine* subgenus soja). Genetics, 144: 329–338.

Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M., and Van de Peer, Y. (2002) The hidden duplication past of Arabidopsis thaliana. Proc. Natl Acad. Sci. USA 99, 13627–13632.

Song, K., Lu, P., Tang, K., and Osborn, T.C. 1995. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc. Natl. Acad. Sci. USA 92: 7719–7723.

Swigonova, Z., J. Lai, J. Ma, W. Ramakrishna, V. Llaca, J.L. Bennetzen, and J. Messing. (2004) Close split of sorghum and maize genome progenitors. Genome Res. 14:1916–1923.

Tuskan G.A., DiFazio S., Jansson S., Bohlmann J. et al. (2006) The genome of black cottonwood (Populus trichocarpa (Torr. & Gray). Science 313:1596–1604.

Vitte, C. and Bennetzen, J.L. (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci U S A. 103(47):17638–43.

Wikstrom, N., Savolainen, V., and Chase, M.W. (2001) Evolution of the angiosperms: calibrating the family tree. Proceedings of the Royal Society of London, series B 268: 2211–2220.

Wikstrom, N., Savolainen, V., and Chase, M.W. (2003) Angiosperm divergence times: congruence and incongruence between fossils and sequence divergence estimates. In P. C. J. Donoghue and M. P. Smith [eds.], Telling the evolutionary time: molecular clocks and the fossil record, 142–165. Taylor & Francis, London, UK.

Yan H.H., Mudge J., Kim D.J., Larsen D., Shoemaker R.C., Cook D.R., and Young N.D. (2003) Estimates of conserved microsynteny among the genomes of *Glycine* max, *Medicago* truncatula and Arabidopsis thaliana. Theor Appl Genet 106(7):1256–65.

Young N.D., Cannon, S.B., Sato, S., Kim, D.J., Cook, D.R., Town, C.D., Roe, B.A., Tabata, S. (2005) Sequencing the genespaces of *Medicago* truncatula and *Lotus* japonicus. Plant Physiology 137:1174–1181.

Yu J., Wang J., Lin W., Li S., Li H., et al. (2005) The genomes of Oryza sativa: A history of duplications. PLoS Biol 3(2):e38.

Yuan Y., SanMiguel P.J., and Bennetzen J.L. (2002) Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in Zea mays. Genome Res. 12(9):1345–9.

Zhu, H., Riely, B.K., Burns, N.J., Ane, J-M. (2006) Tracing Nonlegume Orthologs of Legume Genes Required for Nodulation and Arbuscular Mycorrhizal Symbioses. Genetics 172: 2491–2499.