1-2010

Genome sequence of the palaeopolyploid soybean

Jeremy Schmutz
HudsonAlpha Genome Sequencing Center, Huntsville, AL

Steven B. Cannon
USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, Iowa

Jessica Schlueter
University of North Carolina at Charlotte

Jianxin Ma
Purdue University

Therese Mitros
University of California, Berkeley

See next page for additional authors

Schmutz, Jeremy; Cannon, Steven B.; Schlueter, Jessica; Ma, Jianxin; Mitros, Therese; Nelson, William; Hyten, David L.; Song, Qijian; Thelen, Jay J.; Cheng, Jianlin; Xu, Dong; Hellsten, Uffe; May, Gregory D.; Yu, Yeisoo; Sakurai, Tetsuya; Umezawa, Taishi; Bhattacharyya, Madan K.; Sandhu, Devinder; Valliyodan, Babu; Lindquist, Erika; Peto, Myron; Grant, David; Shu, Shengqiang; Goodstein, David; Barry, Kerrie; Futrell-Griggs, Montona; Abernathy, Brian; Du, Jianchang; Tian, Zhixi; Zhu, Liucun; Gill, Navdeep; Joshi, Trupti; Libault, Marc; Sethuraman, Ananad; Zhang, Xue-Cheng; Shinozaki, Kazuo; Nguyen, Henry T.; Wing, Rod A.; Cregan, Perry; Specht, James E.; Grimwood, Jane; Rokhsar, Dan; Stacey, Gary; Shoemaker, Randy C.; and Jackson, Scott A., "Genome sequence of the palaeopolyploid soybean" (2010). *Agronomy & Horticulture -- Faculty Publications*. 366.
https://digitalcommons.unl.edu/agronomyfacpub/366

ARTICLES

Genome sequence of the palaeopolyploid soybean

Jeremy Schmutz[1,2], Steven B. Cannon[3], Jessica Schlueter[4,5], Jianxin Ma[5], Therese Mitros[6], William Nelson[7], David L. Hyten[8], Qijian Song[8,9], Jay J. Thelen[10], Jianlin Cheng[11], Dong Xu[11], Uffe Hellsten[2], Gregory D. May[12], Yeisoo Yu[13], Tetsuya Sakurai[14], Taishi Umezawa[14], Madan K. Bhattacharyya[15], Devinder Sandhu[16], Babu Valliyodan[17], Erika Lindquist[2], Myron Peto[3], David Grant[3], Shengqiang Shu[2], David Goodstein[2], Kerrie Barry[2], Montona Futrell-Griggs[5], Brian Abernathy[5], Jianchang Du[5], Zhixi Tian[5], Liucun Zhu[5], Navdeep Gill[5], Trupti Joshi[11], Marc Libault[17], Anand Sethuraman[1], Xue-Cheng Zhang[17], Kazuo Shinozaki[14], Henry T. Nguyen[17], Rod A. Wing[13], Perry Cregan[8], James Specht[18], Jane Grimwood[1,2], Dan Rokhsar[2], Gary Stacey[10,17], Randy C. Shoemaker[3] & Scott A. Jackson[5]

Soybean (*Glycine max*) is one of the most important crop plants for seed protein and oil content, and for its capacity to fix atmospheric nitrogen through symbioses with soil-borne microorganisms. We sequenced the 1.1-gigabase genome by a whole-genome shotgun approach and integrated it with physical and high-density genetic maps to create a chromosome-scale draft sequence assembly. We predict 46,430 protein-coding genes, 70% more than *Arabidopsis* and similar to the poplar genome which, like soybean, is an ancient polyploid (palaeopolyploid). About 78% of the predicted genes occur in chromosome ends, which comprise less than one-half of the genome but account for nearly all of the genetic recombination. Genome duplications occurred at approximately 59 and 13 million years ago, resulting in a highly duplicated genome with nearly 75% of the genes present in multiple copies. The two duplication events were followed by gene diversification and loss, and numerous chromosome rearrangements. An accurate soybean genome sequence will facilitate the identification of the genetic basis of many soybean traits, and accelerate the creation of improved soybean varieties.

Legumes are an important part of world agriculture as they fix atmospheric nitrogen by intimate symbioses with microorganisms. The soybean in particular is important worldwide as a predominant plant source of both animal feed protein and cooking oil. We report here a soybean whole-genome shotgun sequence of *Glycine max* var. Williams 82, comprised of 950 megabases (Mb) of assembled and anchored sequence (Fig. 1), representing about 85% of the predicted 1,115-Mb genome[1] (Supplementary Table 3.1). Most of the genome sequence (Fig. 1) is assembled into 20 chromosome-level pseudomolecules containing 397 sequence scaffolds with ordered positions within the 20 soybean linkage groups. An additional 17.7 Mb is present in 1,148 unanchored sequence scaffolds that are mostly repetitive and contain fewer than 450 predicted genes. Scaffold placements were determined with extensive genetic maps, including 4,991 single nucleotide polymorphisms (SNPs) and 874 simple sequence repeats (SSRs)[2–5]. All but 20 of the 397 sequence scaffolds are unambiguously oriented on the chromosomes. Unoriented scaffolds are in repetitive regions where there is a paucity of recombination and genetic markers (see Supplementary Information for assembly details).

The soybean genome is the largest whole-genome shotgun-sequenced plant genome so far and compares favourably to all other high-quality draft whole-genome shotgun-sequenced plant genomes (Supplementary Table 4). A total of 8 of the 20 chromosomes have telomeric repeats (TTTAGGG or CCCTAAA) on both of the distal scaffolds and 11 other chromosomes have telomeric repeats on a single arm, for a total of 27 out of 40 chromosome ends captured in sequence scaffolds. Also, internal scaffolds in 19 of 20 chromosomes contain a large block of characteristic 91- or 92-base-pair (bp) centromeric repeats[6,7] (Fig. 1). Four chromosome assemblies contain several 91/92-bp blocks; this may be the correct physical placements of these sequences, or may reflect the difficulty in assembling these highly repetitive regions.

Gene composition and repetitive DNA

A striking feature of the soybean genome is that 57% of the genomic sequence occurs in repeat-rich, low-recombination heterochromatic regions surrounding the centromeres. The average ratio of genetic-to-physical distance is 1 cM per 197 kb in euchromatic regions, and 1 cM per 3.5 Mb in heterochromatic regions (see Supplementary Information section 1.8). For reference, these proportions are similar to those in *Sorghum*, in which 62% of the sequence is heterochromatic, and different than in rice, with 15% in heterochromatin[8]. In

[1]HudsonAlpha Genome Sequencing Center, 601 Genome Way, Huntsville, Alabama 35806, USA. [2]Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [3]USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, Iowa 50011, USA. [4]Department of Bioinformatics and Genomics, 9201 University City Blvd, University of North Carolina at Charlotte, Charlotte, North Carolina 28223, USA. [5]Department of Agronomy, Purdue University, 915 W. State Street, West Lafayette, Indiana 47906, USA. [6]Center for Integrative Genomics, University of California, Berkeley, California 94720, USA. [7]Arizona Genomics Computational Laboratory, BIO5 Institute, 1657 E. Helen Street, The University of Arizona, Tucson, Arizona 85721, USA. [8]USDA, ARS, Soybean Genomics and Improvement Laboratory, B006, BARC-West, Beltsville, Maryland 20705, USA. [9]Department Plant Science and Landscape Architecture, University of Maryland, College Park, Maryland 20742, USA. [10]Division of Biochemistry & Interdisciplinary Plant Group, 109 Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA. [11]Department of Computer Science, University of Missouri, Columbia, Missouri 65211, USA. [12]The National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505, USA. [13]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, Arizona 85721, USA. [14]RIKEN Plant Science Center, Yokohama 230-0045, Japan. [15]Department of Agronomy, Iowa State University, Ames, Iowa 50011, USA. [16]Department of Biology, University of Wisconsin-Stevens Point, Stevens Point, Wisconsin 54481, USA. [17]National Center for Soybean Biotechnology, Division of Plant Sciences, University of Missouri, Columbia, Missouri 65211, USA. [18]Department of Agronomy and Horticulture, University of Nebraska, Lincoln, Nebraska 68583, USA.

Figure 1 | Genomic landscape of the 20 assembled soybean chromosomes. Major DNA components are categorized into genes (blue), DNA transposons (green), *Copia*-like retrotransposons (yellow), *Gypsy*-like retrotransposons (cyan) and Cent91/92 (a soybean-specific centromeric repeat (pink)), with respective DNA contents of 18%, 17%, 13%, 30% and 1% of the genome sequence. Unclassified DNA content is coloured grey. Categories were determined for 0.5-Mb windows with a 0.1-Mb shift.

general, these boundaries, determined on the basis of suppressed recombination, correlate with transitions in gene density and transposon density. Ninety-three per cent of the recombination occurs in the repeat-poor, gene-rich euchromatic genomic region that only accounts for 43% of the genome. Nevertheless, 21.6% of the high-confidence genes are found in the repeat- and transposon-rich regions in the chromosome centres.

We identified 46,430 high-confidence protein-coding loci in the soybean genome, using a combination of full-length complementary DNAs[9], expressed sequence tags, homology and *ab initio* methods (Supplementary Information section 2). Another ~20,000 loci were predicted with lower confidence; this set is enriched for hypothetical, partial and/or transposon-related sequences, and possess shorter coding sequences and fewer introns than the high-confidence set. The exon–intron structure of genes shows high conservation among soybean, poplar and grapevine, consistent with a high degree of position and phase conservation found more broadly across angiosperms[10]. Introns in soybean gene pairs retained in duplicate have a strong tendency to persist. Of 19,775 introns shared by poplar and grapevine (diverged more than 90 million years (Myr) ago[11]), and hence by the last common ancestor of soybean and grapevine, 19,666 (99.45%) were preserved in both copies in soybean. Of the remaining 0.55%, 78% are absent in both recent soybean copies (that is, lost before the ~13-Myr-ago duplication) and 22% are found only in one paralogue (that is, other copy lost). We find a slower intron loss rate in poplar (0.4%) than in soybean (0.6%) since the last common rosid ancestor, which is consistent with the slower rate of sequence evolution in the poplar lineage thought to be associated with its perennial, clonal habit, global distribution and wind pollination[12]. Intron size is also highly conserved in recent soybean paralogues, indicating that few insertions and deletions have accumulated within introns over the past 13 Myr.

Of the 46,430 high-confidence loci, 34,073 (73%) are clearly orthologous with one or more sequences in other angiosperms, and can be assigned to 12,253 gene families (Supplementary Table 5). Among pan-angiosperm or pan-rosid gene families that also have members outside the legumes, soybean is particularly enriched (using a Fisher's exact test relative to *Arabidopsis*) in genes containing NB-ARC (nucleotide-binding-site-APAF1-R-Ced) and LRR (leucine-rich-repeat) domains. These genes are associated with the plant immune system, and are known to be dynamic[13]. Tandem gene family expansions are common in soybean and include NBS-LRR, F-box, auxin-responsive protein, and other domains commonly found in large gene families in plants. The ages of genes in these tandem families, inferred from intrafamily sequence divergence, indicate that they originated at various times in the evolutionary history of soybean, rather than in a discrete burst.

From protein families in the sequenced angiosperms (http://www.phytozome.net) (Supplementary Table 4), we identified 283 putative legume-specific gene families containing 448 high-confidence soybean genes (Supplementary Information section 2). These gene families include soybean and *Medicago* representatives, but no representatives from grapevine, poplar, *Arabidopsis*, papaya, or grass (*Sorghum*, rice, maize, *Brachypodium*). The top domains in this set are the AP2 domain, protein kinase domain, cytochrome P450, and PPR repeat. An additional 741 putatively soybean-specific gene families (each consisting of two or more high-confidence soybean genes) may also include legume-specific genes that have not yet been sequenced in the ongoing *Medicago* sequencing project, or may represent bona fide soybean-specific genes. The top domains in this list include protein kinase and protein tyrosine kinase, AP2, LRR, MYB-like DNA binding domain, cytochrome P450 (the same domains most common in the entire soybean proteome) as well as GDSL-like lipase/acylhydrolase and stress-upregulated Nod19.

A combination of structure-based analyses and homology-based comparisons resulted in identification of 38,581 repetitive elements, covering most types of plant transposable elements. These elements, together with numerous truncated elements and other fragments, make up ~59% of the soybean genome (Supplementary Table 6).

Long terminal repeat (LTR) retrotransposons are the most abundant class of transposable elements. The soybean genome contains ~42% LTR retrotransposons, fewer than *Sorghum*[8] and maize[14], but higher than rice[15]. The intact element sizes range from 1 kb to 21 kb, with an average size of 8.7 kb (Supplementary Fig. 2). Of the 510 families containing 14,106 intact elements, 69% are *Gypsy*-like and the remainder *Copia*-like. However, most (~78%) of these families are present at low copy numbers, typically fewer than 10 copies. The genome also contains an estimated 18,264 solo LTRs, probably caused by homologous recombination between LTRs from a single element. Nested retrotransposons are common, with 4,552 nested insertion events identified. The copy numbers within each block range from one to six.

The genome consists of ~17% transposable elements, divided into *Tc1/Mariner*, *haT*, *Mutator*, *PIF/Harbinger*, *Pong*, *CACTA* superfamilies and *Helitrons*. Of these superfamilies, those containing more than 65 complete copies, *Tc1/Mariner* and *Pong*, comprise ~0.1% of the genome sequence, and seem to have not undergone recent amplification, indicating that they may be inactive and relatively old. Conversely, other families seem to have amplified recently and may still be active, indicated by the high similarity (>98%) of multiple elements.

Multiple whole-genome duplication events

Timing and phylogenetic position. A striking feature of the soybean genome is the extent to which blocks of duplicated genes have been retained. On the basis of previous studies that examined pairwise synonymous distance (K_s values) of paralogues[16,17], and targeted sequencing of duplicated regions within the soybean genome[18], we expected that large homologous regions would be identified in the genome. Using a pattern-matching search, gene families of sizes from two to six were identified, and K_s values were calculated for these genes,

here displayed as a histogram plot (Fig. 2), which shows two distinct peaks. Similarly, nucleotide diversity for the fourfold synonymous third-codon transversion position, 4dTv, was calculated. Both metrics give a measure of divergence between two genes, but the 4dTv uses a subset of the sites (transitions/transversion) used in the computation of K_s. 31,264 high-confidence soybean genes have recent paralogues with $K_s \approx 0.13$ synonymous substitutions per site and 4dTv ≈ 0.0566 synonymous transversions per site (Fig. 3), corresponding to a soybean-lineage-specific palaeotetraploidization. This was probably an allotetraploidy event based on chromosomal evidence[19]. Of the 46,430 high-confidence genes, 31,264 exist as paralogues and 15,166 have reverted to singletons. We infer that the pre-duplication proto-soybean genome possessed ~30,000 genes: half of $(2 \times 15{,}166 + 2 \times 15{,}632) = 30{,}798$. This number is comparable to the modern *Arabidopsis* gene complement. A second paralogue peak at $K_s \approx 0.59$ (4dTv ≈ 0.26) corresponds to the early-legume duplication, which several lines of evidence suggest occurred near the origins of the papilionoid lineage[20]. The papilionoid origin has been dated to approximately 59 Myr ago[21]. A third highly diffuse peak is seen when the plot is expanded past a K_s value of 1.5 (data not shown) and most probably corresponds to the 'gamma' event[22], shown to be a triplication in *Vitis*[23] and in other angiosperms[24].

Owing to the existence of macrofossils in the legumes and allies, the timing of clade origins in the legumes is better established than other plant families. A fossil-calibrated molecular clock for the legumes places the origin of the legume stem clade and the oldest papilionoid crown clade at 58 to 60 Myr ago[21]. If the early-legume whole-genome duplication (WGD) occurred outside the papilionoid lineage, as suggested by map evidence from *Arachis* (an early-diverging genus in the papilionoid clade)[20], then the duplication occurred within the narrow window of time between the origin of the legumes and the papilionoid radiation. If the older duplication is assumed to have occurred around 58 Myr ago, then the calculated rate of silent mutations extending back to the duplication would be 5.17×10^{-3}, similar to previous estimates of 5.2×10^{-3} (ref. 21). The *Glycine*-specific duplication is estimated to have occurred ~13 Myr ago, an age consistent with previous estimates[16,17].

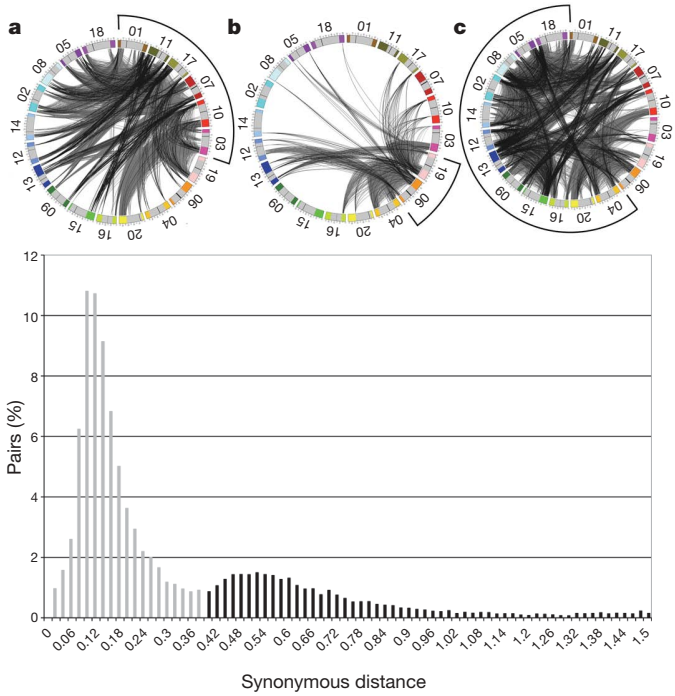


Figure 2 | Homologous relationships between the 20 soybean chromosomes. The bottom histogram plot shows pairwise K_s values for gene family sizes 2 to 6. Top panels show the 20 chromosomes in a circle with lines connecting homologous genes. Gene-rich regions (euchromatin) of each chromosome are coded a different colour around the circle. Grey represents K_s values of 0.06–0.39, 13-Myr genome duplication; black represents K_s values of 0.40–0.80, 59-Myr genome duplication. These correspond to the grey and black bars in the histogram. **a**, Chromosomes 1, 11, 17, 7, 10 and 3, which contain centromeric repeat Sb91. **b**, Chromosomes 19 and 6, which contain both Sb91 and Sb92 centromeric repeats. **c**, Chromosomes 18, 5, 8, 2, 14, 12, 13, 9, 15, 16, 20 and 4, which contain Sb92.

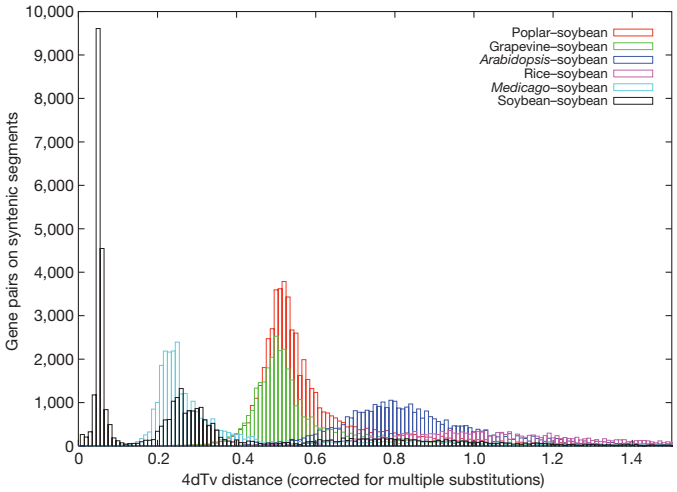


Figure 3 | Distribution of 4dTv distance between syntenically orthologous genes. Segments were found by locating blocks of BLAST hits with significance 1×10^{-18} or better with less than 10 intervening genes between such hits. The 4dTv distance between orthologous genes on these segments is reported.

Structural organization. We identified homologous blocks within the genome using i-ADHoRe[25]. Using relatively stringent parameters, 442 multiplicons (that is, duplicated segments) were identified within the soybean genome and visualized using Circos[26] (Fig. 2). Owing to the multiple rounds of duplication and diploidization in the genome, as well as chromosomal rearrangements, multiplicons (or blocks) between chromosomes can involve more than just two chromosomes. On average, 61.4% of the homologous genes are found in blocks involving only two chromosomes, only 5.63% spanning three chromosomes, and 21.53% traversing four chromosomes. Two notable exceptions to this pattern are chromosome 14, which has 11.8% of its genes retained across three chromosomes, and chromosome 20 with 7.08% of the homologues (gene pairs resulting from genome duplication) retained across four chromosomes. Chromosome 14 seems to be a highly fragmented chromosome with block matches to 14 other chromosomes, the highest number of all chromosomes. Conversely, chromosome 20 is highly homologous to the long arm of chromosome 10, with few matches elsewhere in the genome.

Retention of homologues across the genome is exceptionally high; blocks retained in two or more chromosomes can be clearly observed (Fig. 2 and Supplementary Figs 5 and 6). The number of homologues (gene pairs) within a block average 31, although any given block may contain from 6 to 736 homologues. Given that not all genes within a block are retained as homologues (owing to loss of duplicated genes over time (fractionation)), the average number of genes in a block is ~75 genes and ranges from 8 to 1,377 genes.

Repeated duplications in the soybean genome make it possible to determine rates of gene loss following each round of polyploidy. In homologous segments from the 13-Myr-old *Glycine* duplication, 43.4% of genes have matches in the corresponding region, in contrast to 25.9% in blocks from the early legume duplication. Combining these gene-loss rates with WGD dates of 13 Myr ago and 59 Myr ago, the rate of gene loss has been 4.36% of genes per Myr following the *Glycine* WGD and 1.28% of genes per Myr following the early-legume

WGD. This differential in gene-loss rates indicates an exponential decay pattern of rapid gene loss after duplication, slowing over time.

Nodulation and oil biosynthesis genes

A unique feature of legumes is their ability to establish nitrogen-fixing symbioses with soil bacteria of the family Rhizobiaceae. Therefore, information on the nodulation functions of the soybean genome is of particular interest. Sequence comparisons with previously identified nodulation genes identified 28 nodulin genes and 24 key regulatory genes, which probably represent true orthologues of known nodulation genes in other legume species (Supplementary section 3 and Supplementary Table 8). Among this list of 52 genes, 32 have at least one highly conserved homologue gene. We hypothesize that these are homologous gene pairs arising from the *Glycine* WGD (that is, ~13 Myr ago). Further analysis shows that seven soybean nodulin genes produce transcript variants. The exceptional example is nodulin-24 (Glyma14g05690), which seems to produce ten transcript variants (Supplementary Table 8). In total, 25% of the examined nodulin genes produce transcript variants, which is slightly higher than the incidence of alternative splicing in *Arabidopsis* (~21.8%) and rice (~21.2%)[27]. However, none of the soybean regulatory nodulation genes produces transcript variants (Supplementary Table 8).

Mining the soybean genome for genes governing metabolic steps in triacylglycerol biosynthesis could prove beneficial in efforts to modify soybean oil composition or content. Genomic analysis of acyl lipid biosynthesis in *Arabidopsis* revealed 614 genes involved in pathways beginning with plastid acetyl-CoA production for *de novo* fatty acid synthesis through cuticular wax deposition[28]. Comparison of these sequences to the soybean genome identified 1,127 putative orthologous and paralogous genes in soybean. This is probably a low estimate owing to the high stringency conditions used for gene mining. The distribution of these genes according to various functional classes of acyl lipid biosynthesis is shown in Table 1. Comparing *Arabidopsis* to soybean, the number of genes involved in storage lipid synthesis, fatty acid elongation and wax/cutin production was similar. For all other subclasses, the soybean genome contained substantially higher numbers of genes. Interestingly, the number of genes involved in lipid signalling, degradation of storage lipids, and membrane lipid synthesis were two- to threefold higher in soybean than *Arabidopsis*, indicating that these areas of acyl lipid synthesis are more complex in soybean. The number of genes involved in plastid *de novo* fatty acid synthesis was 63% higher in soybean compared to *Arabidopsis*. Many single-gene activities in

Table 1 | Putative acyl lipid genes in *Arabidopsis* and soybean

Function category of acyl lipid genes	Number in *Arabidopsis*	Number in soybean
Synthesis of fatty acids in plastids	46	75
Synthesis of membrane lipids in plastids	20	33
Synthesis of membrane lipids in endomembrane system	56	117
Metabolism of acyl lipids in mitochondria	29	69
Synthesis and storage of oil	19	22
Degradation of storage lipids and straight fatty acids	43	155
Lipid signalling	153	312
Fatty acid elongation and wax and cutin metabolism	73	70
Miscellaneous	175	274
Total	614	1,127

Arabidopsis are encoded by multigene families in soybean, including ketoacyl-ACP synthase II (12 copies in soybean), malonyl-CoA:ACP malonyltransferase (2 copies), enoyl-ACP reductase (5 copies), acyl-ACP thioesterase FatB (6 copies) and plastid homomeric acetyl-CoA carboxylase (3 copies). Long-chain acyl-CoA synthetases, ER acyltransferases, mitochondrial glycerol-phosphate acyltransferases, and lipoxygenases are all unusually large gene families in soybean, containing as many as 24, 21, 20 and 52 members, respectively. The multigenic nature of these and many other activities involved in acyl lipid metabolism suggests the potential for more complex transcriptional control in soybean compared to *Arabidopsis*.

Transcription factor diversity

We identified soybean transcription factor genes by sequence comparison to known transcription factor gene families, as well as by searching for known DNA-binding domains. In total, 5,671 putative soybean transcription factor genes, distributed in 63 families, were identified (Fig. 4a and Supplementary Table 9). This number represents 12.2% of the 46,430 predicted soybean protein-coding loci. A similar analysis performed on the *Arabidopsis* genome identified 2,315 putative *Arabidopsis* transcription factor genes, representing 7.1% of the 32,825 predicted *Arabidopsis* protein-coding loci (Fig. 4b). Transcription factor genes are homogeneously distributed across the chromosomes in both soybean and *Arabidopsis*, with an average relative abundance of 8–10% transcription factor genes on each chromosome. On rare occasions, regions were identified in both genomes that had a relatively low (<5%) or high density (>12%) of transcription factor genes. Among the transcription factor genes identified, 9.5% of soybean genes (538 transcription factor genes) and 8.2% of *Arabidopsis* genes (190 *Arabidopsis* transcription factor genes)

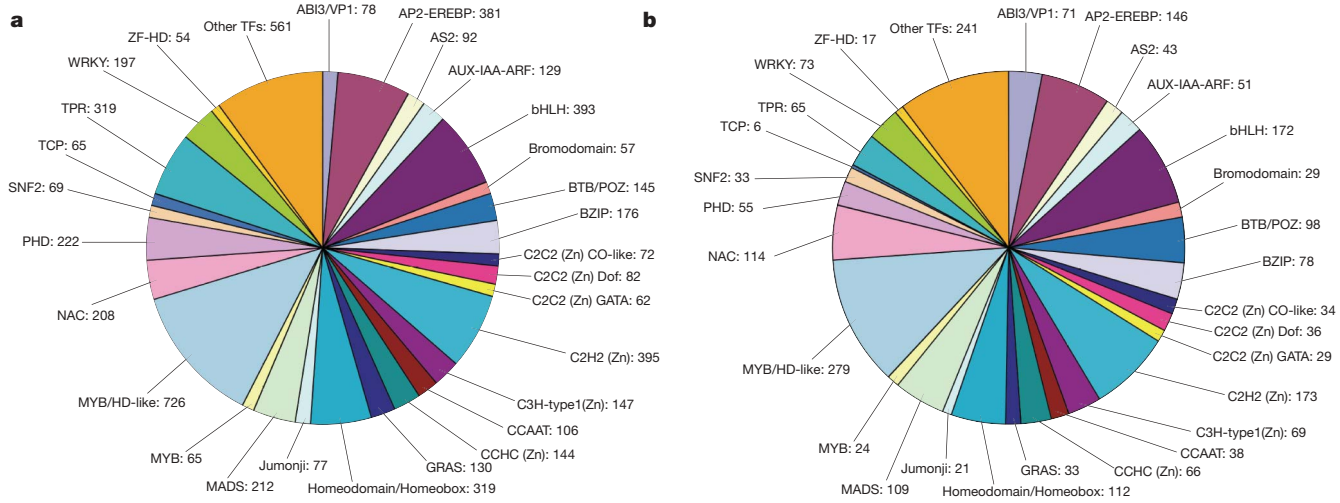


Figure 4 | Distribution of soybean (a) and *Arabidopsis* (b) transcription factor genes in different transcription factor families. Only the distribution of the most representative transcription families is detailed here. AUX-IAA-ARF, indole-3-acetic acid-auxin response factor; BTB/POZ, bric-à-brac tramtrack broad complex/pox viruses and zinc fingers; BZIP, basic leucine zipper; GRAS, (GAI, RGA, SCR); NAC, (NAM, ATAF1/2, CUC2); PHD, plant homeodomain-finger transcription factor; TCP, (TB1, CYC, PCF); TFs, transcription factors; TPR, tetratricopeptide repeat; WRKY, conserved amino acid sequence WRKYGQK at its N-terminal end.

are tandemly duplicated. By way of example, only one region in *Arabidopsis* has more than five duplicated transcription factor genes in tandem (seven ABI3/VP1 genes (At4G31610 to At4G31660)), whereas in soybean several such regions are present (for example, 13 C3H-type 1 (Zn) (Glyma15g19120 to Glyma15g19240); six MYB/HD-like (Glyma06g45520 to Glyma06g45570); and five MADS (Glyma20g27320 to Glyma20g27360); Supplementary Table 8). The overall distribution of soybean transcription factor genes among the various known protein families is very similar between *Arabidopsis* and soybean (Supplementary Fig. 10a, b). However, some families are relatively sparser or more abundant in soybean, perhaps reflecting differences in biological function. For example, members of the ABI3/VP1 family are 2.2-times more abundant in *Arabidopsis*, whereas members of the TCP family are 4.4-times more abundant in soybean. In addition, those gene families with fewer members are differentially represented between soybean and *Arabidopsis*. FHA, HD-Zip (homeodomain/leucine zipper), PLATZ, SRS and TUB transcription factor genes are more abundant in soybean (2.7, 2.9, 4.1, 3, and 4.9 times, respectively) and HTH-ARAC (helix–turn–helix araC/xylS-type) genes were identified exclusively in soybean. In contrast, HSF, HTH-FIS (helix–turn–helix-factor for inversion stimulation), TAZ and U1-type (Zn) genes are present in relatively larger numbers in *Arabidopsis* (5.4, 4.9, 24.5 and 2.9 times, respectively). Notably, both ABI3/VP1, TCP, SRS and Tubby transcription factor genes were shown to have critical roles in plant development (for example, ABI3/VP1 during seed development; TCP, SRS and Tubby affect overall plant development[29–33]). The differences seen in relative transcription factor gene abundance indicates that regulatory pathways in soybean may differ from those described in *Arabidopsis*.

Impact on agriculture

Hundreds of qualitatively inherited (single gene) traits have been characterized in soybean and many genetically mapped. However, most important crop production traits and those important to seed quality for human health, animal nutrition and biofuel production are quantitatively inherited. The regions of the genome containing DNA sequence affecting these traits are called quantitative trait loci (QTL). QTL mapping studies have been ongoing for more than 90 distinct traits of soybean including plant developmental and reproductive characters, disease resistance, seed quality and nutritional traits. In most cases, the causal functional gene or transcription factor underlying the QTL is unknown. However, the integration of the whole genome sequence with the dense genetic marker map that now exists in soybean[2–5] (http://www.Soybase.org) will allow the association of mapped phenotypic effectors with the causal DNA sequence. There are already examples where the availability of the soybean genomic sequence has accelerated these discovery efforts. Having access to the sequence allowed cloning and identification of the *rsm1* (raffinose synthase) mutation that can be used to select for low-stachyose-containing soybean lines that will improve the ability of animals and humans to digest soybeans[34]. Using a comparative genomics approach between soybean and maize, a single-base mutation was found that causes a reduction in phytate production in soybean[35]. Phytate reduction could result in a reduction of a major environmental runoff contaminant from swine and poultry waste. Perhaps most exciting for the soybean community, the first resistance gene for the devastating disease Asian soybean rust (ASR) has been cloned with the aid of the soybean genomic sequence and confirmed with viral-induced gene silencing[36]. In countries where ASR is well established, soybean yield losses due to the disease can range from 10% to 80%[36] and the development of soybean strains resistant to ASR will greatly benefit world soybean production.

Soybean, one of the most important global sources of protein and oil, is now the first legume species with a complete genome sequence. It is, therefore, a key reference for the more than 20,000 legume species, and for the remarkable evolutionary innovation of nitrogen-fixing symbiosis. This genome, with a common ancestor only 20 million years removed from many other domesticated bean species, will allow us to knit together knowledge about traits observed and mapped in all of the beans and relatives. The genome sequence is also an essential framework for vast new experimental information such as tissue-specific expression and whole-genome association data. With knowledge of this genome's billion-plus nucleotides, we approach an understanding of the plant's capacity to turn carbon dioxide, water, sunlight and elemental nitrogen and minerals into concentrated energy, protein and nutrients for human and animal use. The genome sequence opens the door to crop improvements that are needed for sustainable human and animal food production, energy production and environmental balance in agriculture worldwide.

METHODS SUMMARY

Seeds from cultivar Williams 82 were grown in a growth chamber for 2 weeks and etiolated for 5 days before harvest. A standard phenol/chloroform leaf extraction was performed. DNA was treated with RNase A and proteinase K and precipitated with ethanol.

All sequencing reads were collected with Sanger sequencing protocols on ABI 3730XL capillary sequencing machines, a majority at the Joint Genome Institute in Walnut Creek, California.

A total of 15,332,163 sequence reads were assembled using Arachne v.20071016 (ref. 37) to form 3,363 scaffolds covering 969.6 Mb of the soybean genome. The resulting assembly was integrated with the genetic and physical maps previously built for soybean and a newly constructed genetic map to produce 20 chromosome-scale scaffolds covering 937.3 Mb and an additional 1,148 unmapped scaffolds that cover 17.7 Mb of the genome.

Genes were annotated using Fgenesh+[38] and GenomeScan[39] informed by EST alignments and peptide matches to genome from *Arabidopsis*, rice and grapevine. Models were reconciled with EST alignments and UTR added using PASA[40]. Models were filtered for high confidence by penalizing genes which were transposable-element-related, had low sequence entropy, short introns, incomplete start or stop, low C-score, no UniGene hit at 1×10^{-5}, or the model was less than 30% the length of its best hit.

LTR retrotransposons were identified by the program LTR_STRUC[41], manually inspected to check structure features and classified into distinct families based on the similarities to LTR sequences. DNA transposons were identified using conserved protein domains as queries in TBLASTN[42] searches of the genome. Identified elements were used as a custom library for RepeatMasker (current version: open 3.2.8; http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) to detect missed intact elements, truncated elements and fragments.

Virtual suffix trees with six-frame translation were generated using Vmatch[43] and then clustered into families. Pairwise alignments between gene family members were performed using ClustalW[44]. Identification of homologous blocks was performed using i-ADHoRe v2.1 (ref. 25). Visualization of blocks was performed with Circos[26].

Received 19 August; accepted 12 November 2009.

1. Arumuganathan, K. & Earle, E. D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218 (1991).
2. Choi, I. Y. *et al.* A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* **176**, 685–696 (2007).
3. Hyten, D. L. *et al.* High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* (in the press).
4. Hyten, D. L. *et al.* A high density integrated genetic linkage map of soybean and the development of a 1,536 Universal Soy Linkage Panel for QTL mapping. *Crop Sci.* (in the press).
5. Song, Q. J. *et al.* A new integrated genetic linkage map of the soybean. *Theor. Appl. Genet.* **109**, 122–128 (2004).
6. Lin, J. Y. *et al.* Pericentromeric regions of soybean (*Glycine max* L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. *Genetics* **170**, 1221–1230 (2005).
7. Vahedian, M. *et al.* Genomic organization and evolution of the soybean SB92 satellite sequence. *Plant Mol. Biol.* **29**, 857–862 (1995).
8. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
9. Umezawa, T. *et al.* Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. *DNA Res.* **15**, 333–346 (2008).
10. Roy, S. W. & Penny, D. Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. *Mol. Biol. Evol.* **24**, 171–181 (2007).
11. Wang, H. *et al.* Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl Acad. Sci. USA* **106**, 3853–3858 (2009).

12. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313,** 1596–1604 (2006).
13. Michelmore, R. & Meyers, B. C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8,** 1113–1130 (1998).
14. Bruggmann, R. *et al.* Uneven chromosome contraction and expansion in the maize genome. *Genome Res.* **16,** 1241–1251 (2006).
15. Ma, J., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14,** 860–869 (2004).
16. Pfeil, B. E., Schlueter, J. A., Shoemaker, R. C. & Doyle, J. J. Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. *Syst. Biol.* **54,** 441–454 (2005).
17. Schlueter, J. A. *et al.* Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47,** 868–876 (2004).
18. Schlueter, J. A., Scheffler, B. E., Jackson, S. & Shoemaker, R. C. Fractionation of synteny in a genomic region containing tandemly duplicated genes across *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*. *J. Hered.* **99,** 390–395 (2008).
19. Gill, N. *et al.* Molecular and chromosomal evidence for allopolyploidy in soybean, *Glycine max* (L.) Merr. *Plant Physiol.* **151,** 1167–1174 (2009).
20. Bertioli, D. J. *et al.* An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. *BMC Genomics* **10,** 45 (2009).
21. Lavin, M., Herendeen, P. S. & Wojciechowski, M. F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* **54,** 575–594 (2005).
22. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422,** 433–438 (2003).
23. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).
24. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18,** 1944–1954 (2008).
25. Simillion, C., Janssens, K., Sterck, L. & Van de Peer, Y. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* **24,** 127–128 (2008).
26. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res* **19,** 1639–1645 (2009).
27. Wang, B. B. & Brendel, V. Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl Acad. Sci. USA* **103,** 7175–7180 (2006).
28. Beisson, F. *et al. Arabidopsis* genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. *Plant Physiol.* **132,** 681–697 (2003).
29. Fridborg, I., Kuusk, S., Moritz, T. & Sundberg, E. The *Arabidopsis* dwarf mutant *shi* exhibits reduced gibberellin responses conferred by overexpression of a new putative zinc finger protein. *Plant Cell* **11,** 1019–1032 (1999).
30. Barkoulas, M., Galinha, C., Grigg, S. P. & Tsiantis, M. From genes to shape: regulatory interactions in leaf development. *Curr. Opin. Plant. Biol.* **10,** 660–666 (2007).
31. Lai, C. P. *et al.* Molecular analyses of the *Arabidopsis* TUBBY-like protein gene family. *Plant Physiol.* **134,** 1586–1597 (2004).
32. Herve, C. *et al. In vivo* interference with AtTCP20 function induces severe plant growth alterations and deregulates the expression of many genes important for development. *Plant Physiol.* **149,** 1462–1477 (2009).
33. Stone, S. L. *et al.* LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. *Proc. Natl Acad. Sci. USA* **98,** 11806–11811 (2001).
34. Skoneczka, J., Saghai Maroof, M. A., Shang, C. & Buss, G. R. Identification of candidate gene mutation associated with low stachyose phenotype in soybean line PI 200508. *Crop Sci.* **49,** 247–255 (2009).
35. Saghai Maroof, M. A., Glover, N. M., Biyashev, R. M., Buss, G. R. & Grabau, E. A. Genetic basis of the low-phytate trait in the soybean line CX1834. *Crop Sci.* **49,** 69–76 (2009).
36. Meyer, J. D. F. *et al.* Identification and analyses of candidate genes for Rpp4-mediated resistance to Asian soybean rust in soybean (*Glycine max* (L.) Merr.). *Plant Physiol.* **150,** 295–307 (2009).
37. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
38. Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10,** 516–522 (2000).
39. Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11,** 803–816 (2001).
40. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31,** 5654–5666 (2003).
41. McCarthy, E. M. & McDonald, J. F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19,** 362–367 (2003).
42. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
43. Beckstette, M., Homann, R., Giegerich, R. & Kurtz, S. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* **7,** 389 (2006).
44. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Weeks for informatics support and C. Gunter for critical reading of the manuscript. We acknowledge funding from the National Science Foundation (DBI-0421620 to G.S.; DBI-0501877 and 082225 to S.A.J.) and the United Soybean Board.

Author Contributions Sequencing, assembly and integration: J. Schmutz, S.B.C., J. Schlueter, W.N., U.H., E.L., M.P., D. Grant, S.S., D. Goodstein, K.B., A.S., J.G. and D.R. Annotation: J.M., T.M., J.J.T., J.C., D.X., J.D., Z.T., L.Z., N.G., T.J., M.L., X.-C.Z. and G.S. EST sequencing: G.D.M., T.S., T.U., M.B., D.S., B.V., K.S. and H.T.N. Physical mapping: Y.Y., M.F.G., R.A.W. and R.C.S. Genetic mapping: D.H., J. Specht, Q.S. and P.C. Writing/coordination: S.A.J.

Author Information This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession ACUP00000000. The version described here is the first version, ACUP01000000. Full annotation is available at http://www.phytozome.net. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.A.J. (sjackson@purdue.edu).

Supplementary information for

Genome sequence of the palaeopolyploid soybean

Jeremy Schmutz, Steven B. Cannon, Jessica Schlueter, Jianxin Ma, Therese Mitros, William Nelson, David L. Hyten, Qijian Song, Jay J. Thelen, Jianlin Cheng, Dong Xu, Uffe Hellsten, Gregory D. May, Yeisoo Yu, Tetsuya Sakurai, Taishi Umezawa, Madan K. Bhattacharyya, Devinder Sandhu, Babu Valliyodan, Erika Lindquist, Myron Peto, David Grant, Shengqiang Shu, David Goodstein, Kerrie Barry, Montona Futrell-Griggs, Brian Abernathy, Jianchang Du, Zhixi Tian, Liucun Zhu, Navdeep Gill, Trupti Joshi, Marc Libault, Anand Sethuraman, Xue-Cheng Zhang, Kazuo Shinozaki, Henry T. Nguyen, Rod A. Wing, Perry Cregan, James Specht, Jane Grimwood, Dan Rokhsar, Gary Stacey, Randy C. Shoemaker & Scott A. Jackson

Nature 463, 178-183(14 January 2010); doi:10.1038/nature08670

Supplementary Information [pages 1–25]

This Supplement contains Genome sequencing, assembly and mapping; Annotation results; Computational analysis of nodulation genes in the soybean genome; Soybean transcription factors and Detail of QTL trait experiments bolstered by the available genome sequence. It also includes Supplementary Figures S1-S11 with Legends, Supplementary Tables S1-S4, S6-S7 and S9 (please see separate files for Tables S5 and S8) and Supplementary References.

Supplementary Table S8 [pages 29–29]

This supplemental table shows computational analysis of putative nodulation genes in soybean.

Supplementary Table S5

This is data file showing gene clusters of angiosperm and soybean. It is not printed here, but is attached to the main download page as an “Additional file”.
Filename = **Specht NATURE 2010 Suppl Table S5.txt**

SUPPLEMENTARY INFORMATION

S1. Genome Sequencing, Assembly and Mapping

Ordering, orienting, and validation of genomic WGS scaffolds to produce chromosomal pseudomolecule sequences required a combination of resources: a large consensus genetic map, a new genetic map built specifically to aid in the pseudomolecule assembly, a manually curated physical map, additional clone pairs not used in the primary assembly, genetic sequence landmarks such as centromeres and telomeres, and synteny comparisons within soybean.

S1.1. Construction of (preliminary) 6.5x scaffold assemblies, using the Arachne assembler [1]. Provisional scaffold assemblies were constructed at the stage of 6.5x WGS coverage in order to assess sequencing and assembly characteristics, and extent of marker anchoring. The 6.5x assembly consisted of 993,511,522 bases, in 3,119 scaffolds. Most of the sequence was contained in large scaffolds: the N50 was 6.5 Mb, and 97.6% of the sequence was in 364 scaffolds larger than 100 kb.

S1.2. Placement of markers from the 4.0 composite genetic map on the 6.5x scaffold assemblies, and assessment of scaffolds with insufficient or uneven marker coverage. Markers from the 4.0 consensus map [2] were placed on the scaffolds using the top e-PCR match [3] (parameters -n 3 -g1 -t3 -m 400 -d400-800, with the best-scoring match chosen in cases of multiple matches), then the top blastn hit [4]. At this stage, a total of 4,634 markers could be located on the draft scaffolds. If markers had been distributed evenly in the genomic sequence, markers would have been located approximately every 214 kb in the scaffolds. However, marker coverage in genomic sequence is uneven, with generally higher densities in euchromatic regions, where repeat densities are lower and unique primers are more easily identified. In actuality, 83 of the 6.5x scaffolds greater than 100 kb had no markers, with the largest marker-less scaffold being 1.2 Mb.

S1.3. Design of additional markers, using SNPs designed from placement of *G. soja* genomic sequence reads on the 6.5x *G. max* 'Williams 82' scaffold sequences. To provide additional marker coverage, new SNP markers were designed from short sequence reads of the *G. soja* genotype PI 468916 mapped onto the *G. max* 'Williams 82' scaffolds. The 33 bp sequence reads were generated using the Illumina Genome Analyzer sequencing platform, from a reduced representation genomic sequence library [5]. Markers were selected from 491,115 reads that mapped to non-chloroplast, non-

mitochondrial soybean genomic DNA. Read mappings were determined with the MAQ software [6]. A total of 1,536 markers were chosen from 20,119 identified SNPs, according to the following priorities (in order of decreasing importance): scaffolds > 100 kb without markers; scaffold regions without markers within 500 kbp of a scaffold end; scaffolds with less than 1 cM separation between distal markers; scaffolds with internal marker-less gaps greater than 1 Mbp. The 1,536 selected markers were used to construct an Illumina GoldenGate "oligo pool all" (SoyOPA-4), for mapping, as described in the next step.

S1.4. Construction of a genetic map in the 'Williams 82' x *G. soja* mapping population. A high-resolution genetic map was created using a mapping population of 444 recombinant inbred lines (RILs) from the cross of 'Williams 82' x *G. soja* PI 468916 (the 'W82x486' map) [5]. These were genotyped with 1804 markers: 550 of them from an earlier map, to help associate the previous and new maps; and 1,254 of them from the set of 1,536 designed on the 6.5x scaffolds.

S1.5. Construction of the 8x scaffold assemblies. Once the final data collection phase of sequencing was completed based on an estimate of additional sequence needed from the 6x assembly, a total of 15,332,163 reads (see Table S1 for clone size breakdowns) were assembled with a modified version of Arachne (Jaffe et al., 2003) v.20071016 with parameters maxcliq1=90, correct1_passes=0 and BINGE_AND_PURGE=True to form 3,363 scaffolds covering 969.6Mb of the soybean genome (see Table S2 for scaffold and contigs totals).

Library Type	Average Insert Size	Read Number	Assembled Sequence Coverage (X)
3kb	3,287	5,337,808	2.91
8kb (1)	6,547	2,626,278	1.36
8kb (2)	6,806	3,261,934	1.76
8kb (3)	8,106	1,954,318	0.96
Fosmid (1)	35,461	499,391	0.24
Fosmid (2)	36,675	991,002	0.47
Fosmid (3)	37,447	305,275	0.15
BAC (GM_WBa)	113,756	59,286	0.03
BAC (GM_WBb)	133,543	121,680	0.09
BAC (GM_WBc)	135,292	175,191	0.07
Total		15,332,163	8.05

Table S1. Genomic libraries included in the soybean genome assembly and their respective assembled sequence coverage levels in the final release.

Size	Number	Contigs	Scaffold Size	Basepairs	% Non-gap Basepairs
5,000,000	64	10,124	591,926,032	587,605,388	99.27%
2,500,000	111	12,873	756,646,961	751,670,694	99.34%
1,000,000	194	15,366	894,763,142	889,167,252	99.37%
500,000	249	16,465	934,295,162	927,815,889	99.31%
250,000	292	17,546	950,004,922	942,041,880	99.16%
100,000	368	18,923	962,641,918	951,116,174	98.80%
50,000	439	19,623	967,493,271	954,172,409	98.62%

25,000	559	20,378	972,479,545	956,358,413	98.34%
10,000	889	21,355	977,117,651	960,461,820	98.30%
5,000	1,787	23,555	983,239,271	965,409,031	98.19%
2,500	2,994	25,894	987,884,399	969,135,187	98.10%
1,000	3,363	26,341	988,480,365	969,681,045	98.10%
0	4,262	27,240	989,039,777	970,240,457	98.10%

Table S2. Summary statistics of the output of the whole genome shotgun assembly, before breaking and constructing chromosome scale pieces. The table shows total contigs and total assembled basepairs for each set of scaffolds greater than the given size.

S1.6. Placement of markers from the 4.0 composite map and the W82x486 map on the 8x scaffold assemblies. Markers from the 4.0 consensus map [2] and the W82x486 map [5] were placed on the scaffolds using the top e-PCR match [3] where possible, and otherwise using the top BLAST hit [4].

S1.7. Construction of provisional pseudomolecules by ordering and orienting (O&O) the 8x scaffold assemblies. A first approximate ordering and orienting (O&O) was made by positioning scaffolds by average cM value of marker positions in scaffolds. Initial orientations were determined by comparing cM values of first and last markers in the scaffold. Both genetic maps described above were used for this stage. An O&O based exclusively on genetic map data is subject to several kinds of problems. First, although the consensus map is remarkably large, the genetic resolution is low because the mapping populations are small. The map was constructed from five mapping populations, each having approximately 100 individuals. The W82x468 has higher resolution in some regions, but fewer markers, and apparent distortion in some chromosomal regions. Second, in pericentromeric regions, which constitute roughly half of the sequence space, there is very little recombination. In 37% of the mapped scaffolds, the maximum cM separation is less than 2 cM. Third, markers are rare in the large pericentromeres because unique sequences suitable for marker design are relatively rare in these highly repetitive regions. This means many large scaffolds have too few markers to determine an orientation, and some scaffolds have no markers in the consensus map. Fourth, outlying markers at the scaffold ends may distort the calculation of orientation. For example, a scaffold might have markers with cM values 9, 8.7, 10, 10.1, 10.3, 8.5 (in order along the scaffold). A simple automated ordering procedure

might conclude that this scaffold should be flipped, based on the first and last positions (9 and 8.5), whereas a close look at the markers would suggest that a positive orientation is more likely.

S1.8. Assessment of scaffold integrities for marker contiguity, and draft pseudomolecules for correctness of scaffold O&O. To address the deficiencies in the first-pass ordering on the consensus map, we used a combination of additional physical map resources, synteny comparisons, and genomic characteristics and landmarks.

S1.8.1. Evaluation of synteny in genome self-comparisons. Even using the higher-resolution map as well as physical map information, many scaffolds were either too small or the genomic region had too little recombination to allow precise placements or orientations. To resolve such cases, we evaluated dot plots of the soybean genome to itself, as well as genomic landmarks (ribosomal arrays, centromeric and telomeric sequences, and densities of repeats, genes, and GC content). In a dot plot of sequence matches between two chromosomes, regions with extended homology appear as diagonal features or "synteny blocks." Multiple episodes of polyploidy in the extended evolutionary history of *Glycine* mean that any given region usually matches at least one, and usually three or more other regions. Older synteny blocks are more degraded than recently derived blocks and can be obscured by repetitive DNA. To reduce noise from repetitive DNA, we used custom Perl scripts to mask all but predicted genes, and then and compared translations of the remaining genic nucleotide sequences using the promer program from the MUMmer package [7]. Further, we considered only the top reciprocal best matches between any two chromosomes in a comparison. Plots for each chromosome pair were generated using custom Perl scripts and gnuplot. For each evaluation round, we examined all plots visually (400 per genome comparison in total, or 210 unique comparisons). Potential scaffold misorientations appear as a slope sign change, with the inversion breakpoints occurring precisely at scaffold boundaries. Potential scaffold misplacements appear as horizontal or vertical translations (shifts) in part of a synteny block. Potential scaffold shifts or reorientations were checked against marker and physical map constraints before revising the O&O.

S1.8.2. Evaluation of genomic landmarks and features. Eight of the chromosomes have telomeric repeats (TTTAGGG or CCCTAAA) in both of the distal scaffolds and 11 other chromosome have telomeric repeats found on a single arm. Also, internal scaffolds in 19 of 20 chromosomes contain a

large block of characteristic 91- or 92-basepair pericentromeric repeats [8, 9]. These arrays of centromeric satellite repeats were checked for contiguity and placement within each chromosome. Other features that were considered included gene- and repeat-density gradients, as well as dinucleotide signature [10].

S1.8.3. Evaluation of additional clone pairs. Not all clone pairs are included in the Arachne assembly because they fall outside of the expected BAC clone size range, but may still be considered and used to provide supporting evidence for O&O. Using blast and stringent match parameters (require a single match at >= 99% identity and >= 99% of BAC-end length), these "extra" clone-pairs spanned and validated approximately 26% of the scaffold pairs in the assembly. In some cases, large numbers of clone pairs established the association. In others, single clones span the gap, in which case we required additional information before determining scaffold O&O.

S1.8.4. Comparison with the physical map. The soybean cv. Williams 82 physical map, a product of the NSF SoyMap project [11], comprises, as of the June 2008 version, 1,745 contigs, incorporating 141,617 BACs from primarily three libraries: GM_WBa (35,145), GM_WBb (61,379), GM_WBc (37,658), as well as 7,435 from a minimal tile of an existing FPC map of the Forrest cultivar. The Williams 82 map was constructed using the SNaPshot restriction enzyme fingerprinting technique [12] and assembled using the FPC software FingerPrinted Contigs; [13]. The FPC software program contains tools that were developed during the course of the soybean project [14] to facilitate and analyze the alignment of draft sequence to an FPC map. For the 8x assembly, a detailed analysis indicated 51 potential scaffold edits, including 32 scaffold breaks (misassemblies) and 19 scaffold merges (i.e., scaffold pairs bridged by an FPC contig).

S1.8.5. Comparison of genetic and genomic distances. A final quality control step used comparisons of genetic and physical (sequence) distances, in the form of plots of genetic vs. physical distances (Figure S1). All show similar patterns of markedly diminished recombination (flat central slope) in broad pericentromeric regions, and consistent recombination (rising slopes) at chromosome ends. In several assembly iterations, scaffold misplacements were visible as discontinuities or negative slopes.

S1.9. Splitting of chimeric scaffolds, and iterations of steps 5 through 9 until reaching a satisfactory 1.0 candidate. Evaluation of a draft O&O identified chimeric scaffolds – evident, for

example, when one end of a scaffold contained multiple markers from another linkage group. To determine the location of the misassembly, we considered clone coverage (particularly, looking for spots with little or no coverage), synteny comparisons, physical map comparisons, and other genomic gradients and features. After identifying a region in which to make a scaffold split, the Arachne assembler was run again, with constraints against particular contig joins. Over the course of five intermediate assemblies, 46 scaffolds were broken relative to the initial Arachne 8x build.

S1.10. Generation of the 1.01 pseudomolecules in Arachne. We then ordered the scaffolds with Arachne, making 322 joins to create 20 chromosome size pseudomolecules. We orientated each scaffold and joined them with 1000 N bps. We then compared the scaffolds again against the genetic map to verify the accuracy of our ordering, and reordered the scaffolds for the release according to order of the soybean genetic map. We classified the remaining scaffolds in various bins depending on sequence content. We identified contamination using megablast against Genbank NR and blastp against a set of known microbial proteins. 90 scaffolds were identified as prokaryotic contamination. We classified additional scaffolds as unanchored rDNA (143), mitochondrion (18), chloroplast (81), small unanchored repetitive scaffolds as defined by 95% of the 24mers occurring greater than four times in the large scaffolds (788). We also removed 542 scaffolds that were less than 1kb in sequence length. We appended the remaining 1,148 scaffolds putatively soybean scaffolds to the 20 chromosome scaffolds. The resulting final statistics are shown in Table S3.

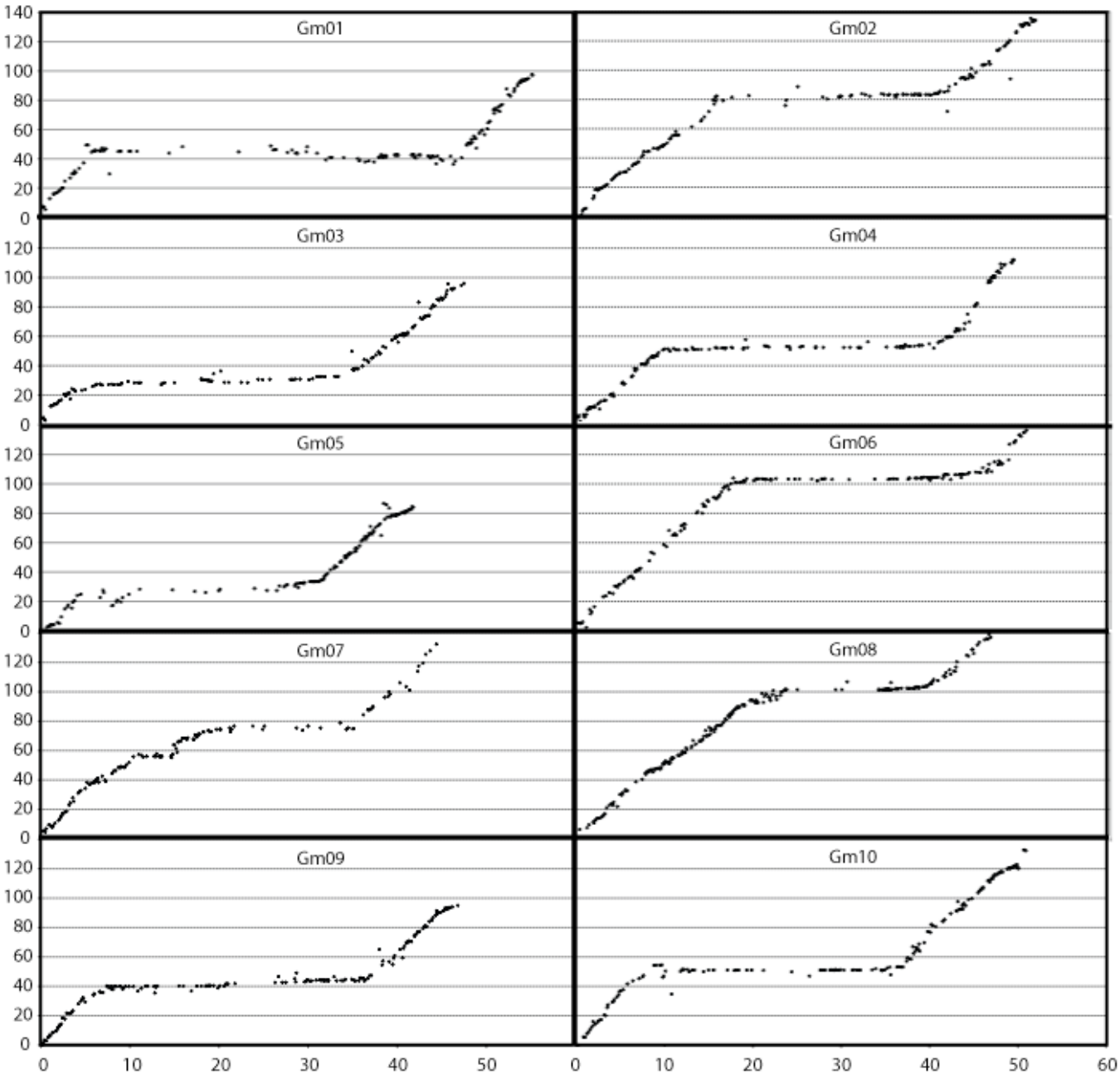
Scaffold total	1,168
Contig total	16,311
Scaffold sequence total	973.3 MB
Contig sequence total	955.1 MB (1.9% gap)
Scaffold N/L50	10/47.8 MB
Contig N/L50	1492/189.4 KB

Table S3. Final summary assembly statistics for chromosome scale assembly.

	Raw Bps (phred20s)	Assembled Bps	Coverage	Estimated Mbps at Genome Coverage
Assembled chromosomes	7,538,817,779	937,319,251	8.04	937.32
Unmapped scaffolds	98,135,309	17,735,586	5.53	12.20
Unanchored rDNA	76,472,810	1,128,931	67.74	9.51
Excluded repetitive scaffolds	28,890,682	6,380,793	4.53	3.59
Unassembled repetitive reads (>50% >=40x	787,023,270	787,023,270	1.00	97.85

24mers)				
Total sampled				1,060.47

Table S3.1 Amount of soybean genome sampled by WGS sequence.



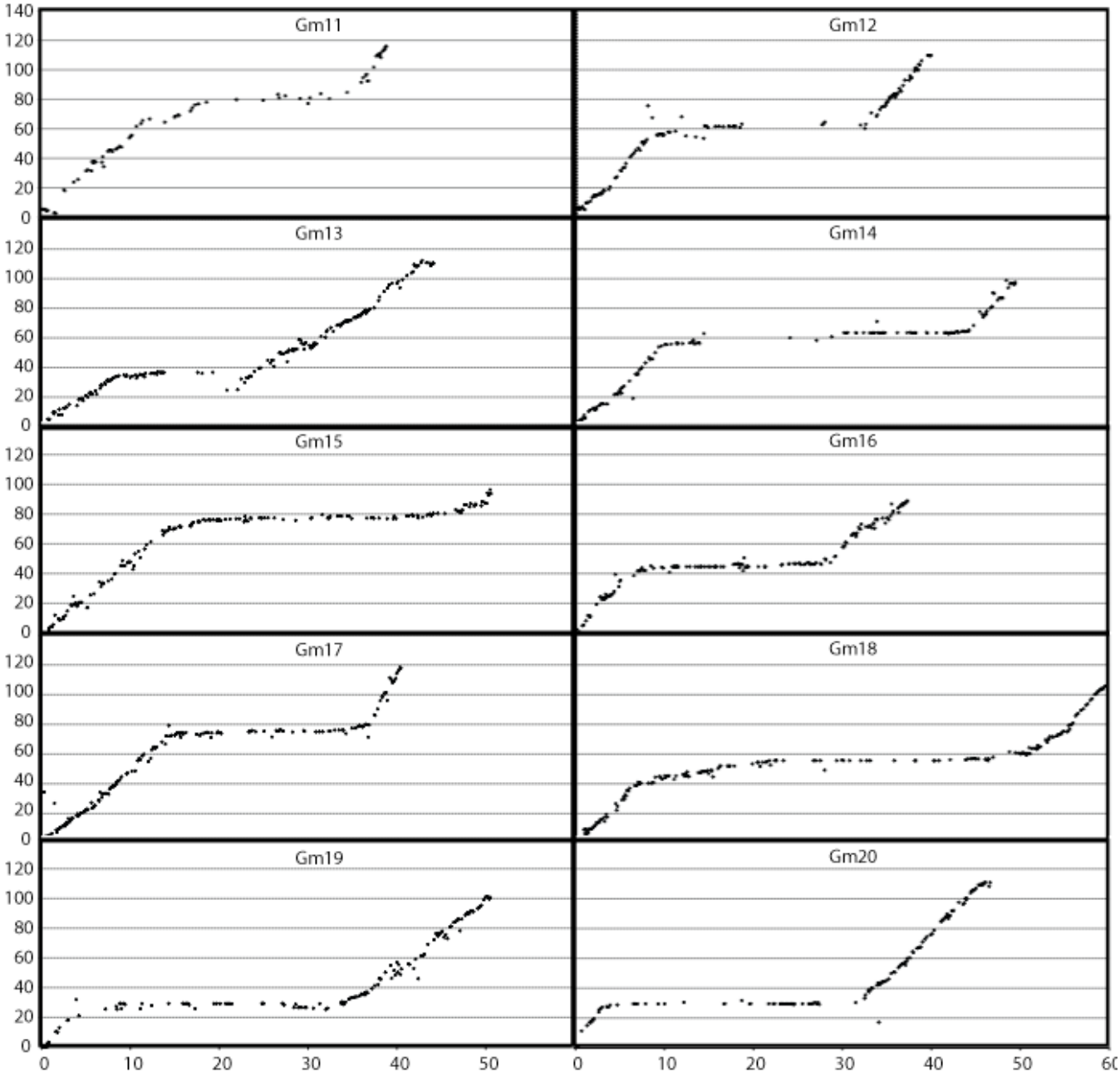


Figure S1. Plots of genetic-by-physical distances. In each plot, physical distance along the indicated pseudomolecule is on the horizontal axis (in 100 kb), and genetic distance is on the vertical axis (in cM). Dots show the locations of markers from the soybean 4.0 genetic map [2] on the sequence. There are 4,697 markers shown in these comparisons. Of these, 85% are SNP and the remainder SSR markers. The SNP markers were placed on the sequence with e-PCR [3], and SSR markers with BLAST [4]. The average genetic-to-physical ratios are approximately 193 kb/cM in the euchromatic chromosome arms, and 4.2 Mb/cM (i.e. nearly flat) in the pericentromeric regions.

1.11. Comparison to other whole genome shotgun plant genome assemblies.
Table S4. Assembly statistics from published WGS plant genomes

Genome Name	Estimated Genome Size (Mbp)	Assembled Sequence Coverage	Assembled (Mbp)	Portion Mapped (Mbp) [A]	Scaff -old N50	Scaffold L50 (Mbp)	Contig N50	Contig L50 (kb)
Soybean *(Glycine Max)*	1,115	8.04x	955.1	937.3	10	47.8	1,492	189.4
Sorghum bicolor[1,2]	818	8.50x	697.6	625.6	6	62.4	958	195.4
Poplar (*Populus Trichocarpa)* [3]	550	7.45x [B]	403.8 [B]	370.4 [B]	9 [B]	18.8 [B]	448 [B]	242.2 [B]
Physcomitrella patans[4]	511	8.92x [B]	466.7 [B]	-	86 [B]	1.7 [B]	369 [B]	291.8 [B]
Grape (*Vitus vinifera)* [5]	475	8.4x	467.5 [C]	290.2 [C]	14 [C]	13.9 [C]	2,012 [C]	66.4 [C]
Rice (*Oryza sativa*) [6]	490	5.87x	410.7 [C]	359.4 [C]	6 [C]	31.2 [C]	4,918 [C]	23.2 [C]
Papaya (*Carica papaya*)[7]	>372Mb	<3x	271.7 [C]	235	74 [C]	1.3 [C]	7,109 [C]	10.6 [C]

Notes:
A. This is the amount of sequence assigned to a distinct chromosome location; *Physcomitrella* is an unmapped release.
B. These statistics represent a newer release than that published, but one that uses the same data set as the original release and was similarly assembled to soybean.
C. These statistics were recalculated from the release to better match the ones presented for soybean, rather than presenting the numbers from the publication.

References for Table S4:

[1] Price et al. Genome evolution in the genus Sorghum (Poaceae). *Ann Bot*. **95**(1):219-27 (2005)
[2] Paterson et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551-556 (2009)
[3] Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006)
[4] Rensing *et al*. The *Physcomitrella* Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. Science **319**, 64 – 69 (2008)
[5] Jaillon, C. O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007)

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402

Gill N, Findley S, Walling JG, Hans C, Ma J, Doyle J, Stacey G, Jackson SA (2009) Molecular and chromosomal evidence for allopolyploidy in soybean, Glycine max (L.) Merr. Plant Physiol

Hyten DL, Cannon SB, Song Q, Weeks NT, Fickus EC, Shoemaker RC, Specht JE, May GD, Cregan PB (submitted) High-Throughput SNP Discovery through Deep Resequencing of a Reduced Representation Library to Anchor and Orient Scaffolds in the Soybean Whole Genome Sequence. BMC Genomics **(submitted)**

Hyten DL, Choi I-Y, Song Q, Specht JE, Carter TE, Shoemaker RC, Hwang E-Y, Matukumalli LK, Cregan PB (2009) A high density integrated genetic linkage map of soybean and the development of a 1,536 Universal Soy Linkage Panel for QTL mapping. Genetics **(submitted)**

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res **13:** 91-96

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol **5:** R12

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res **18:** 1851-1858

Lin JY, Jacobus BH, SanMiguel P, Walling JG, Yuan Y, Shoemaker RC, Young ND, Jackson SA (2005) Pericentromeric regions of soybean (Glycine max L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics **170:** 1221-1230

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics **82:** 378-389

Nelson W, Soderlund C (2009) Integrating sequence with FPC fingerprint maps. Nucleic Acids Res **37:** e36

Peto M, Shoemaker RC, Cannon SB (in preparation) Applying small-scale DNA signatures as an aid in assembling soybean chromosome sequences. Advances in Bioinformatics

Schuler GD (1997) Sequence mapping by electronic PCR. Genome Res **7:** 541-550

Shoemaker RC, Grant D, Olson T, Warren WC, Wing R, Yu Y, Kim H, Cregan P, Joseph B, Futrell-Griggs M, Nelson W, Davito J, Walker J, Wallis J, Kremitski C, Scheer D, Clifton SW, Graves T, Nguyen H, Wu X, Luo M, Dvorak J, Nelson R, Cannon S, Tomkins J, Schmutz J, Stacey G, Jackson S (2008) Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. Genome **51:** 294-302

Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. Genome Res **10:** 1772-1787

S2. Annotation Results

S2.1. Phytozome clustering method. The evolutionary based gene family is a collection of modern day genes which represent one gene in the common ancestor at that node. Families are constructed at each node on the species tree. For example, at the angiosperm node each gene family is a collection of grape, Arabidopsis, poplar, rice, and sorghum genes which are the modern descendants of one gene in

the common angiosperm ancestor. When constructing the gene families we consider two avenues for gene creation: speciation and duplication. When speciation occurs each daughter lineage receives one copy of each gene. A pair of genes, one from each lineage, coming from the same mother gene are said to be orthologs. Speciation events happen at the nodes of the species tree. Gene duplication events happen along the edges of the species tree. They result from the duplication of a gene within a species. This can either be whole-genome duplication or a local duplication. Genes related via duplication are called paralogs.

The Phytozome clustering algorithm accommodates multiple rounds of genome-wide duplications amongst the plants by using synteny to assign orthologs. The 4DTV (4-fold degenerate transversions) distance of speciation is well-defined for the angiosperm and more recent nodes. This can be used to find orthologous segments from the era of the speciation. Blast alignments of the peptide sets are performed between all species and within each species. Syntenic segments with a maximum of 10 non-aligning genes between pairs of aligning genes (E-value < 1e-18) are found within and between species. Orthologs are assigned as genes that occur on syntenic segments from the appropriate 4DTV era in which mutual-best hits account for at least 20% of hits on that segment. Mutual-best hits not in syntenic segments are also considered orthologs. Paralogs were added to these orthologous clusters by examining all hits from genes not already in clusters to these orthologous clusters. Genes are added as paralogs to their best-hitting cluster if the blast score to its putative paralog is better than the best blast score between genes already in that cluster. The clustering algorithm is hierarchical, starting at the tree tips with the modern day organisms, and marches backwards in time capturing duplications along the edges and orthologs at the nodes. It is also nested, such that any cluster of genes in recent clusters remain together in more ancient clusters.

Gene family numbers represented in this paper are the phytozome v. 4 clusterings at the angiosperm node. Legume-specific gene families refer to families that contain at least two legume (soybean or medicago) genes but no other angiosperm representatives.

S2.2. Assigning gene confidence. The initial annotation set was examined for common problems. Genes were then scored on the basis of the following common negative characteristics: contains a TE-related domain annotation, has low sequence entropy, contains an intron shorter than 40 bp, is homologous to and less than 30% the length of another gene annotation in the genome, is missing a start

or stop codon, and has a low C-score (where c-score = (blast score of hit)/(best blast score)) to reference proteomes of poplar, grapevine, and Arabidopsis.

Genes were also scored for the following positive qualities: contains an annotated domain, best hit unigene coverage is > 70%, clusters with at least three rosids represented, and is covered by an EST. Each gene was given a combined log-odds score by determining the log odds ratio assuming genes with syntenic paralogous (within soybean) or orthologous genes are true positives. The log odds scores for each parameter are reported in Supplemental Table S.7. The log odds scores were added to give a total score. After examination of the distribution of scores in Supplemental Figure S.9 a cutoff of -2 was chosen to distinguish high confidence from lower confidence genes.

Table S5. Gene clusters of angiosperm and soybean (included as separate data file).

Classification	Copy Number	DNA Content (bp)	DNA Content %
Class I :Retrotransposon	313,125	403,374,706	42.24
LTR-Retrotransposon	309,705	401,002,695	41.99
Ty1/copia	124,516	119,103,911	12.47
Intact elements	4,913	32,223,498	3.37
Solo LTRs	8,405	12,877,970	1.35
Truncated elements/fragments	111,198	74,002,443	7.75
Ty3/gypsy	185,189	281,898,784	29.52
Intact elements	9,193	97,259,046	10.18
Solo LTRs	9,859	14,510,106	1.52
Truncated elements/fragments	166,137	170,129,632	17.82
Non-LTR Retrotransposon	3,420	2,372,011	0.25
LINE	3,420	2,372,011	0.25
Class II DNA Transposon	294,937	157,551,529	16.50
Subclass I:	287,809	152,481,672	15.97
Tc1/Mariner	536	243,206	0.03
hAT	938	379,651	0.04
Mutator	100,571	43,259,871	4.53
PIF/Harbinger	10,207	2,810,474	0.29
Pong	1,755	851,918	0.09
CACTA	127,467	97,015,991	10.16
MITE	46,335	7,920,560	0.83
Tourist	19,168	3,192,179	0.33
Stowaway	27,167	4,728,381	0.50

Subclass II:	7,128	5,069,857	0.53
Helitron	7,128	5,069,857	0.53
Satellites*	11,004	11,315,839	1.18
Simple repeats	91,939	4,410,941	0.46
Low complexity	126,553	10,448,074	1.09
Total	837,558	587,101,089	61.47

*copy numbers underestimated due to the intergration of larger tandem arrays in a single unit by the Repeatmasker program.

Table S6. Repetitive composition and major components of the soybean genome.

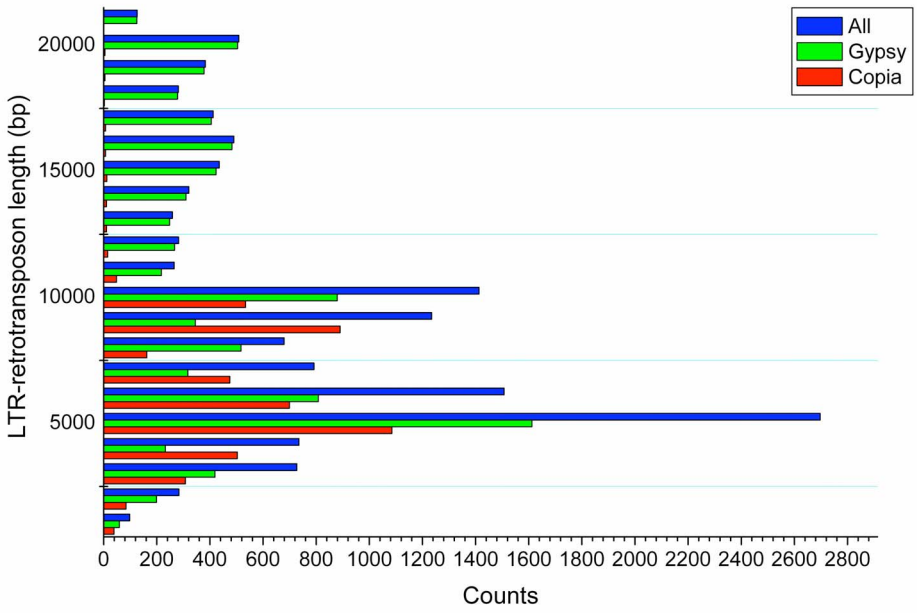


Table S.7. Log odds scores associated with presence of gene characteristics.

	TE-related	Low entropy	short intron	small copy	incomplete	low c-score	no unigene	annotated domain	good unigene coverage	high confidence cluster	est coverage
present	-1.29	-2.38	.-0.42	-3.83	-1.27	-2.87	-4.13	0.68	5.39	3.49	0.99
absent	0.09	0.03	0.09	0.39	1.19	2.25	0.15	-1.21	-2.26	-1.39	-1.20

Figure S2. LTR-retrotransposon counts in the soybean genome

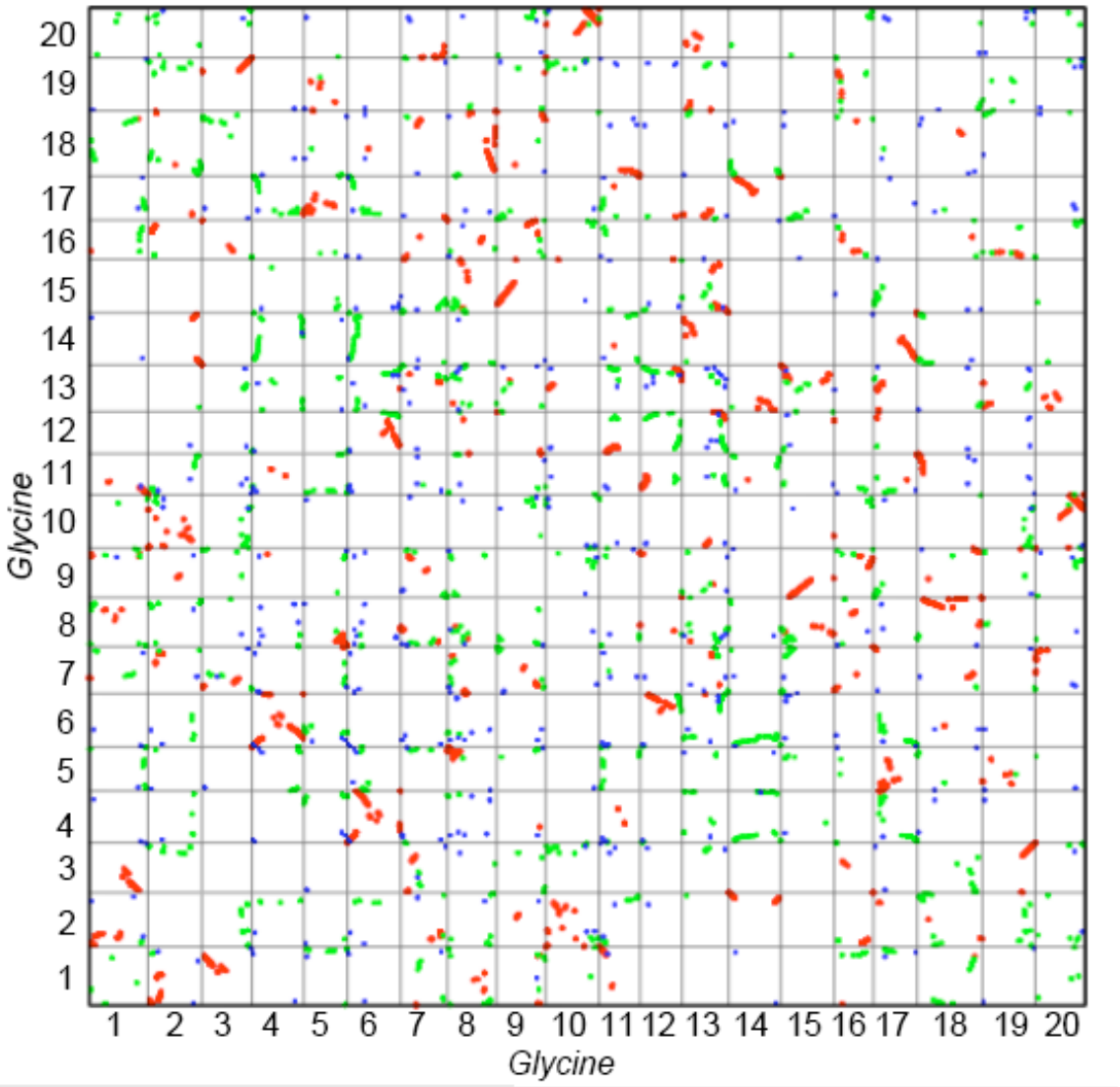


Figure S5. Syntenic paralogs from the three whole genome duplications in *Glycine max*. Red dots represent genes from the 4DTv range of 0.029-0.11, green dots from 0.11-0.3, and blue dots from 0.33-0.5. Syntenically orthologous pairs areplotted based on chromosomal position. Syntenic orthologs were calculated by finding segments for which there were a maximum of 5 non-aligning genes between aligning genes within segment pairs. Aligning genes were defined bya blast e-value of 1e-18 or less. Genes are plotted if they form segments larger than those observed in those found in a simulated randomized genome with the same numbers of genes.

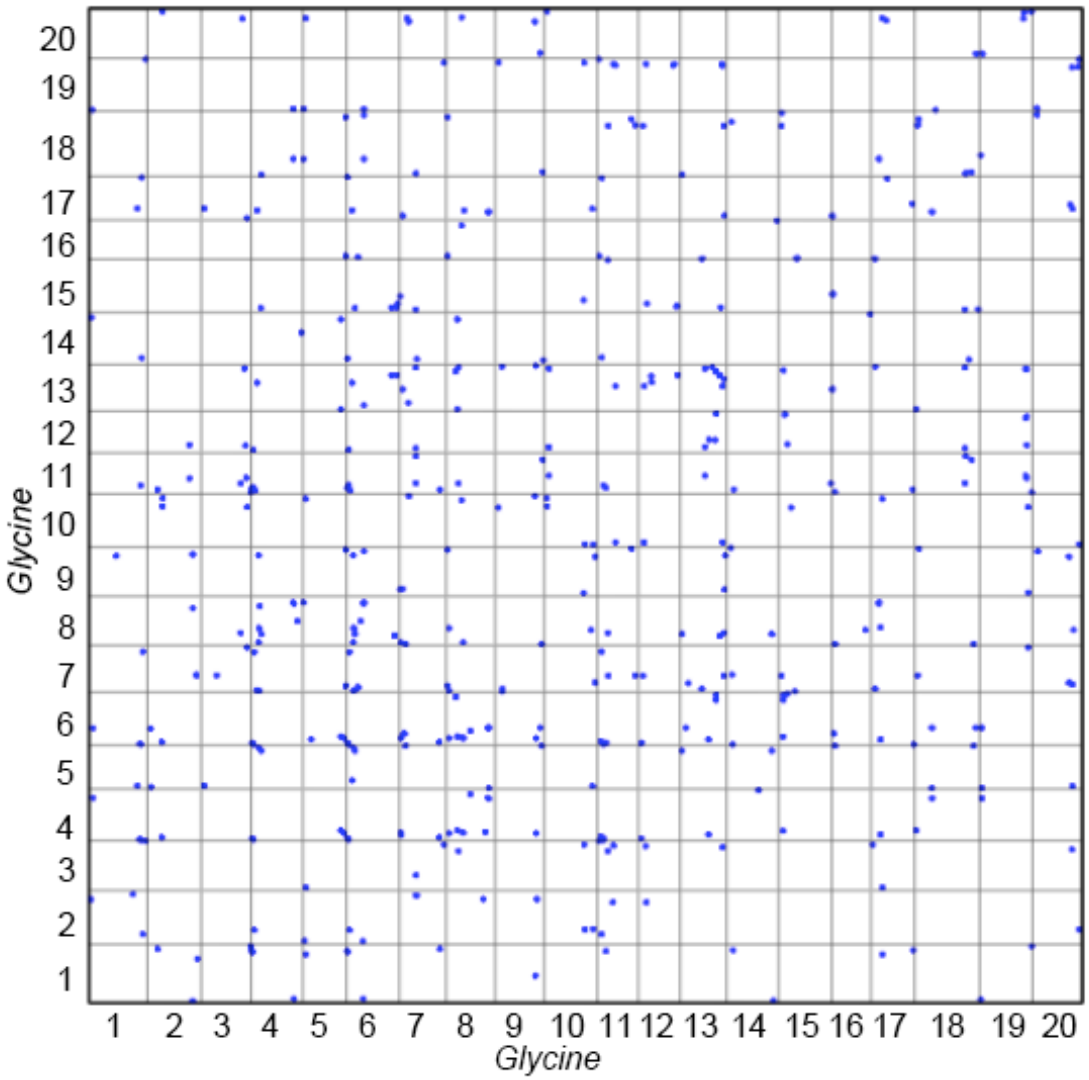


Figure S6. Syntenic paralogs from the oldest genome duplication(s) in *Glycine max*. Dots represent paralogous regions with 4DTv in the range of 0.33-0.5. These may date to the pre-rosid triplication event described in Jalion et al. (2007).

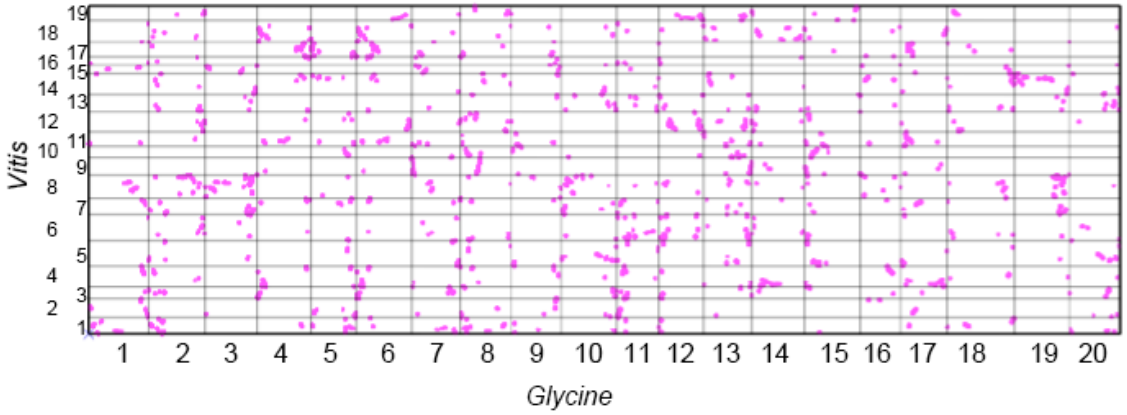


Figure S7. Syntenic orthologs between *Glycine max* and *Vitis vinifera*. Dots represent paralogous regions with

4DTv in the range of 0.36 to 0.52, characteristic of this species divergence.

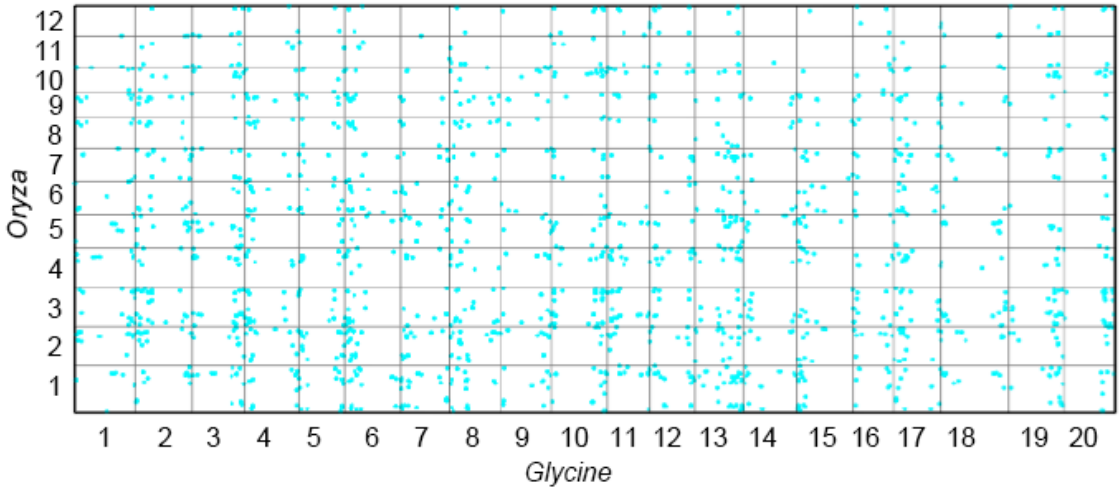


Figure S8. Syntenic orthologs between *Glycine max* and *Oryza sativa*. Dots represent paralogous regions with 4DTv in the range of 0.25 to 0.36, characteristic of this species divergence.

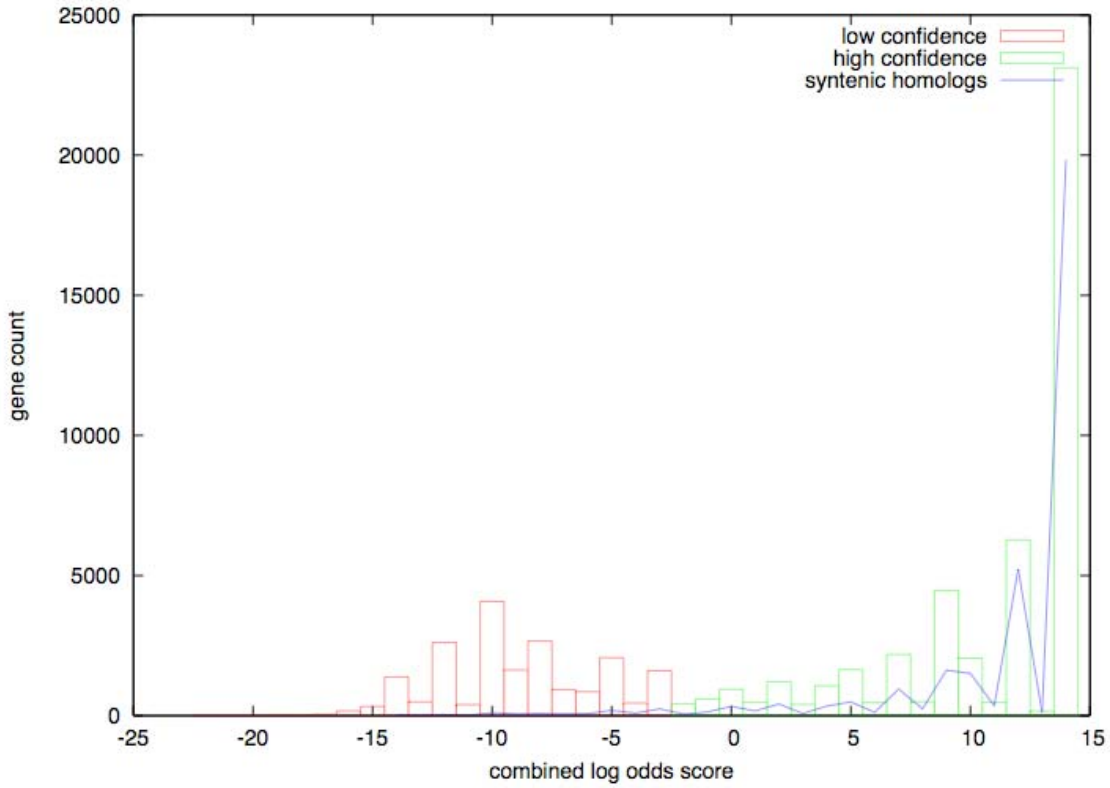


Figure S9. Combined log odds scores for all initial gene predictions.

S3. Computational analysis of nodulation genes in the soybean genome

In order to identify soybean genes involved in nodulation, we searched for soybean genes orthologous to nodulation-related genes identified through studies of other legumes. Thirty-four sequences of soybean nodulin genes were extracted from the NCBI database. A BLASTp search, with an E-value cutoff at 1e-15, identified 23 out of 34 soybean nodulins in the Glyma1 dataset. The remaining 11 soybean nodulins were not found in the Glyma1 dataset, even though the E-value cutoff was set at 1e-5. The absence of these genes is likely due to fragmentation of the query sequences or the unsaturated annotation of the soybean genome. Similarly, we extracted 18 nodulin sequences from other legumes. BLASTp (e -15) identified putative orthologs of six of these genes in the Glyma1 dataset. The remaining 12 nodulin genes, mostly representing ENOD2 and ENOD40-like sequences, are very short and did not match the Glyma1 dataset. To identify the key regulatory genes whose orthologs are required for nodulation in other legumes, we extracted 36 genes from the NCBI database, most of them encoding receptor-like kinases and transcription factors. However, some orthologous genes carry different names in different species. Therefore, for our analysis it was important to first identify a unique gene set for comparison to soybean. For example, our analyses showed that seven genes in *Lotus japonicus,* NFR1, NFR5, SYMRK, Pollux, CCaMK, NIN, and HAR1 have orthologs in *Medicago truncatula* and pea (Table S7). *Lotus* NFR5 is orthologous to *Medicago* NFP and *Pisum* SYM10[15, 16]. Taking into account these orthologous relationships, a total of 23 key regulatory, nodulation genes were identified from Lotus, Medicago and pea. BLASTp (E -50) identified putative orthologs of all 23 genes in the Glyma1 dataset.

Starting from this initial list of putative soybean nodulation gene orthologs, we tentatively defined the orthologous and paralogous relationships of these genes based on the sequence alignments, phylogenetic analyses, and non-synonymous nucleotide substitution levels (Ks). Sequences were aligned using MUSCLE3.6[17], and manually inspected using Jalview [18]. Majority-role parsimony trees were calculated using the program "protpars" in the PHYLIP package [19]. The annotations were performed by querying the protein sequences against the Pfam database (http://pfam.sanger.ac.uk/) and the InterProScan database (http://www.ebi.ac.uk/Tools/InterProScan/). This analysis led to the identification of 26 nodulin genes and 24 key regulatory genes in the Glyma1 dataset, which likely represent true orthologs of known nodulation related genes identified in other species (Table S8).

Among this list of 50 identified soybean nodulation genes, 32 genes have at least one highly conserved homolog gene (Table S7). Sequence comparisons between these gene pairs showed that peptide sequence identity ranges from 80~98% (Data not shown). Given the fact that the soybean genome underwent one round of recent homeologous duplication about 13 million years ago (Mya) [20-22], we hypothesize that these are homeologous gene pairs. Indeed, pairwise comparisons of the non-synonymous nucleotide substitution level (Ks) reveal a peak at 0.0996±0.0082 (Figure S9), suggesting the presence of a large-scale gene duplication. Assuming a rate of 6.1 synonymous substitutions per site every one billion years [23], the values estimate this duplication event at approximately 13 Mya, which is consistent with previous studies [20-22]. We also identified paralogs of 13 soybean nodulation-related genes (Table S8).

Analysis of 8 of the soybean nodulation genes suggests the presence of different transcript variants (2-11 depending on the gene). The exceptional example is Nodulin-24 (Glyma14g05690.1), which produces 10 transcript variants (Table S8). In total, 16% of the examined nodulation-related genes produce alternatively spliced transcripts, which is slightly lower than the incidence of alternative splicing in *Arabidopsis* (~21.8%) and rice (~21.2%) [24].

REFERENCES

1. Jaffe, D. B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res 13, 91-6 (2003).
2. Hyten, D. L. et al. A high density integrated genetic linkage map of soybean and the development of a 1,536 Universal Soy Linkage Panel for QTL mapping. Genetics (submitted) (2009).
3. Schuler, G. D. Sequence mapping by electronic PCR. Genome Res 7, 541-50 (1997).
4. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402 (1997).
5. Hyten, D. L. et al. High-Throughput SNP Discovery through Deep Resequencing of a Reduced Representation Library to Anchor and Orient Scaffolds in the Soybean Whole Genome Sequence. BMC Genomics (submitted) (submitted).
6. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18, 1851-8 (2008).
7. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol 5, R12 (2004).
8. Gill, N. et al. Molecular and chromosomal evidence for allopolyploidy in soybean, Glycine max (L.) Merr. Plant Physiol (2009).

9. Lin, J. Y. et al. Pericentromeric regions of soybean (Glycine max L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics 170, 1221-30 (2005).
10. Peto, M., Shoemaker, R. C. & Cannon, S. B. Applying small-scale DNA signatures as an aid in assembling soybean chromosome sequences. Advances in Bioinformatics (in preparation).
11. Shoemaker, R. C. et al. Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. Genome 51, 294-302 (2008).
12. Luo, M. C. et al. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82, 378-89 (2003).
13. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res 10, 1772-87 (2000).
14. Nelson, W. & Soderlund, C. Integrating sequence with FPC fingerprint maps. Nucleic Acids Res 37, e36 (2009).
15. Arrighi, J. et al. The *Medicago truncatula* lysin motif-receptor-like kinase gene family includes NFP and new nodule-expressed genes. Plant Physiol. 142, 265-279 (2006).
16. Madsen, E. B. et al. A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals. Nature 425, 637-640 (2003).
17. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids Res. 32, 1792-1797 (2004).
18. Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. Bioinformatics 20, 426-427 (2004).
19. Felsenstein, J. PHYLIP (Phylogeny Inference Package) (**Distributed by the Author**. Seattle:Department of Genetics, University of Washington, 2000).
20. Schlueter, J. A. et al. Mining EST databases to resolve evolutionary events in major crop species. Genome 47, 868-876 (2004).
21. Pfeil, B. E., Schlueter, J. A., Shoemaker, R. C. & Doyle, J. J. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. System. Biol. 54, 441-454 (2005).
22. Lavin, M., Herendeen, P. S. & Wojciechowski, M. F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. System. Biol. 54, 575-94 (2005).
23. Lynch, M. & Connery, J. S. The evolutionary fate and consequences of duplicate genes. Science 209, 1151-1155 (2000).
24. Wang, B. B. & Brendel, V. Genomewide comparative analysis of alternative splicing in plants. Proc. Natil. Aca. Sci. USA 103, 7175-7180 (2006).
25. Choi, I. Y. et al. A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176, 685-96 (2007).
26. Song, Q. J. et al. A new integrated genetic linkage map of the soybean. Theor Appl Genet 109, 122-8 (2004).
27. Skoneczka, J., Saghai Maroof, M. A., Shang, C. & Buss, G. R. Identification of candidate gene mutation associated with low stachyose phenotype in soybean line PI 200508. Crop Sci. 49, in press (2009).
28. Saghai Maroof, M. A., Glover, N. M., Biyashev, R. M., Buss, G. R. & Grabau, E. A. Genetic basis of the phytate trait in the soybean line CX1834. Crop Sci. 49, in press (2009).
29. Connaly, E. L. & Guerinot, M. L. Iron stress in plants. Genome Biol 3, 1024.1-1024.4 (2002).

30. Schmidt, W. Iron solutions: acquisition strategies and signaling pathways in plants. Trends Plant Sci 8, 188-93 (2003).
31. Lin, S., Baumer, J. S., Ivers, D., Cianzio, S. & Shoemaker, R. C. Field and nutrient solution tests measure similar mechanisms controlling iron deficiency chlorosis in soybean. Crop Sci. 38, 254-259 (1998).
32. Lin, S., Cianzio, S. & Shoemaker, R. C. Mapping genetic loci for iron deficiency chlorosis in soybean. Mol Breeding 3, 219-229 (1997).
33. Bauer, P., Ling, H. Q. & Guerinot, M. L. FIT, the FER-LIKE IRON DEFICIENCY INDUCED TRANSCRIPTION FACTOR in Arabidopsis. Plant Physiol Biochem 45, 260-1 (2007).
34. Vert, G. et al. IRT1, an Arabidopsis transporter essential for iron uptake from the soil and for plant growth. Plant Cell 14, 1223-33 (2002).

Table S8. Computational analysis of putative nodulation genes in soybean (included as separate file)

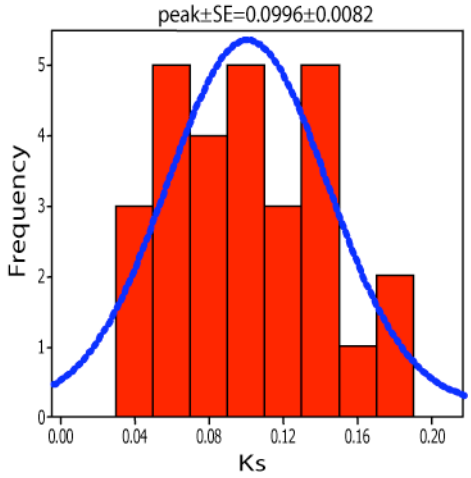


Figure S10. Histogram plots of pairwise synonymous distance of nodulation related genes in soybean. Y-axis denotes the frequency and x-axis denotes the synonymous distance (Ks). The blue line indicates the presence of one round of large-scale gene duplications.

S4. Soybean transcription factors

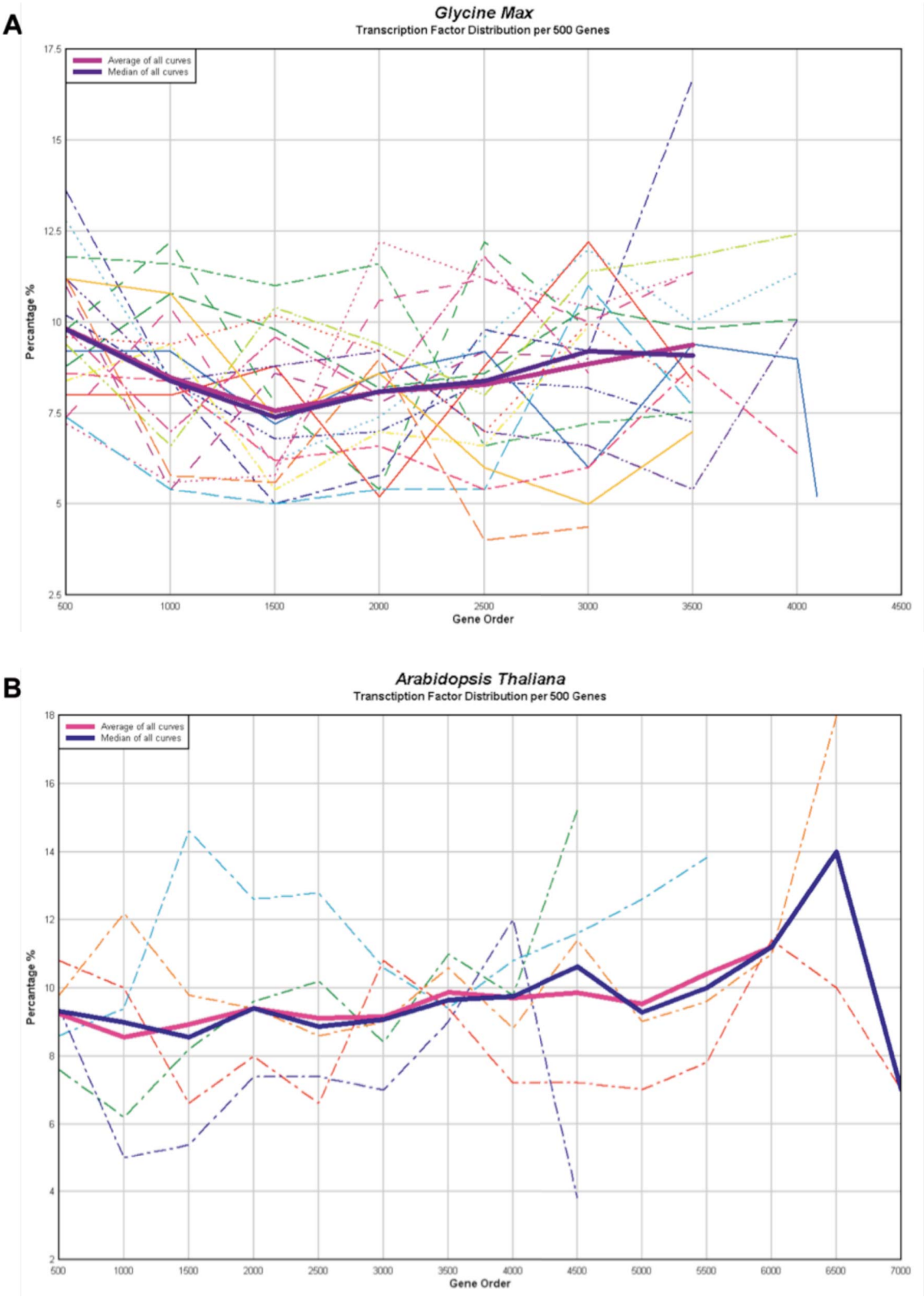


Figure S11. Transcription factor genes density compared to total soybean (A) and *Arabidopsis* (B) genes density. The percentage of TF genes is shown for each of 20 soybean and 5 *Arabidopsis*

chromosomes (thin lines). The average and mean densitfy of TF genes between chromosomes are indicated for both species (bold lines).

	Soybean		*A. thaliana*	
	number	%	number	%
ABI3/VP1	78	1.4%	71	3.1%
AP2-EREBP	381	7%	146	6%
AS2	92	2%	43	2%
AUX-IAA-ARF	129	2%	51	2%
bHLH	393	7%	172	7%
BROMODOMAIN	57	1%	29	1%
BTB/POZ	145	3%	98	4%
BZIP	176	3%	78	3%
C2C2 (Zn) CO-like	72	1%	34	1%
C2C2 (Zn) Dof	82	1%	36	2%
C2C2 (Zn) GATA	62	1%	29	1%
C_2H_2 (Zn)	395	7%	173	7%
C_3H-type1(Zn)	147	3%	69	3%
CCAAT	106	2%	38	2%
CCHC (Zn)	144	3%	66	3%
GRAS	130	2%	33	1%
Homeodomain/HOMEOBOX	319	6%	112	5%
JUMONJI	77	1%	21	1%
MADS	212	4%	109	5%
MYB	65	1%	24	1%
MYB/HD-like	726	13%	279	12%
NAC	208	4%	114	5%
PHD	222	4%	55	2%
SNF2	69	1%	33	1%
TCP	65	1.1%	6	0.30%
TPR	319	6%	65	3%
WRKY	197	3%	73	3%
ZF-HD	54	1%	17	1%
others	561	10%	241	10%
TOTAL	**5683**		**2315**	

Table S9. List of soybean and *Arabidopsis* TF genes. TF gene distribution among the different TF families and identification of tandemly duplicated TF genes both species.

S5. Detail of QTL trait experiments bolstered by the available genome sequence

QTL mapping studies have been ongoing for more than 90 distinct traits of soybean. These traits include plant developmental and reproductive characters, disease resistance, seed quality, and nutritional traits. In most cases, the causal functional gene or transcription factor underlying the QTL still remains a mystery. However, the integration of the whole genome sequence with the dense genetic marker map

that now exists in soybean [25, 26] (http://www.soybase.org) will allow the association of mapped phenotypic effectors with the causal DNA sequence. There are already examples where the availability of the soybean genomic sequence has accelerated these discovery efforts to improve soybean as a crop species.

In one case, the genome sequence allowed the positioning of a low stachyose locus to the whole genome assembly which permitted the cloning and identification of the locus termed *rsm*1 (raffinose synthase mutation). Two soybean seed carbohydrate components, raffinose and stachyose, are difficult for animals and humans to digest, and can cause off-flavors, gastric discomfort, and flatulence. Skoneczka et al. [27] mapped low seed stachyose, a qualitatively inherited trait present in accession PI 200508 (i.e., the recessive *sty1a*) to a 12 cM interval between publicly available microsatellite markers. The authors used the sequences of the markers flanking the locus to identify a segment within the genomic sequence potentially containing the causal gene. Fine mapping and analysis of the predicted genes in the mapped interval identified one gene with homology to galactosyltransferase, which seemed to be the lone candidate gene of 66 possibilities. Candidate gene sequence comparison between three wild-type seed stachose levels (two parental lines, plus Williams 82) versus the low stachyose parental line, revealed a sequence polymorphism (i.e., a 3 bp deletion) found only in the low stachyose line. The authors then developed a gene-specific microsatellite marker that can be conveniently used for marker-assisted selection of low stachyose lines.

In another example, Saghai-Maroof et al. [28] used the Williams 82 WGS to search for gene candidates underlying two QTL for low seed phytate in the germplasm line CX1834, one of which mapped to chromosome 19 and the other to chromosome 12. Phytate is a major storage form of phosphorus in the soybean seed, but is poorly digested by non-ruminants (i.e., swine and poultry). The phytate excreted in the manure of non-ruminants eventually reaches wetlands, rivers, and lakes, which causes serious environmental consequences. The authors, who were aware that a defective MRP ABC transporter was the candidate gene for the maize low phytate mutant *lpa1,* used the maize MRP4 gene to search against the soybean WGS and obtained positive matches on soybean chromosome 12 and 19, both tightly linked to phytate QTL. Comparative sequencing of the MRP homolog in CX1834 and in Williams 82 revealed a single base mutation from A (wild-type) to T (mutant), resulting in a stop codon (a truncated protein product). In yet another example, the first resistance gene for the devastating disease Asian Soybean

Rust (ASR) has been cloned with the aid of the soybean genomic sequence and confirmed with viral induced gene silencing (Meyer et al. 2009). In countries where ASR is well established, soybean yield losses due to the disease can range from 10 to 80 percent [35-37] and the development of soybean strains resistant to ASR will greatly benefit the world.

As a final example, it is known that iron is essential for plant function and plays important roles in electron transport chains of photosynthesis and respiration [29]. Iron is abundant in the earth mantle yet remains in a largely inaccessible form for plants to utilize. Nearly two-thirds of the world's population is at risk for iron-induced anemia and the major source of iron for humans is through plants [30]. Iron-deficient chlorosis (IDC) causes yield reduction in many crop plants and is a complex quantitative trait whose manipulation has been recalcitrant for breeders. Several QTL for IDC have been mapped in soybean populations [31, 32]. Although many of the genes involved in the mechanisms of iron acquisition, uptake and transport have been identified [30], the causal gene(s) underlying the phenotypes associated with each IDC QTL remain largely unknown. Searches of the WGS assembly have generated some noteworthy possibilties: A FIT homolog (FER-like Iron Deficiency Induced Transcription Factor; [33] lies within an iron QTL on chromosome 12 and an IRT [Iron Transporter[34]] homolog lies within a QTL on chromosome 14 (unpublished). Whether these candidates are causal to the trait locus remains to be proven, however with the availability of the soybean genome sequence the flood-gates are open to uncover nearly limitless possibilities for genetic discovery and soybean crop improvement.

Table 1. Computational analysis of putative nodulation genes in soybean.

Glyma1 loci	Gm genes	Lotus japonicus alias	Medicago truncatula alias	Pisum sativum alias	Protein Annotations	Putative homeologues	Putative paralogues
Glyma13g44100.1	NodE27(X03979)[a] (*1*)[b]				nodulin		
Glyma10g23790.1	Nod35(D86930) (*2*)				uricase	Glyma20g17440.1[c]	
Glyma02g36580.1	Nod55-1(CAA48908) (*3*)		ENOD16(X99466) (*4*)		Plastocyanin-like domain; B3 DNA binding	Glyma17g08110.1	
Glyma17g08110.1	Nod55-2(X69157) (*3*)		ENOd20(X99467) (*4*)		Plastocyanin-like domain; B3 DNA binding	Glyma02g36580.1	
Glyma13g17420.2 Glyma13g17420.1	Nod100(AF030231) (*5*)				sucrose synthase; Glycosyl transferases	Glyma15g20180.1 Glyma15g20180.2 Glyma09g08550.3	
Glyma07g33090.1	SAN1A(DQ418880) (*6*)				2OG-Fe(II) oxygenase	Glyma02g15370.1	Glyma02g15360.1[d] Glyma02g15380.1 Glyma02g15390.1 Glyma02g15400.1
Glyma07g33070.1	SAN1B(DQ418879) (*6*)				2OG-Fe(II) oxygenase	Glyma07g33090.1	Glyma02g15360.1 Glyma02g15380.1 Glyma02g15390.1 Glyma02g15400.1
Glyma01g03470.1	N36a(D13503) (*7*)				N/A		
Glyma13g12500.1	N56(D38015) (*8*)				isopropylmalate synthase and homocitrate synthase	Glyma13g12440.1	Glyma13g12490.1
Glyma18g02230.1 Glyma18g02230.2	N70(D13505) (*7*)				Sulfate transporter family		
Glyma06g24760.1	N93(D13506) (*7*)				nodulin	Glyma05g08400.1	Glyma05g08380.1 Glyma17g12610.1
Glyma17g08110.1	N315(D13502) (*7*)				Plastocyanin-like domain; B3 DNA binding domain		
Glyma02g43320.1	Nodulin-16(X54307) (*9*)				Glycine rich protein family		Glyma02g43330.1 Glyma02g43300.1
Glyma13g40400.2 Glyma13g40400.1	Nodulin-20(X05020) (*10*)				N/A		Glyma15g05010.1
Glyma05g25010.1	Nodulin-21(X16488) (*11*)				Integral membrane protein DUF125	Glyma08g08120.1	Glyma08g08090.1 Glyma08g08110.1 Glyma05g24970.1 Glyma05g24980.1 Glyma05g24990.1 Glyma05g25000.1
Glyma15g05010.1	Nodulin-22(X05024) (*10*)				N/A	Glyma13g40400.1	
Glyma14g05690.1 Glyma14g05690.2 Glyma14g05690.3 Glyma14g05690.4 Glyma14g05690.5 Glyma14g05690.6 Glyma14g05690.7 Glyma14g05690.8 Glyma14g05690.9 Glyma14g05690.10	Nodulin-24(M10595) (*12*)				nodulin		
Glyma08g12650.1	Nodulin-26a(X04782) (*13*)				Major intrinsic protein; aquaporins	Glyma05g29510.1	
Glyma19g22210.1 Glyma19g22210.2 Glyma19g22210.3	Nodulin-26b(CAA287430) (*14*)				Major intrinsic protein; aquaporins		
Glyma10g39450.1	Nodulin-33(AJ518837) (*15*)				haloacid dehalogenase-like hydrolase	Glyma20g28320.1 Glyma20g28320.2	
Glyma10g06810.1	Nodulin-61(AF434718) (*16*)		N6(AJ133118) (*17*)		Amidohydrolase		Glyma10g06820.1

Glyma06g06930.2 Glyma06g06930.1	Nodulin(AF065435)				SPFH/Band 7 family; integral membrane protein		Glyma06g06920.1
Glyma13g40400.2 Glyma13g40400.1	Nodulin(X57247) (*18*)				nodulin		
Glyma14g23780.1			ENOD8(AF064775) (*19*)		GDSL-like Lipase/Acylhydrolase	Glyma13g03300.1	
Glyma17g03340.1			N13(Y10455) (*20*)		Pathogenesis-related protein Bet v I family		
Glyma15g00620.1			Nod26(AY605123) (*21*)		Major intrinsic protein	Glyma08g23230.1	
Glyma14g38170.1			RIP1(MTU16727) (*22*)		peroxidase	Glyma02g40020.1	
Glyma02g43860.1		NFR1(AJ575248) (*23*)	LYK3(AY372402) (*24*)/HCL (*25*)	SYM37(EU564102) (*26*)	LysM RLK	Glyma14g05060.1	
Glyma11g06740.1		NFR5(AJ575255) (*27*)	NFP(DQ496250) (*28*)	SYM10(AJ575253) (*27*)	LysM RLK	Glyma01g38560.1	
Glyma09g33510.1	GmNORK(DQ341398) (*29*)	SYMRK(AF492655) (*30*)	DMI2(AJ418369) (*31*)	SYM19(AF491997) (*31*)	LRR RLK	Glyma01g02450.1	
Glyma12g28860.1		pollux(BAD89020) (*32*)	DMI1(DQ341395) (*33*)	SYM8(EF447277) (*34*)	ion-channel protein	Glyma16g00500.1	
Glyma19g45310.1		castor(BAD89019) (*32*)			ion-channel protein		Glyma12g28860.1[d] Glyma16g00500.1
Glyma15g35070.1		CCaMK(AM230793) (*35*)	CCaMK[DMI3](AY496049) (*36*)	SYM9(AY502067) (*37*)	calcium calmodulin binding kinase	Glyma08g24360.1	
Glyma07g04430.1		NSP1(ABK35066) (*38*)			GRAS TF	Glyma16g01020.1	
Glyma04g43090.1		NSP2(ABG49438) (*39*)			GRAS TF	Glyma13g02840.1	
Glyma04g00210.1		NIN(AJ238956) (*40*)		SYM35(AJ493066) (*41*)	RWP-RK/PB1; putative TF	Glyma06g00240.1	
Glyma12g05390.1		NIN2(CAE30325) (*42*)			RWP-RK/PB1; putative TF	Glyma11g13390.1	Glyma13g42160.1 Glyma15g03220.1
Glyma12g04390.1	GmNARK(AY166655) (*43*)	HAR1(AJ580824) (*44*)	SUNN(AY769943)	SYM29(EU270068) (*45*)	LRR RLK	Glyma11g12190.1	
Glyma10g06610.1			SKL1(EU709495) (*46*)		AtEIN2-like	Glyma13g20810.1 Glyma13g20810.2	Glyma03g33850.1
Glyma18g14750.1		Astray(AB092677) (*47*)			bZIP TF	Glyma08g41450.1	
Glyma17g14220.1			CCS52(AF134835) (*48*)		WD40 protein	Glyma05g03710.1	
Glyma01g35250.1[e] Glyma01g35260.1		Cyclops(EF569221) (*49*)			bZIP TF	Glyma09g34690.1 Glyma09g34700.1	
Glyma08g05370.1		HK1(DQ848999) (*50*)			CHASE domain His kinase A	Glyma05g34310.1	Glyma02g09550.1 Glyma07g27540.1
Glyma01g36080.1			HMGR1(EU302813) (*51*)		Hydroxymethylglutaryl-coenzyme A reductase		
Glyma11g09330.1			HMGR2(EU302814) (*51*)		Hydroxymethylglutaryl-coenzyme A reductase		
Glyma11g09330.2 Glyma11g09330.1			HMGR3(EU302815) (*51*)		Hydroxymethylglutaryl-coenzyme A reductase		
Glyma16g21620.1			HMGR4(EU302816) (*51*)		Hydroxymethylglutaryl-coenzyme A reductase		
Glyma20g05530.1			HMGR5(EU302817) (*51*)		Hydroxymethylglutaryl-coenzyme A reductase	Glyma02g44070.1	
Glyma14g01160.1		NUP133(AJ890251) (*52*)			nucleoporin	Glyma02g47530.1	
Glyma17g27490.1		NUP85(AB284835) (*53*)			nucleoporin		
Glyma loci	Gm genes	Lotus japonicus alias	Medicago truncatula alias	Pisum alias	protein annotations		

[a]Denotes the GenBank accession numbers.
[b]References are listed below this table.
[c]Genes shown in red represent the putative homeologues.
[d]Genes shown in blue represent putative paralogues.
[e]These tow gene models need to merge to match the Lotus Cyclops genes. The same is true for it's homeologous.

REFERENCES:

1. C. Sengupta-Gopalan, J. W. Pitas, D. V. Thompson, L. M. Hoffman, *Mol. Gen. Genet.* **203**, 410 (1986).
2. T. Nguyen, M. Zelechowska, V. Foster, H. Bergmann, D. P. Verma, *Proc. Natl. Acad. Sci. USA* **82**, 5040 (1985).
3. C. de Blank *et al.*, *Plant Mio. Biol.* **22**, 1167 (1993).
4. E. A. Greene, M. Erard, A. Dedieu, D. G. Barker, *Plant Mol. Biol.* **36**, 775 (1998).
5. X.-Q. Zhang *et al.*, *Plant Physiol.* **115**, 1729 (1997).
6. C. J. Webb, C. Chan-Weiher, D. A. Johnson, *Plant Physiol.* **165**, 1736 (2008).
7. H. Kouchi, S. Hata, *Mol. Gen. Genet.* **238**, 106 (1993).
8. H. Kouchi, S. Hata, *Mol. Plant Microbe Interact.* **8**, 172 (1995).
9. W. Nirunsuksiri, C. Sengupta-Gopalan, *Plant Mio. Biol.* **15**, 835 (1990).
10. N. N. Sandal, K. Bojsen, K. A. Marcker, *Nucleic acids Res.* **15**, 1507 (1987).
11. A. J. Delauney, C. I. Cheon, P. J. Snyder, D. P. Verma, *Plant Mio. Biol.* **14**, 449 (1990).
12. P. Katinakis, D. Verma, *Proc. Natil. Aca. Sci. USA* **82**, 4157 (1985).
13. M. G. Fortin, N. A. Morrison, D. P. Verma, *Nucleic acids Res.* **15**, 813 (1987).
14. F. A. Jacobs, M. Zhang, M. G. Fortin, D. P. Verma, *Nucleic acids Res.* **15**, 1271 (1987).
15. A. Roussis *et al.*, *Plant Physiol. Biochem* **41**, 719 (2003).
16. B. Trevaskis, M. Wandrey, G. Colebatch, M. K. Udvardi, *Mol. Plant Microbe Interact.* **15**, 630 (2002).
17. R. Mathis, C. Grosjean, F. de Billy, T. Huguet, P. Gamas, *Mol. Plant Microbe Interact.* **12**, 544 (1999).
18. S. G. Gottlob-Mchugh, D. A. Johnson, *Can. J. Bot.* **69**, 2263 (1991).
19. C. Liu, A. T. Yeung, R. Dickstein, *Plant Physiol.* **117**, 1127 (1998).
20. P. Gamas, F. de Billy, G. Truchet, *Mol. Plant Microbe Interact.* **11**, 393 (1998).
21. N. Uehlein *et al.*, *Phytochem.* **68**, 122 (2007).
22. D. Cook *et al.*, *Plant Cell* **7**, 43 (1995).
23. S. Radutoiu *et al.*, *Nature* **425**, 585 (2003).
24. E. Limpens *et al.*, *Science* **302**, 630 (2003).
25. P. Smit *et al.*, *Plant Physiol.* **145**, 183 (2007).
26. V. Zhukov *et al.*, *Mol. Plant Microbe Interact.* **21**, 1600 (2008).
27. E. B. Madsen *et al.*, *Nature* **425**, 637 (2003).
28. J. Arrighi *et al.*, *Plant Physiol.* **142**, 265 (2006).
29. H. Zhu, B. K. Riely, N. J. Burns, J. M. Ane, *Genetics* **172**, 2491 (2006).
30. S. Stracke *et al.*, *Nature* **417**, 959 (2002).
31. G. Endre *et al.*, *Nature* **417**, 962 (2002).
32. H. Imaizumi-Anraku *et al.*, *Nature* **433**, 527 (2005).
33. J. M. Ane *et al.*, *Science* **303**, 1364 (2004).
34. A. Edwards, A. B. Heckmann, F. Yousafzai, G. Duc, J. A. Downie, *Mol. Plant Microbe Interact.* **20**, 1183 (2007).
35. L. Tirichine *et al.*, *Nature* **441**, 1153 (2006).
36. R. M. Mitra *et al.*, *Proc. Natl. Acad. Sci. USA* **101**, 4701 (2004).
37. J. Levy *et al.*, *Science* **303**, 1361 (2004).
38. A. B. Heckmann *et al.*, *Plant Physiol.* **142**, 1739 (2006).
39. Y. Murakami *et al.*, *DNA Res.* **13**, 255 (2006).
40. L. Schauser, A. Roussis, J. Stiller, J. Stougaard, *Nature* **402**, 191 (1999).
41. A. Y. Borisov *et al.*, *Plant Physiol.* **131**, 1009 (2003).
42. L. Schauser, W. Wieloch, J. Stougaard, *J. Mol. Evol.* **60**, 229 (2005).
43. I. R. Searle *et al.*, *Science* **299**, 109 (2003).
44. L. Krusell *et al.*, *Nature* **420**, 422 (2002).
45. R. Jing *et al.*, *Genetics* **177**, 2263 (2007).
46. R. V. Penmetsa *et al.*, *Plant J.* **55**, 580 (2008).
47. R. Nishimura, M. Ohmori, H. Fujita, M. Kawaguchi, *Proc. Natil. Aca. Sci. USA* **99**, 15206 (2002).
48. A. Cebolla *et al.*, *EMBO J.* **18**, 4476 (1999).
49. K. Yano *et al.*, *Proc. Natil. Aca. Sci. USA* **105**, 20540 (2008).

50. J. D. Murray *et al.*, *Science* **315**, 101 (2007).
51. Z. Kevei *et al.*, *Plant Cell* **19**, 3974 (2008).
52. N. Kanamori *et al.*, *Proc. Natil. Aca. Sci. USA* **103**, 359 (2006).
53. K. Saito *et al.*, *Plant Cell* **19**, 610 (2007).