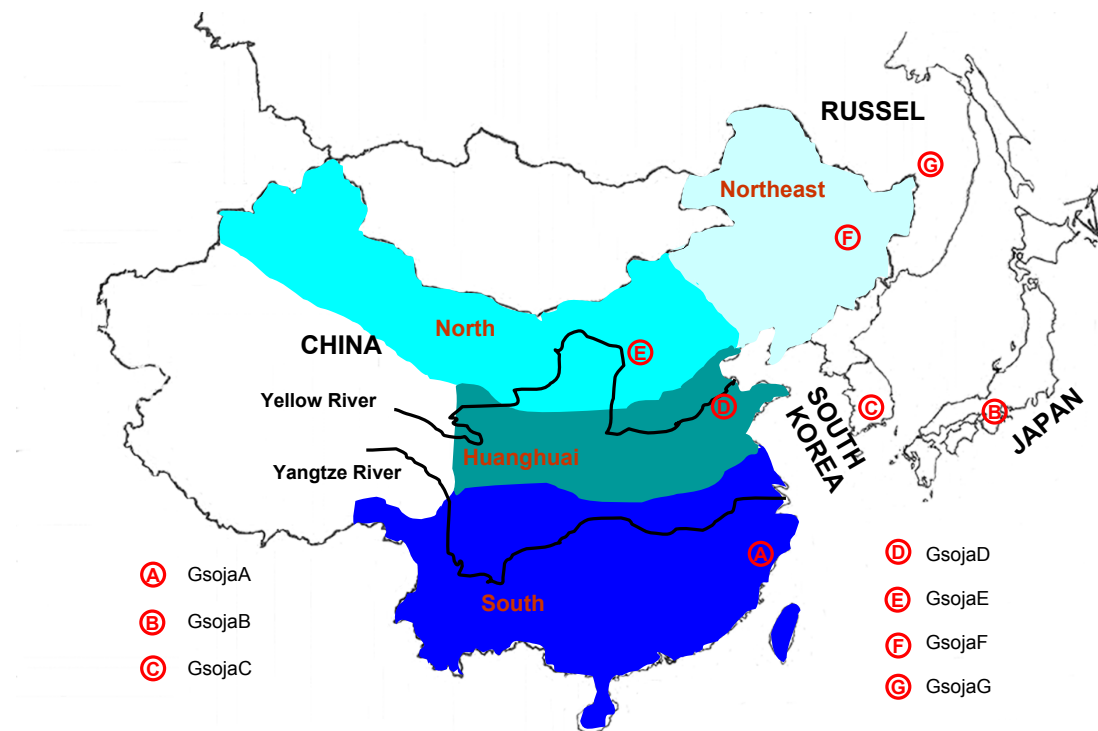
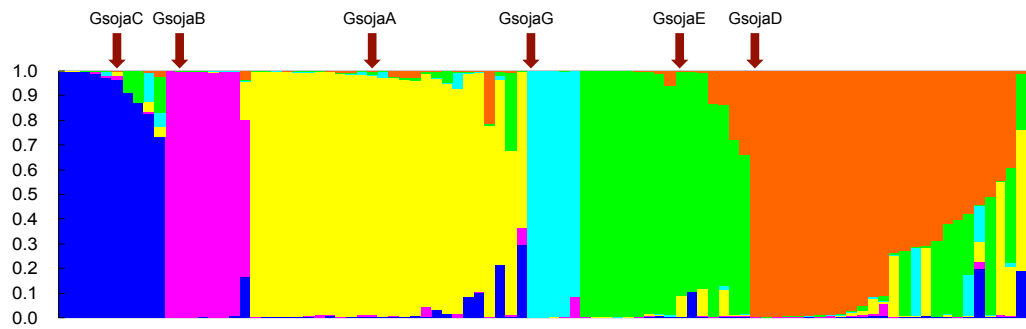


Supplementary Figure 1.



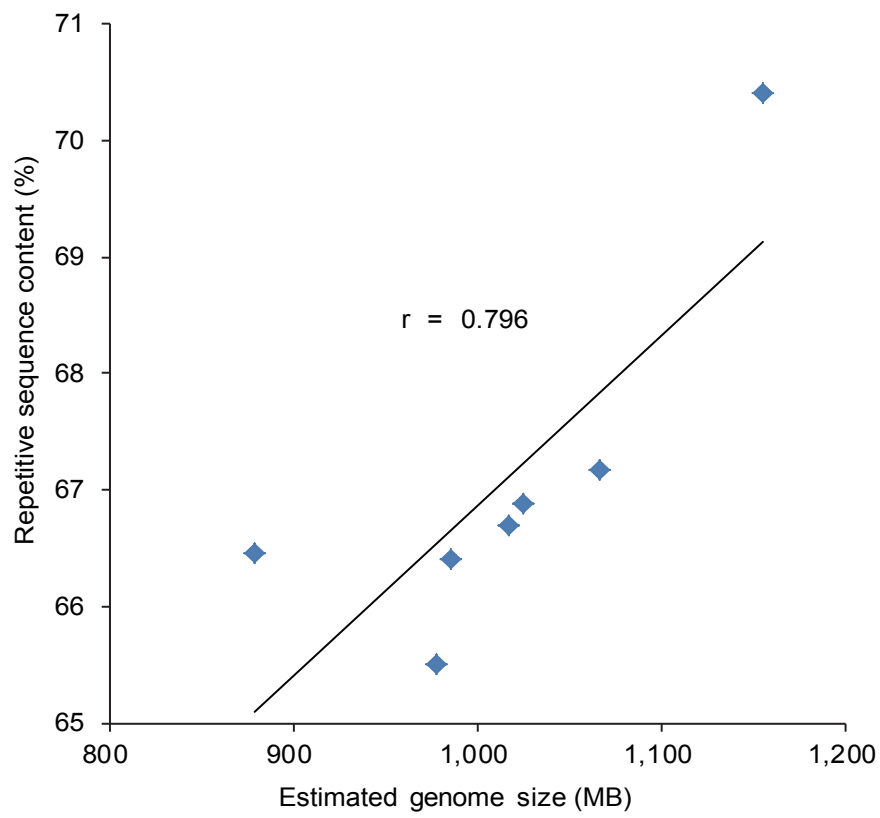
Geographical distribution of the seven sequenced *G. Soja* accessions in China. Four large regions, including Northeast, North, Huanghuai and South region are indicated by different colors.

Supplementary Figure 2.



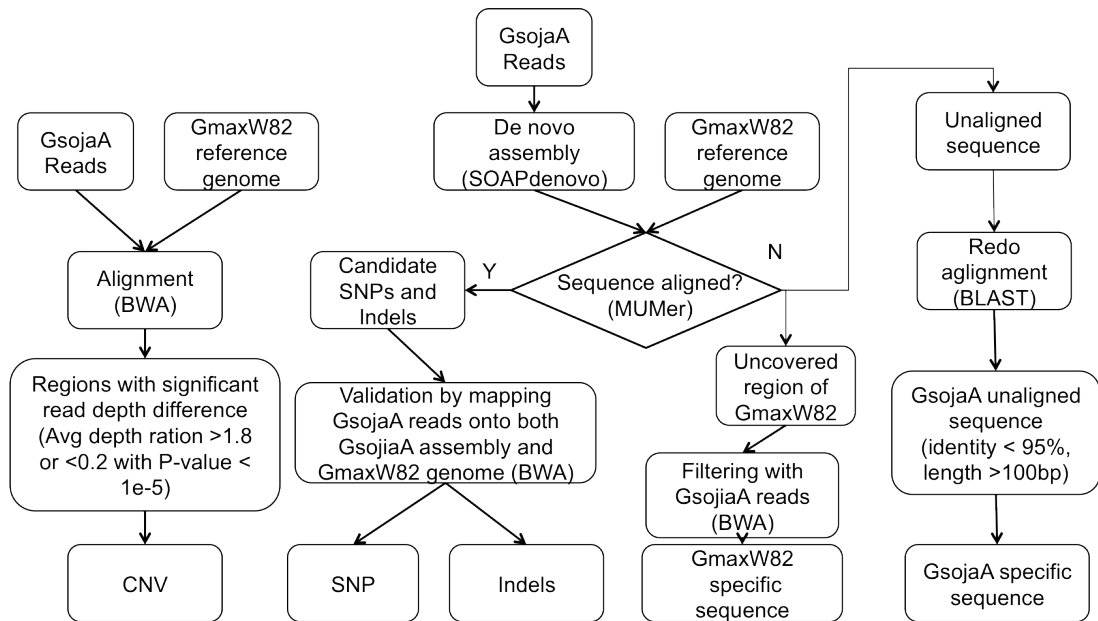
Population structure of 92 representative *G. soja* accessions as inferred by Bayesian clustering based on 99 simple sequence repeat (SSR) markers using STRUCTURE (Li et al. 2010, New Phytologist). Each individual is shown as a thin vertical line partitioned into six colored components, representing inferred membership in six genetic clusters. The six accessions representing six genetic clusters respectively are indicated.

Supplementary Figure 3.



Correlation between estimated genome size and repetitive sequences of the *G. soja* accessions.

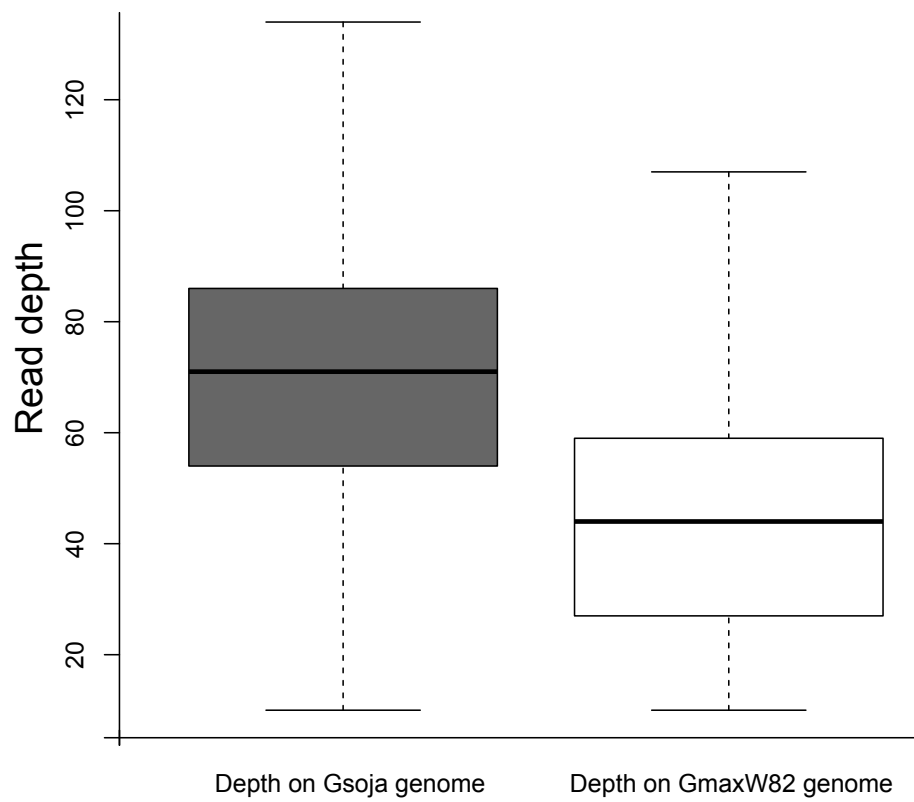
Supplementary Figure 4.



Flowchart for detection of variation between GsojaA and the GmaxW82 genome. The same pipeline was applied for all seven *G. soja* accessions.

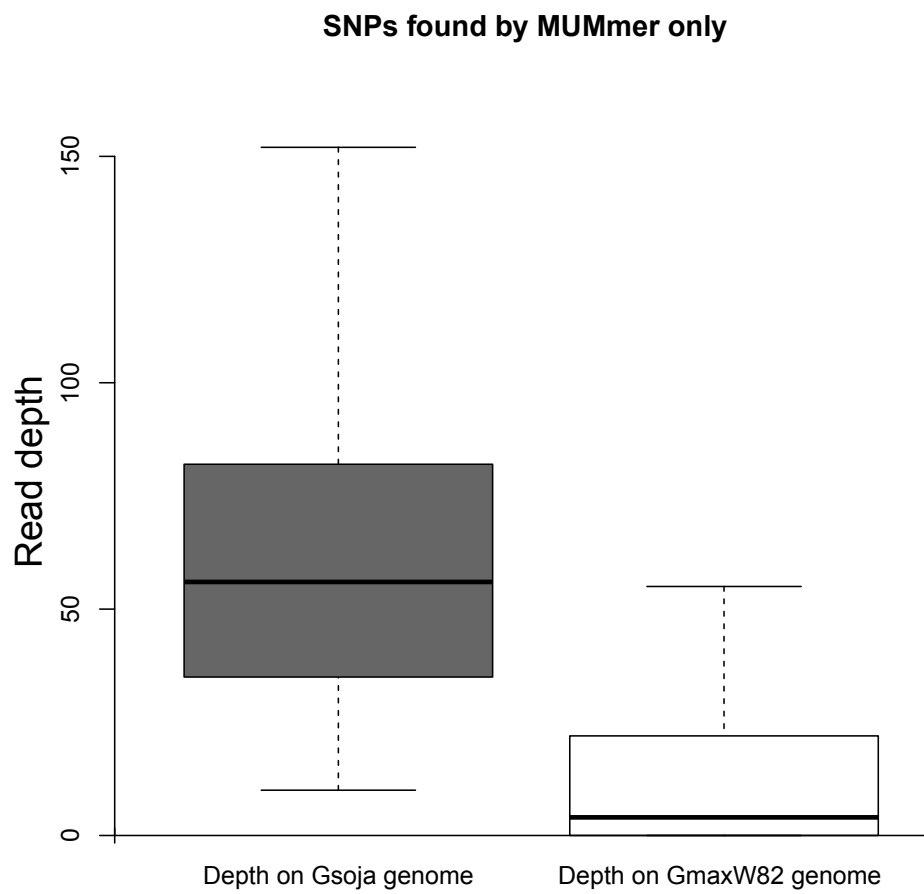
Supplementary Figure 5.

Overlap SNPs found by MUMmer and SAMtools



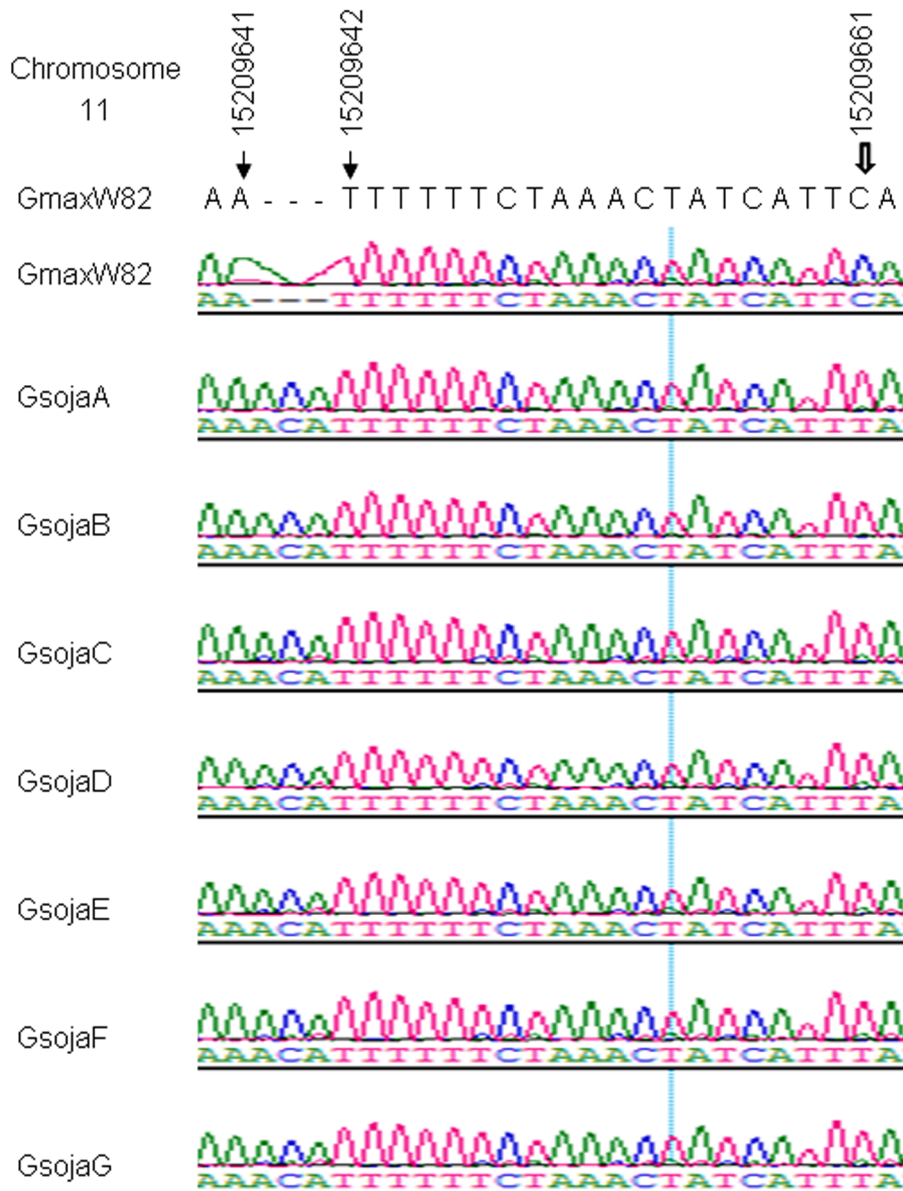
Read depth comparison of SNPs identified by MUMmer (assembly-based method) and SAMtools (alignment-based method).

Supplementary Figure 6.



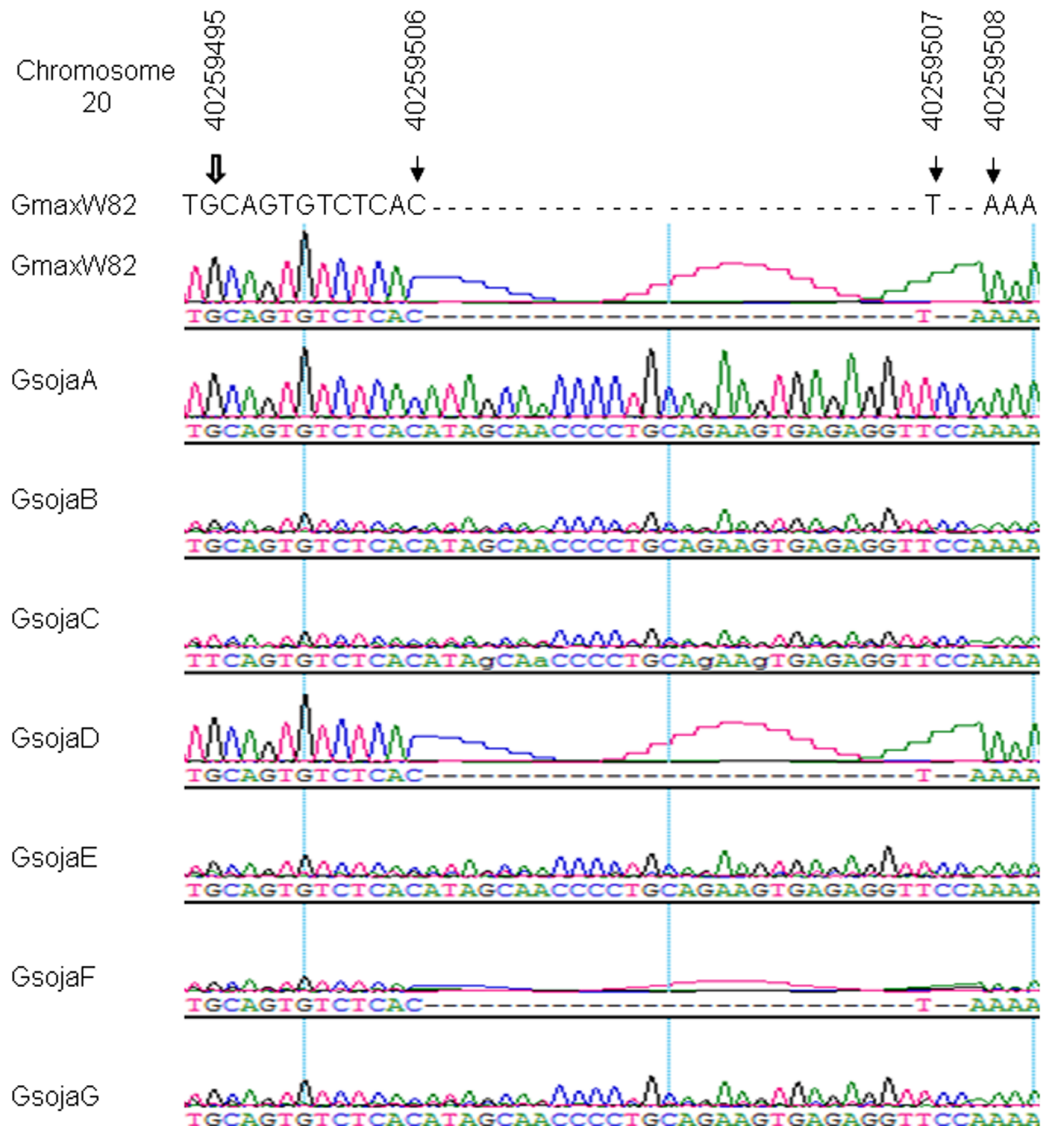
Read depth comparison on SNPs identified only by MUMmer (assembly-based method).

Supplementary Figure 7.



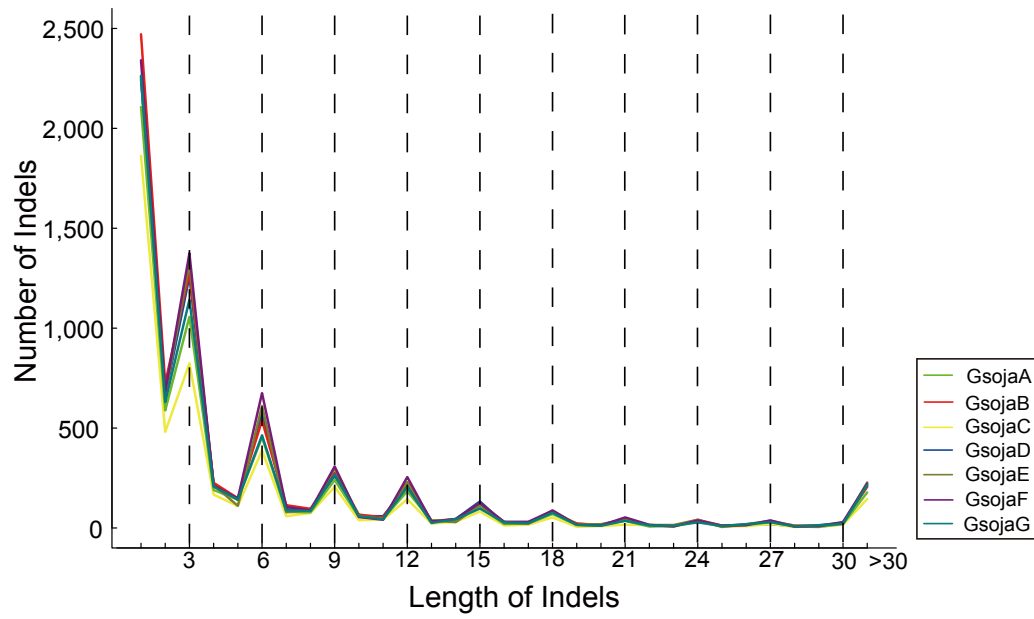
Confirmation of one SNP (C/T) and one indel (--/TT) on chromosome11 identified in the seven assembled *G. soja* genomes and GmaxW82 using Sanger sequencing. The SNP and indel are indicated by hollow arrow and solid arrow, respectively. The location of SNP and indel on GmaxW82 genome are also displayed.

Supplementary Figure 8.



Confirmation of one SNP (G/T) and two indels (22 and 2 nt respectively) on chromosome20 identified in the seven assembled *G. soja* genomes and GmaxW82 using Sanger sequencing. The SNP and indels are indicated by hollow arrow and solid arrow respectively. The location of SNP and indels on GmaxW82 genome are also displayed.

Supplementary Figure 9.



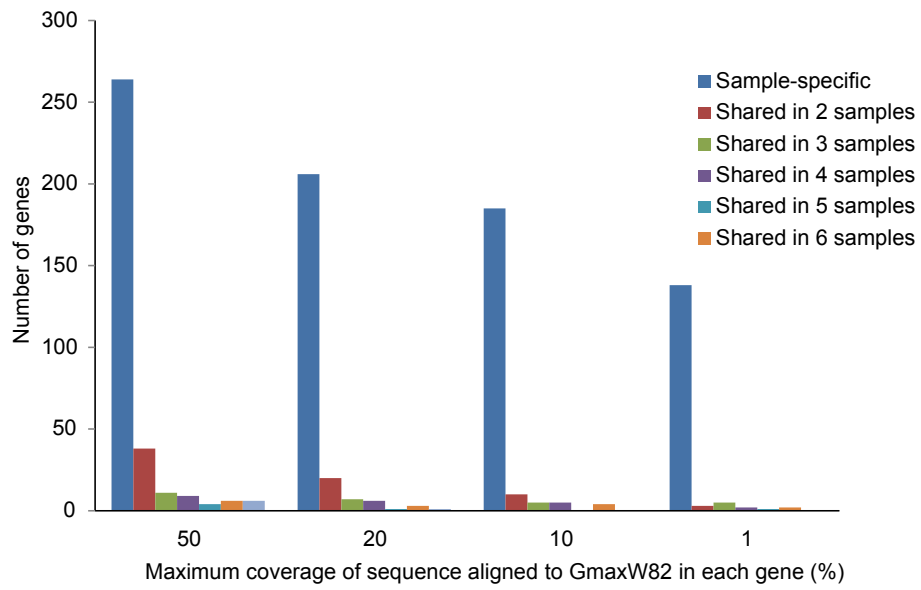
Length distribution of indels in CDS regions. The 3 bp size periodicity is shown by dashed vertical lines.

Supplementary Figure 10.



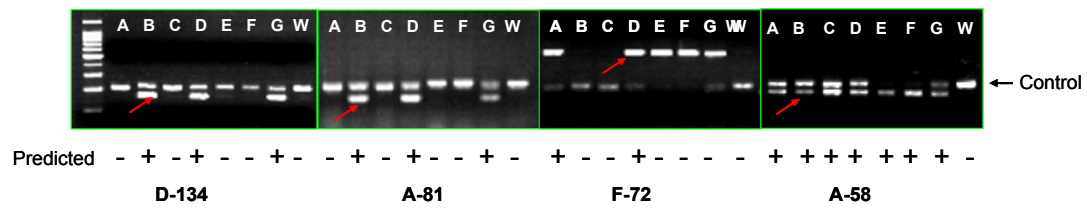
Morphological and physiological differences between *G. soja* and cultivated *G. max*. (*G. soja* plant was supported by a pole.)

Supplementary Figure 11.



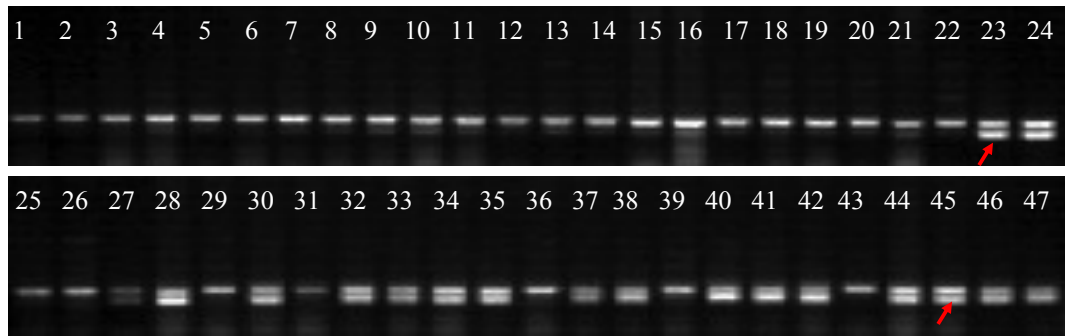
Number of *G. soja*-specific genes under different alignment cutoffs. X-axis represents the least percentage of unaligned length in a gene required to be considered a *G. soja*-specific gene.

Supplementary Figure 12.



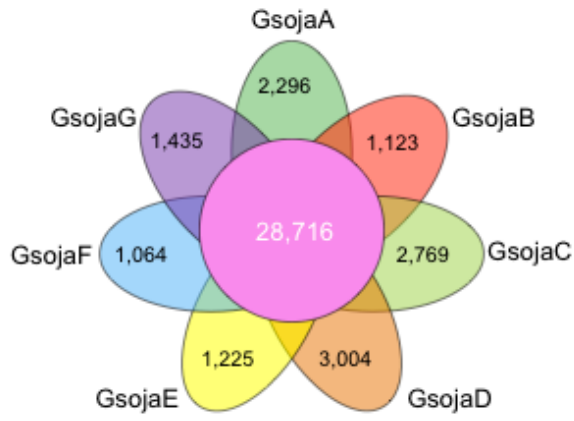
PCR-based validation of presence/absence of *G. soja*-specific sequence in seven *G. soja* accessions and GmaxW82.344 PCR products were obtained from 43 primer pairs on 8 genotypes (covering 37 loci). The predicted PAV within seven *G. soja* and GmaxW82 genomes based on assembly are shown below the agarose gel picture (“-” = absent and “+” = present). A common control sequence (218bp) in wild and cultivated was used as a control, indicated by arrow to right. Red arrow indicates PAV band. 100 bp ladders were used for molecular weight markers, from 100bp to 1000bp. The codes of accessions (A-G and W) replacing of GsojaA-G and GmaxW82 respectively are indicated above the picture. Marker names are indicated below the picture.

Supplementary Figure 13.



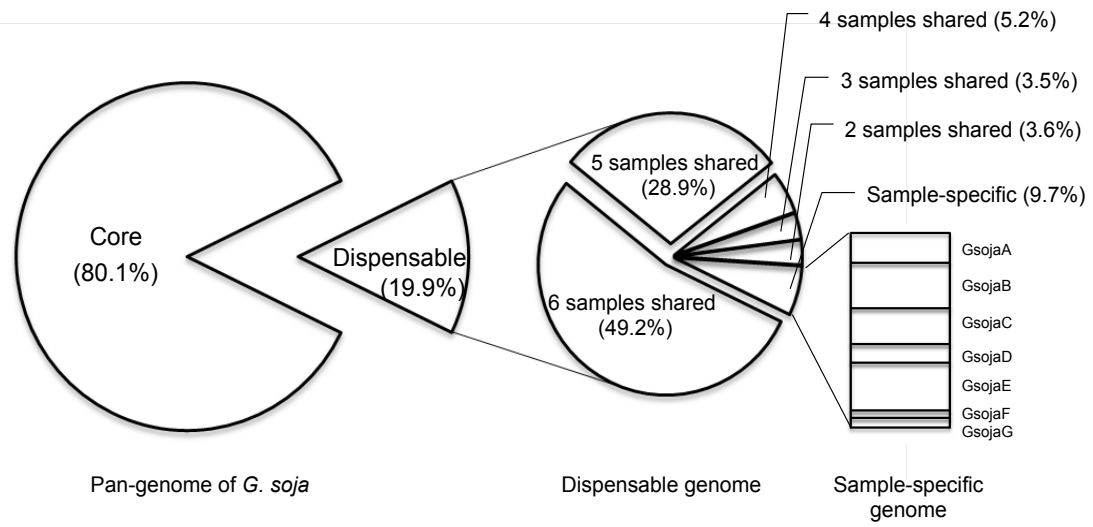
PCR-based validation of PAV variation (GLY locus, A-58) in 47 soybean germplasm accessions (1-22 *G. max* landraces; 23-47 *G. soja* accessions). A common control sequence in wild and cultivated (upper band, 218bp) was used as a control. Red arrow indicates PAV band. 100 bp ladder was used as molecular weight marker.

Supplementary Figure 14.



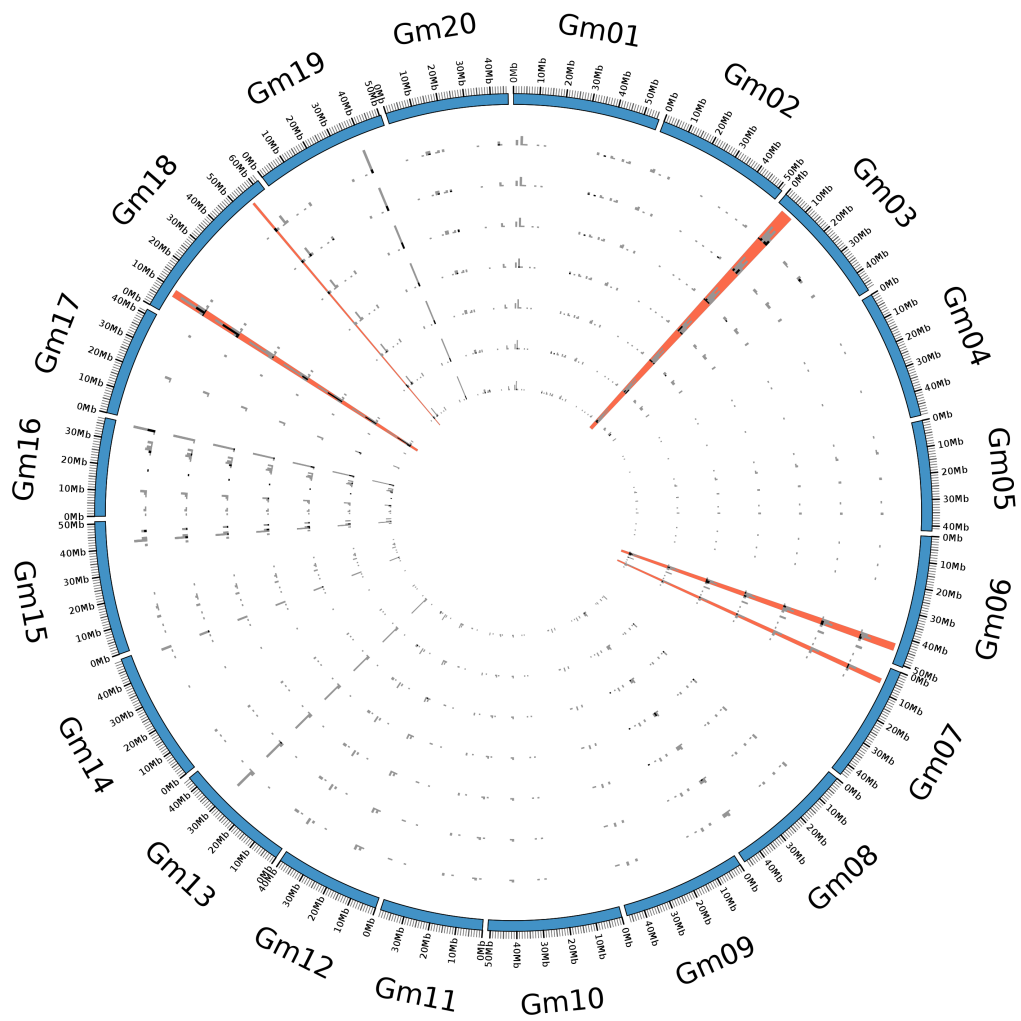
Annotated gene families of the core and dispensable genomes.

Supplementary Figure 15.



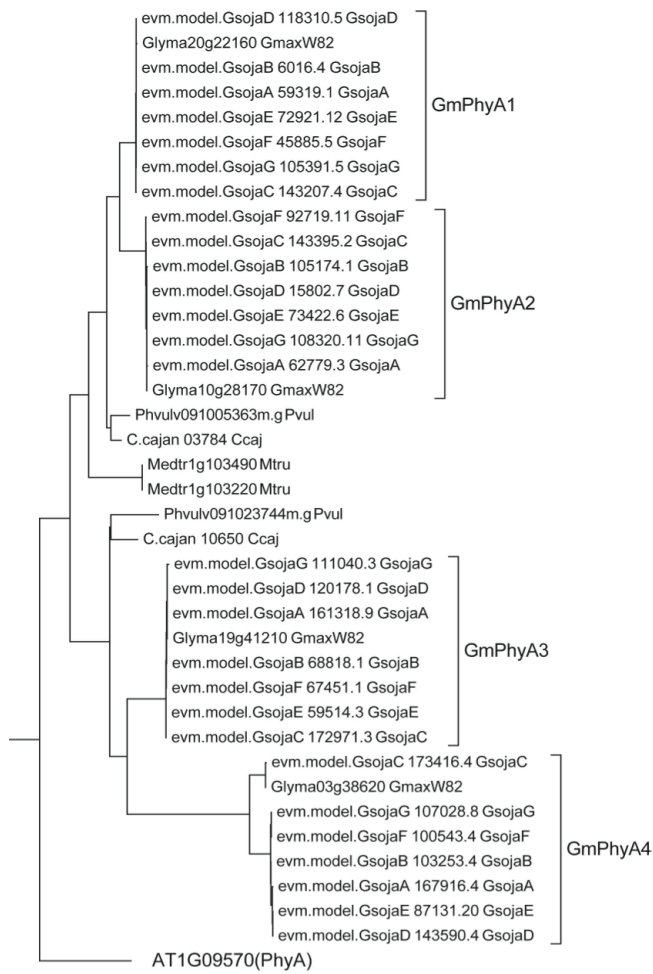
Composition of *G. soja* Pan-genome constructed using all of seven assembled *G. soja* genomes.

Supplementary Figure 16.



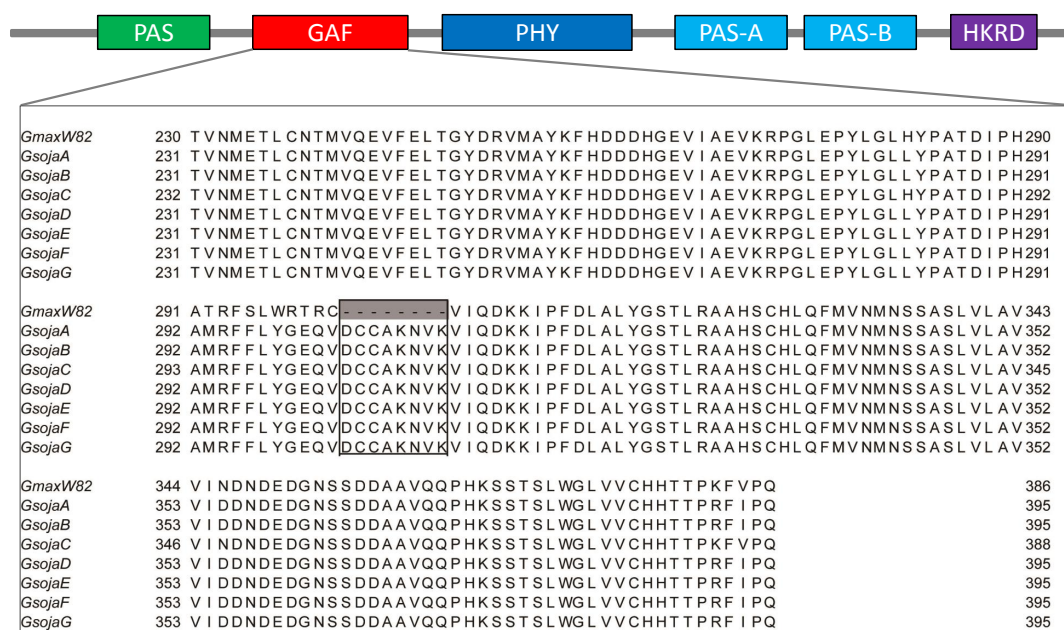
Distribution of R-genes. Red blocks show the five structurally diverged regions reported by McHale *et al.* 2012, black and grey bars show R-genes with/without copy number change. From outer to inner circles: GsojaA to GsojaG.

Supplementary Figure 17.



Phylogenetic tree of *PHYA* family

Supplementary Figure 18.



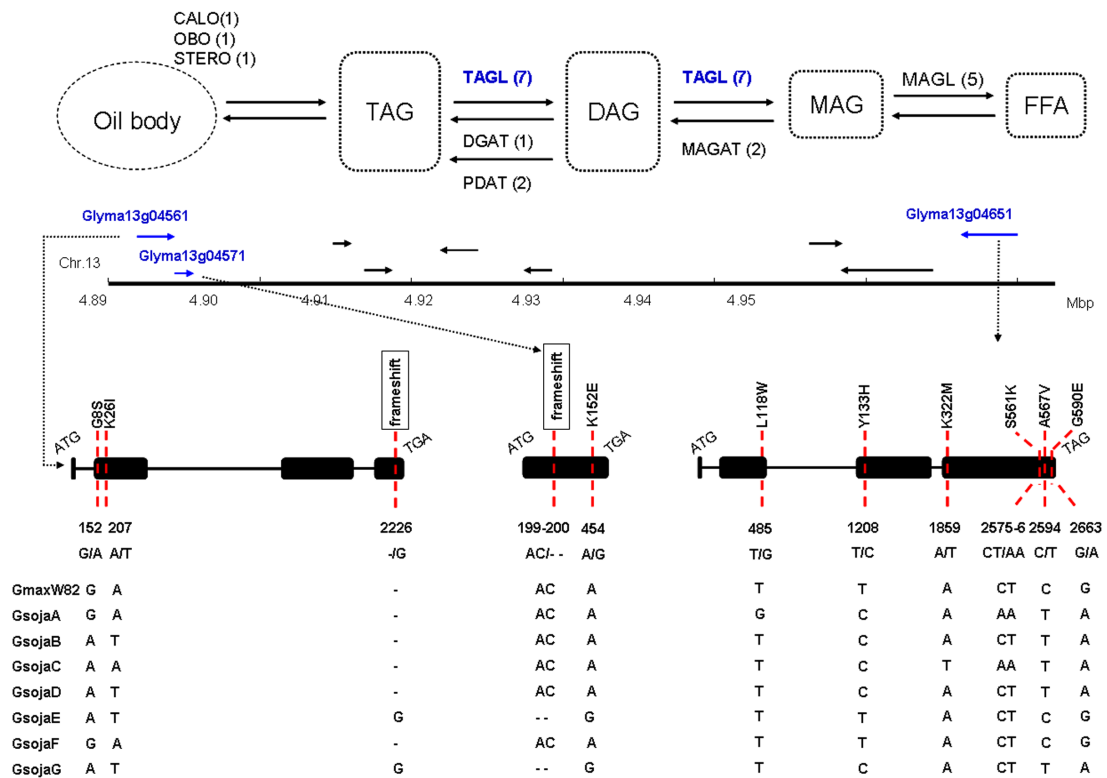
Full alignment of *GmPHYA4* from seven *G. soja* accessions.

Supplementary Figure 19.

	10	20	30	40	50	60	
GmaxW82	-----						
GsojaA	MPSGSRDPLVVGRVIGDVLDPFECSIPMRVSYNNRDVSNNGCEF KPSQVVNQPRINIGGDDL R						62
GsojaB	MPSGSRDPLVVGRVIGDVLDPFECSIPMRVSYNNRDVSNNGCEF KPSQVVNQPRINIGGDDL R						62
GsojaC	MPSGSRDPLVVGRVIGDVLDPFECSIPMRVSYNNRDVSNNGCEF KPSQVVNQPRINIGGDDL R						62
GsojaD	MPSGSRDPLVVGRVIGDVLDPFECSIPMRVSYNNRDVSNNGCEF KPSQVVNQPRINIGGDDL R						62
GsojaE	MPSGSRDPLVVGRVIGDVLDPFECSIPMRVSYNNRDVSNNGCEF KPSQVVNQPRINIGGDDL R						62
GsojaF	MPSGSRDPLVVGRVIGDVLDPFECSIPMRVSYNNRDVSNNGCEF KPSQVVNQPRINIGGDDL R						62
GsojaG	MPSGSRDPLVVGRVIGDVLDPFECSIPMRVSYNNRDVSNNGCEF KPSQVVNQPRINIGGDDL R						62
	70	80	90	100	110	120	
GmaxW82	-----						
GsojaA	NFYTLIAVDPDAPSPSDPNLREYLHWLVTDIPATTGPSFGHEVVTYESPRPMMGIHRLVFL						11
GsojaB	NFYTLIAVDPDAPSPSDPNLREYLHWLVTDIPATTGPSFGHEVVTYESPRPMMGIHRLVFL						124
GsojaC	NFYTLIAVDPDAPSPSDPNLREYLHWLVTDIPATTGPSFGHEVVTYESPRPMMGIHRLVFL						124
GsojaD	NFYTLIAVDPDAPSPSDPNLREYLHWLVTDIPATTGPSFGHEVVTYESPRPMMGIHRLVFL						124
GsojaE	NFYTLIAVDPDAPSPSDPNLREYLHWLVTDIPATTGPSFGHEVVTYESPRPMMGIHRLVFL						124
GsojaF	NFYTLIAVDPDAPSPSDPNLREYLHWLVTDIPATTGPSFGHEVVTYESPRPMMGIHRLVFL						124
GsojaG	NFYTLIAVDPDAPSPSDPNLREYLHWLVTDIPATTGPSFGHEVVTYESPRPMMGIHRLVFL						124
	130	140	150	160	170		
GmaxW82	FRQLGRET VYAPGWRQNFNTREFAELYNLGLPVAAVYFN IQRESGSGGRRLYH						64
GsojaA	FRQLGRET VYAPGWRQNFNTREFAELYNLGLPVAAVYFN IQRESGSGGRRLYH						177
GsojaB	FRQLGRET VYAPGWRQNFNTREFAELYNLGLPVAAVYFN IQRESGSGGRRLYH						177
GsojaC	FRQLGRET VYAPGWRQNFNTREFAELYNLGLPVAAVYFN IQRESGSGGRRLYH						177
GsojaD	FRQLGRET VYAPGWRQNFNTREFAELYNLGLPVAAVYFN IQRESGSGGRRLYH						177
GsojaE	FRQLGRET VYAPGWRQNFNTREFAELYNLGLPVAAVYFN IQRESGSGGRRLYH						177
GsojaF	FRQLGRET VYAPGWRQNFNTREFAELYNLGLPVAAVYFN IQRESGSGGRRLYH						177
GsojaG	FRQLGRET VYAPGWRQNFNTREFAELYNLGLPVAAVYFN IQRESGSGGRRLYH						177

Full alignment of *FT2c* from seven *G. soja* accessions.

Supplementary Figure 20.



Genes with large-effect SNP/indel detected in triacylglycerol biosynthesis and triacylglycerol and fatty acid degradation pathways. Top, diagram of genes involved in the formation and degradation of TAG. Numbers of genes with putative large-effect SNP/indel are shown in brackets. Bottom, comparison of nucleotide and amino acid polymorphisms near by a meta-QTL controlling oil content (chromosome13, 4.89-11.49 Mb) in seven *G. soja* accessions and GmaxW82. Non-synonymous mutation sites (red dashed lines) in three predicted *TAGL* genes (blue) are shown.

Supplementary Table 1. Morphological and physiological traits of seven *G. soja* accessions and GmaxW82.

Name	Code	100-seeds weight (g)	Oil content (%) [*]	Protein content (%) [*]	Flower color	Flowering time [#]	Maturity time [#]	R/V ⁼
GsojaA	ZYD04569	1.9	7.6	49.3	Purple	118	160	0.36
GsojaB	PI507600	1.5	9.0	46.5	Purple	136	-	-
GsojaC	PI407222	1.4	11.2	40.5	Purple	116	166	0.43
GsojaD	ZYD03247	1.2	14.4	42.6	Purple	106	154	0.45
GsojaE	ZYD02878	1.3	5.4	41.6	Purple	81	136	0.68
GsojaF	ZYD00401	1.3	6.4	51.6	Purple	51	103	1.02
GsojaG	PI578344B	1.6	10.7	45.3	Purple	50	98	0.96
GmaxW82	PI 518671 (Williams 82)	15.8	20.3	41.5	White	43	130	2.02

Notes:

^{*}From Chinese National Soybean GeneBank (CNSGB) and USDA-ARS Soybean Germplasm Collection (University of Illinois, Urbana, IL) respectively.[#]Days after emergence; phenotyped in Beijing, China; sowing in the spring (12/05/2012).⁼R, reproductive growth period, days from flowering date to maturity date; V, vegetative growth period, days from emergence date to flowering date. -, Initial pod when first frost was coming.**Supplementary Table 2.** Summary of sequencing data, in Giga base pairs (Gbp).

Library insert-size	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
180bp	91.56	84.64	90.44	79.73	57.46	49.29	58.81
500bp	18.84	18.43	24.46	20.20	20.46	21.88	22.41
2,000bp	12.21	17.49	11.80	43.85	33.16	17.09	22.86
Total	122.61	120.56	126.70	143.78	111.08	88.26	104.08

Supplementary Table 3. Statistics of predicted protein-coding genes for seven *G. soja* accessions and GmaxW82.

	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG	GmaxW82*
Original prediction [#]	103,396	93,534	110,508	106,462	96,125	89,814	93,470	NA
Predicted genes ⁼	58,756	56,655	60,377	62,048	58,414	57,573	58,169	NA
Refined genes ^{&}	55,061	54,256	56,542	57,631	55,901	54,805	54,797	54,175
With homology support (%)	96.19	97.16	96.36	96.15	96.88	97.04	96.93	NA
Average transcript length (bp)	2,596	2,857	2,351	2,530	2,817	2,809	2,738	3,192
Average ORF length (bp)	1,059	1,124	1,041	1,032	1,113	1,108	1,086	1,185
Average exons per gene	4.48	4.75	4.38	4.33	4.69	4.70	4.60	5.17
Average exon length (bp)	237	237	237	239	237	236	236	229
Average intron length (bp)	442	463	441	450	462	459	458	481

Notes:

* Glymav1.1

[#]Original prediction represents original genes models generated by evidence modeler after combing *ab initio* and homology-based predictions⁼Predicted gene represents genes after filtering transposable element related genes and genes with low homology supporting[&]Refined gene represents genes after correction and refinement of fragmental prediction.**Supplementary Table 4.** RNA expression evidence for predicted genes.

Sample	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
Raw data (Gb)	5.92	5.56	5.29	5.06	4.47	3.97	5.2
Mapping rate (%)	91.62	90.34	88.03	89.62	93.54	92	84.96
Gene coverage (%)	82.9	66.3	71.3	75.2	73.9	78.5	75.7
Total genes	58,756	56,655	60,377	62,048	58,414	57,573	58,169
Expressed genes with coverage >50%(%)	74.07	66.98	71.24	73.44	71.46	69.62	70.02
Expressed genes with coverage >90%(%)	71.19	63.94	67.86	67.41	68.12	65.87	66.53

Supplementary Table 5. Summary of identified SNPs in seven *G. soja* genomes.

	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
Total	4,423,892	4,468,029	3,984,568	4,724,306	3,905,164	4,139,688	3,629,763
Homozygous	3,246,925	4,344,518	2,479,337	4,228,749	3,781,671	3,992,754	3,490,194
Heterozygous	1,176,967	123,511	1,505,231	495,557	123,493	146,934	139,569
Intergenic	3,866,846	3,984,418	3,456,077	4,221,259	3,473,693	3,684,739	3,200,072
Intron	401,804	349,160	380,406	363,050	310,723	328,005	299,657
CDS	154,312	133,606	147,268	139,173	120,037	126,173	129,241
Synonymous	64,652	55,758	62,245	57,682	49,584	52,255	53,950
Nonsynonymous	87,369	75,865	82,864	79,541	68,684	72,027	73,363
Splicing	930	845	817	825	711	771	794
Stop gain/loss	2,286	1,979	2,155	1,947	1,764	1,886	1,923

Supplementary Table 6. Prediction and validation of 24 indels in seven *G. soja* genomes as compared to the GmaxW82 genome.

Loci	Chr.	Position and indel information	Predicted by alignment or Validated by Sanger sequencing	Insertion (INS) and/or deletion (DEL) of indels in seven <i>G. soja</i> genomes							
				GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG	
1	1	49611857-49611963 DEL:6:ATATAT*	Predicted Validated [#]	DEL DEL							
2	1	49611859-49611863 DEL:4:ATAT	Predicted Validated				DEL DEL				
3	1	49611861-49611863 DEL:2:AT	Predicted Validated			DEL DEL		DEL DEL	DEL DEL	DEL DEL	
4	1	54855722-54855723 INS:3:TAC	Predicted Validated		INS INS						
5	1	54855725-54855727 DEL:2:CA	Predicted Validated			DEL DEL					
6	1	54855732-54855734 DEL:2:TA	Predicted Validated	DEL DEL	DEL DEL			DEL ?	DEL DEL	DEL DEL	
7	2	14450193-14450195 DEL:2:TA	Predicted Validated						DEL DEL	DEL DEL	
8	2	14452554-14452555 INS:2:CC	Predicted Validated	INS INS	INS INS			INS INS	INS INS	INS INS	
9	2	14452556-14452557 INS:1:T	Predicted Validated	INS INS	INS INS			INS INS	INS INS	INS INS	
10	6	1362992-1362993 INS:19:CAAATTTGTAACACTACATT	Predicted Validated			INS INS		INS ?	INS INS	INS INS	INS INS
11	8	14064218-14064219 INS:5:ATATT	Predicted Validated	INS INS		INS INS	INS INS	INS ?	INS INS	INS INS	INS INS
12	11	15209641-15209642 INS:3:CAA	Predicted Validated	INS INS	INS INS	INS INS	INS INS	INS INS	INS INS	INS INS	INS INS
13	13	35725358-35725362 DEL:4:AAAA	Predicted Validated		DEL DEL						
14	13	35725361-35725362 INS:7:AAAAAAA	Predicted Validated	INS INS			INS INS		INS INS	INS INS	INS INS
15	14	4543368-4543369 INS:12:ACTCTTGATCTT	Predicted Validated	INS INS							

16	14	4543409-4543415	Predicted							DEL
		DEL:6:CTGTAC	Validated							DEL
17	15	10101443-10101444	Predicted				INS	INS	INS	INS
		INS:5:GTCCT	Validated				INS	INS	INS	INS
18	15	10101444-10101445	Predicted				INS	INS	INS	INS
		INS:2:GT	Validated				INS	INS	INS	INS
19	16	7107494-7197495	Predicted	INS			INS	INS	INS	INS
		INS:9:GTTTTGGTA	Validated	INS			INS	?	INS	INS
20	17	37682857-37682858	Predicted		INS		INS			
		INS:17:GCTAGCGCGACGAAAAC	Validated		INS		INS			
21	17	37682861-37682862	Predicted		INS		INS			
		INS4:GCGT	Validated		INS		INS			
22	17	37682861-37682862	Predicted	INS		INS			INS	
		INS:21:AGCAGCGCGATCGCGAATG	Validated	?		INS			INS	
		TG								
23	20	40259506-40259507	Predicted	INS	INS	INS		INS		INS
		INS:27:TGGAGAGTGAAGACGTCCC	Validated	INS	INS	INS		INS		INS
		CAACGATA								
24	20	40259507-40259508	Predicted	INS	INS	INS		INS		INS
		INS:2:CC	Validated	INS	INS	INS		INS		INS

* "INS" notes for insertion, "DEL" notes for deletion, "?"notes for missing data; digits following INS or DEL means the size of indels; the nucleotides following digits represent insert/deleted sequences.

#Validation was based on Sanger sequencing results.

Supplementary Table 7. Summary of identified indels.

	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
Total	619,132	772,978	532,597	735,258	706,611	736,717	504,081
Intergenic	538,362	680,395	465,562	651,131	621,830	648,598	431,787
Intron	74,524	85,096	61,836	77,023	77,531	80,532	65,517
CDS	6,095	7,304	5,053	6,943	7,070	7,404	6,577
Frameshift	3,497	4,181	2,989	3,863	3,802	3,958	3,781
Splicing	151	183	146	161	180	183	200
Stop gain/loss	251	296	194	262	287	297	239

Supplementary Table 8. Summary of large-effect mutations resulting in stop codon gain/ loss or frameshift.

	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG	Intersection	Union total
Stop gain/ loss	2,167	1,879	1,991	1,871	1,761	1,836	1,829	221	5,647
Frameshift	2,428	2,914	2,031	2,671	2,636	2,771	2,682	717	6,321
Total	4,129	4,245	3,609	4,030	3,908	4,088	4,007	898	9,811

Supplementary Table 9. Summary of identified CNV genes.

	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
Copy number loss	321	448	303	301	335	313	469
Copy number gain	224	215	253	186	249	242	387

Supplementary Table 10. Shared and sample-specific CNV genes.

	# Genes shared across <i>G. soja</i> accessions							Total CNV genes
	N=7	N=6	N=5	N=4	N=3	N=2	N=1	
Loss	13	26	47	88	144	249	612	1,179
Gain	56	16	35	41	73	103	402	726
Loss/Gain	2	1	3	11	25	31	-	73
Sum	71	43	85	140	242	383	1,014	1,978

Supplementary Table 11. Statistics of *G. soja*-specific sequences.

Sample Name		GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
Raw novel sequence	Number [*]	41,942	30,765	46,692	42,758	30,527	31,094	30,621
	Length(bp) [#]	12,284,124	10,136,359	13,040,955	12,809,910	11,342,412	9,735,843	9,658,322
	Max length(bp) ⁼	19,263	43,106	25,262	45,404	65,863	9,294	7,003
Filtered novel sequence	Number	9,359	7,213	8,634	9,527	7,030	7,161	6,982
	Length(bp)	3,096,875	2,718,224	2,811,760	3,230,823	3,911,409	2,323,724	2,366,349
	Max length(bp)	18,859	14,823	25,262	19,178	65,863	9,294	4,465

Notes:

Novel sequences with length > 100 were retained;

^{*}Number: total number of novel sequence;[#]Length: total length of all novel sequence;⁼Max length: max length of novel sequence.**Supplementary Table 12.** Statistics of *G. soja*-specific genes.

	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
Specific genes	80	48	79	67	126	46	61
Expressed genes (coverage >50%)	20	16	25	24	20	13	25
Expressed genes (coverage >90%)	15	15	20	15	12	8	17

Supplementary Table 13. Primer sequences and expected PCR product sizes for 43*G.soja*-specific sequences/genes and the control sequences.

ID	Primer sequence (5'-3')		Predicted size of PCR products (bp)	Ta(°C)*	Note
	Forward	Reverse			
A-107	GCCTGTCCTTTAGATCCCATAC	ACCTCATGACCGTTGAAGAC	424	55	
A-3	TCATCCCAACTTGACCTATGC	TCCTTCTCCTTCTAGTGTCAATATTC	172	55	
A-33	TGAGATTCAGGTGGAAGGCG	GCTGTTCGTGACTACCTTGCCA	320	50	With common band in length of 500bp
A-35	GCTTTCATTGGTCTCAAATACTCC	ACTCAATTCAGTCATACTCCCATAC	316	55	
A-58	CAGATTATGAGCACAGACA	AGGTGGTTGTGCTTGGAG	168	55	
A-71	CACACTCTTTGATTGCATCGTG	AGCAACCACAACCTCTACAACC	162	55	
A-81	CCACGAAGTGTCTCCGG	CAGCCACTGATCACCGAC	162	55	
A-85	CCAATCTCTAAACGCTCCCTG	TGTTCTGCTTGAAAGGGACC	497	55	
A-94	CGACTACCACATCCAATATAGAGC	ATTCAGTGAGCAGTGTGGAG	296	55	
B-102	AAGGCAATTCCTCCACAT	AGGTGTTACTCATGCATC	182	50	
B-120	CATCCCATTCTCCCCACATAC	GCCATCTGTGCAATTCATCTTG	323	55	
B-122	AGAAGTCATTGCTGGAGGA	GGTTAACATCCTCGCCGCC	157	50	
B-13	TTATTCAGACTTAGCGACACCC	TCGACGTTGGAGTTTCTCATG	330	55	
B-139	TGTTCTGCTTGAAAGGGACC	CCAATCTCTAAACGCTCCCTG	499	55	
B-142	CATCCTTCGTGGTGAACCTCT	ATGGTAGTTCTTGTCCGGTATT	179	55	
B-41	ATATTCACGCACCTTACCCG	GGAAACCACGCACTTATCAG	154	55	
B-56	TCTGTGGAAAGGGAACCG	CACCGAGCCATCCCATAAAG	169	55	
B-97	CAATTGTTCTGTTGGCCTAG	TCTTCAAAATGAAACCAACTGGC	343	55	
C-23	TCTGTATGTTGAAGGTGGGC	TGGAATGTAACCTAAGGCTTGGG	340	55	
C-31	GAAACCCGCAGAGAAAATCAG	CAGGCAAAGATAAAGGGAATGTG	172	55	

C-41	TGTTCGTCTTGAAAGGGACC	CCAATCTCTAAACGCTCCCTG	499	55	
C-52	GTTACCTCTAACTCAGCCTTC	AATGGCGGTTGAGTCTAGC	368/400	55	Two alleles
C-73	AAAATAAGGAGCTAGAAGGCC	AGTTCATTTGCCCTACCATG	378	55	
D-118	CAAGGCGAGACAGAGTAGAC	ACAAGTAGAGGAACCAAGCG	184	55	
D-13	CATTGCTGGAGGAGGAGG	TCACGCGGGTTAACATCC	157	55	
D-134	CGGGTTGCACATTTAGTTGG	ACGGACTATGCAGCTTGTG	163	55	
D-142	AATCTGGCTCTTCTGCTTCG	ACACAGATCCGCTATGGTAGT	302	55	
D-2	ACCTCATGACCGTTGAAGAC	GCCTGTCCTTTAGATCCCATAC	424	55	
D-4	GCAACCACAATTCTACAACCTTG	CACACTCTTTGATTGCATCGTG	171	55	
D-98	TCGACGTTGGAGTTTCTCATG	TTATTCAGACTTAGCGACACCC	329	55	
E-107	CAAACCAAATGACTCCAACCAC	CTCCTCGTAAAGCCGCAG	305	55	
E-114	AATCTGGCTCTTCTGCTTCG	CCGCTATGGTAGTTCTTGTGCG	294	55	
E-119	CACAGGCAAAGATAAAGGGAATG	GAAACCCGCAGAGAAAATCAG	175	55	
E-62	GAGCTGGCATTCAATCAAG	ATTCAGTGAGCAGTGTGGAG	276	55	
E-89	TGTATCCATGCTTTGTCCGC	CAGACACTAACAATTCGAACTGC	159	55	
F-64	GGGTTAACATCCTCGCCG	GATATGTTATCTTGC GGCTTTCTC	103	55	
F-72	AAAGATTTGATGGGTTGGCAC	GATGTTAGTCCAGGCGGTAAG	457	55	
F-74	GTGACCCTTGTGAGTATAGCC	ACTTGGCTTCTCACTGGATTC	105	55	
F-78	CAATGTTTCGTGTTGGCTAG	CCAAGTGGCTTCAATTCATGC	329	55	
G-113	GTGACCCTTGTGAGTATAGCC	ACTTGGCTTCTCACTGGATTC	105	55	
G-38	CCAATCTCTAAACGCTCCCTG	TGTTCTGCTTGAAAGGGACC	499	55	
G-42	ACTCTTTGGTATTGGTCCTTGTAG	GATCTAAGATTGGCAAGTTGTCG	195	55	
G-89	TCATCCCAACTTGACCTATGC	TCCTTCTCCTTCTAGTGTCAATATTC	173	55	
C1	GTCATTGTAACAGGTGGGGAGA	ACTGCGACTTTATTTAAGATAG	218		

Note: *Ta is the optimized annealing temperature for PCR.

Supplementary Table 14. Prediction and validation of 37 *G. soja*-specific PAV using 43 markers.

Loci	Predicted by alignment or Validated by PCR	The marker used to validated the prediction	Presence (P) and/or absence (A) of PAV in seven <i>G. soja</i> and GmaxW82 genomes							
			GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG	GmaxW82
1	Predicted		P	A	A	P	A	A	A	A
	Validated	A-35	A	A	A	P	A	A	A	A
		A-71	A	A	A	P	A	A	A	A
2	Predicted		P	P	P	P	P	P	P	A
	Validated	A-58	P	P	P	P	P	P	P	A
		C-52	P	P	P	P	P	P	P	A
3	Predicted		P	P	P	A	A	A	P	A
	Validated	A-3	A	P	P	A	A	A	P	A
4	Predicted		P	P	P	P	A	P	A	A
	Validated	B-102	P	P	P	P	A	P	A	A
5	Predicted		A	A	P	P	A	A	A	A
	Validated	C-23	A	P	P	A	A	A	A	A
6	Predicted		P	P	P	A	A	A	P	A
	Validated	C-73	P	P	P	A	P	P	A	A
7	Predicted		P	P	P	A	A	A	P	A
	Validated	A-85	P	P	P	A	A	A	P	A
		B-139	P	P	P	A	A	A	P	A
		C-41	P	P	P	A	A	A	P	A
8		G-38	P	P	P	A	A	A	P	A
	Predicted		P	A	A	P	A	A	A	A
	Validated	D-4	P	A	A	P	A	A	A	A
9	Predicted		A	P	A	P	P	A	A	A
	Validated	E-89	A	P	A	P	P	P	P	A

10	Predicted		P	P	P	P	P	P	P	A
	Validated	C-31	P	P	P	P	P	P	P	A
		E-119	P	P	P	P	P	P	P	A
11	Predicted		P	P	P	P	P	P	A	A
	Validated	B-120	P	P	P	P	P	P	A	A
12	Predicted		P	A	P	P	A	A	P	A
	Validated	A-107	P	A	P	P	A	A	P	A
13	Predicted		P	A	P	P	P	A	P	A
	Validated	A-33	P	P	P	P	A	A	P	A
14	Predicted		P	P	P	P	P	P	P	A
	Validated	A-94	P	P	P	P	P	P	P	A
15	Predicted		P	P	P	P	P	A	P	A
	Validated	B-13	P	P	P	P	P	A	P	A
16	Predicted		P	P	A	P	P	P	A	A
	Validated	B-142	A	P	A	P	A	A	A	A
17	Predicted		A	P	P	P	A	A	P	A
	Validated	B-41	P	P	A	P	A	A	P	A
18	Predicted		A	P	P	P	P	P	P	A
	Validated	B-97	A	P	P	P	P	P	P	A
19	Predicted		P	P	P	P	P	P	P	A
	Validated	D-13	P	P	P	P	P	P	P	A
20	Predicted		P	A	P	P	P	A	P	A
	Validated	A-81	A	P	A	P	A	A	P	A
21	Predicted		P	P	P	P	A	P	A	A
	Validated	B-56	A	P	A	P	P	P	P	A
22	Predicted		A	A	A	P	A	A	A	A
	Validated	D-118	A	A	A	P	P	P	P	A
23	Predicted		P	A	P	P	P	A	P	A
	Validated	D-134	A	P	A	P	A	A	P	A
24	Predicted		P	P	P	P	P	P	P	A

25	Validated	B-122	P	P	A	P	A	A	P	A
	Predicted		P	P	A	P	P	P	A	A
26	Validated	D-142	A	P	A	P	A	A	A	A
	Predicted		P	P	A	P	P	P	A	A
27	Validated	E-114	A	P	A	P	A	A	A	A
	Predicted		P	P	P	A	A	A	P	A
28	Validated	G-89	A	P	P	A	P	P	P	A
	Predicted		P	A	P	P	A	A	P	A
29	Validated	D-2	P	A	P	P	A	A	P	A
	Predicted		P	P	P	P	A	P	A	A
30	Validated	D-98	P	P	P	P	A	A	A	A
	Predicted		P	P	P	A	P	P	P	A
31	Validated	E-107	P	P	P	A	P	P	P	A
	Predicted		P	P	P	P	P	A	P	A
32	Validated	E-62	P	P	P	P	A	A	A	A
	Predicted		P	P	P	P	P	P	P	A
33	Validated	F-64	P	P	P	P	P	A	P	A
	Predicted		P	A	A	P	A	P	A	A
34	Validated	F-72	P	A	A	P	P	P	P	A
	Predicted		A	P	P	P	P	P	P	A
35	Validated	F-74	A	P	P	P	P	P	P	A
	Predicted		A	P	P	P	P	P	P	A
36	Validated	F-78	A	P	P	P	P	P	P	A
	Predicted		A	P	P	P	P	P	P	A
37	Validated	G-113	A	P	P	P	P	P	P	A
	Predicted		A	P	P	P	P	P	P	A
	Validated	G-42	A	P	P	P	P	P	P	A

Supplementary Table 15. Top ten GO categories in biological process, enriched for genes with *G. soja*-specific PAV.

GO number	Category description	Number of genes with PAV
GO:0055114	oxidation-reduction process	22
GO:0008152	metabolic process	19
GO:0006508	proteolysis	14
GO:0006952	defense response	12
GO:0006355	regulation of transcription, DNA-dependent	11
GO:0005975	carbohydrate metabolic process	9
GO:0019684	photosynthesis, light reaction	8
GO:0015979	photosynthesis	7
GO:0044237	cellular metabolic process	5
GO:0006810	transport	5

Supplementary Table 16. Comparison of variation between GmaxW82 and *G. soja* in pericentromere regions and chromosome arms.

	Average in pericentromere regions (%)	Average in chromosome arms (%)	Chi-square test p-value
Large effect mutations	22.27	16.46	<0.001
Copy number gain genes	1.95	1.13	0.128
Copy number loss genes	5.08	1.63	<0.001
GmaxW82 specific sequence	0.19	0.11	0.021

Supplementary Table 17. Enrichment in GO functional classes for genes under positive selection in *G. max*.

GO root	GO name	GO id	FDR	P-value	Frequency in GO	Frequency in total
BP*	base-excision repair	GO:0006284	0.352	0.001829	4/319(1.25%)	40/24322(0.16%)
BP	nitrate transport	GO:0015706	0.4767	0.002494	3/319(0.94%)	21/24322(0.09%)
MF [#]	DNA N-glycosylase activity	GO:0019104	0.3852	0.003269	3/413(0.73%)	23/31473(0.07%)
BP	(1->3)-beta-D-glucan biosynthetic process	GO:0006075	0.6097	0.003688	3/319(0.94%)	24/24322(0.10%)
BP	proline metabolic process	GO:0006560	0.6097	0.003688	3/319(0.94%)	24/24322(0.10%)
MF	1,3-beta-D-glucan synthase activity	GO:0003843	0.4309	0.00367	3/413(0.73%)	24/31473(0.08%)
CM ⁼	1,3-beta-D-glucan synthase complex	GO:0000148	0.2133	0.003861	3/216(1.39%)	24/16181(0.15%)
BP	stomatal complex development	GO:0010374	0.6414	0.004151	3/319(0.94%)	25/24322(0.10%)
BP	response to nitrate	GO:0010167	0.6824	0.004647	3/319(0.94%)	26/24322(0.11%)
BP	amino acid transport	GO:0006865	0.7569	0.006071	5/319(1.57%)	88/24322(0.36%)
MF	cyclic nucleotide binding	GO:0030551	0.603	0.006364	3/413(0.73%)	29/31473(0.09%)
BP	glutamine family amino acid biosynthetic process	GO:0009084	0.7878	0.006646	4/319(1.25%)	57/24322(0.23%)
MF	amino acid transmembrane transporter activity	GO:0015171	0.6171	0.00671	5/413(1.21%)	90/31473(0.29%)
MF	microtubule motor activity	GO:0003777	0.6971	0.008092	7/413(1.69%)	173/31473(0.55%)
BP	covalent chromatin modification	GO:0016569	0.898	0.01028	5/319(1.57%)	100/24322(0.41%)
CM	microtubule associated complex	GO:0005875	0.473	0.01062	7/216(3.24%)	180/16181(1.11%)
BP	mRNA processing	GO:0006397	0.9329	0.01251	5/319(1.57%)	105/24322(0.43%)
BP	microtubule-based movement	GO:0007018	0.9335	0.01263	7/319(2.19%)	189/24322(0.78%)
MF	voltage-gated potassium channel activity	GO:0005249	0.8873	0.01548	3/413(0.73%)	40/31473(0.13%)
BP	post-embryonic morphogenesis	GO:0009886	0.9888	0.02471	4/319(1.25%)	84/24322(0.35%)
MF	ATP-dependent peptidase activity	GO:0004176	0.9796	0.02793	2/413(0.48%)	20/31473(0.06%)
BP	nicotinamide nucleotide biosynthetic process	GO:0019359	0.9975	0.03056	2/319(0.63%)	21/24322(0.09%)
BP	root hair elongation	GO:0048767	0.9975	0.03056	2/319(0.63%)	21/24322(0.09%)
CM	nuclear speck	GO:0016607	0.8168	0.03153	2/216(0.93%)	21/16181(0.13%)
MF	ribonuclease activity	GO:0004540	0.9876	0.03202	4/413(0.97%)	91/31473(0.29%)
BP	inositoltrisphosphate metabolic process	GO:0032957	0.9989	0.03333	2/319(0.63%)	22/24322(0.09%)
BP	ATP synthesis coupled electron transport	GO:0042773	0.9989	0.03333	2/319(0.63%)	22/24322(0.09%)
BP	polyol catabolic process	GO:0046174	0.9989	0.03333	2/319(0.63%)	22/24322(0.09%)

MF	phosphotransferase activity, carboxyl group as acceptor	GO:0016774	0.9944	0.036243	2/413(0.48%)	23/31473(0.07%)
BP	histone modification	GO:0016570	0.9993	0.03895	4/319(1.25%)	97/24322(0.40%)
MF	actin filament binding	GO:0051015	0.9963	0.03920	2/413(0.48%)	24/31473(0.08%)
BP	phosphatidylinositol phosphorylation	GO:0046854	0.9997	0.04063	3/319(0.94%)	58/24322(0.24%)
BP	response to cyclopentenone	GO:0010583	0.9998	0.04155	5/319(1.57%)	144/24322(0.59%)
BP	transition metal ion transport	GO:0000041	0.9999	0.04410	4/319(1.25%)	101/24322(0.42%)
BP	cell-cell signaling	GO:0007267	0.9999	0.04532	2/319(0.63%)	26/24322(0.11%)
BP	response to hypoxia	GO:0001666	0.9999	0.04853	2/319(0.63%)	27/24322(0.11%)

Notes:

*BP: biological process;

#MF: molecular function;

=CM: cellular component.

Supplementary Table 18. Positively selected genes related to stress tolerance.

Functional category	Term ID	Term name	Gene ID	Best hit in Arabidopsis	Arabidopsis ID name	Arabidopsis ID description
KEGG	ath00073	cutin, suberine and wax biosynthesis	Glyma10g38480	AT5G51950		Glucose-methanol-choline (GMC) oxidoreductase family protein
KEGG	ath00073	cutin, suberine and wax biosynthesis	Glyma07g01220	AT1G72970	EDA17,HTH	Glucose-methanol-choline (GMC) oxidoreductase family protein
KEGG	ath00073	cutin, suberine and wax biosynthesis	Glyma04g37470	AT5G41040	ASFT, HHT,	feruloyl-CoA transferase required for suberin synthesis
KEGG	ath00073	cutin, suberine and wax biosynthesis	Glyma01g42480	AT1G63020	NRPD1,NRPD1A,POLIVA,SDE4,SMD2	nuclear RNA polymerase D1A
GO	GO:0010374	stomatal complex development	Glyma13g44270	AT1G80080	AtRLP17,TMM	Leucine-rich repeat (LRR) family protein
GO	GO:0010374	stomatal complex development	Glyma02g14290	AT3G24140	FMA	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
GO	GO:0010374	stomatal complex development	Glyma01g23230	AT3G24140	FMA	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
GO	GO:0005249	voltage-gated potassium channel activity	Glyma17g12740	AT2G26650	AKT1,ATAKT1,KT1	K ⁺ transporter 1
GO	GO:0005249	voltage-gated potassium channel activity	Glyma05g08230	AT2G26650	AKT1,ATAKT1,KT1	K ⁺ transporter 1
GO	GO:0005249	voltage-gated potassium channel activity	Glyma04g24950	AT4G30360	ATCNGC17,CNGC17	cyclic nucleotide-gated channel 17
GO	GO:0006560	proline metabolic process	Glyma18g03830	AT3G55610	P5CS2	delta 1-pyrroline-5-carboxylate synthase 2
GO	GO:0006560	proline metabolic process	Glyma02g41850	AT3G55610	P5CS2	delta 1-pyrroline-5-carboxylate synthase 2
GO	GO:0006560	proline metabolic process	Glyma05g27360	AT5G4618	DELTA-OAT	ornithine-delta-aminotransferase
GO	GO:0015706	nitrate transport	Glyma05g06130	AT1G32450	NTR1.5	nitrate transporter 1.5

GO	GO:0015706	nitrate transport	Glyma16g06190	AT3G27170	ATCLC-B,CLC-B	chloride channel B
GO	GO:0015706	nitrate transport	Glyma09g37230	AT1G32450	NTR1.5	nitrate transporter 1.5

Supplementary Table 19. Summary statistics of *G. soja* pan-genome.

	Sequence length (bp)	No. of gene family
Shared in 7 samples	790,135,846	28,716
Shared in 6 samples	96,619,808	4,657
Shared in 5 samples	56,619,081	2,527
Shared in 4 samples	10,124,621	2,287
Shared in 3 samples	6,909,803	2,962
Shared in 2 samples	6,990,020	5,015
GsojaA specific	3,473,075	2,296
GsojaB specific	1,457,723	1,123
GsojaC specific	4,575,947	2,769
GsojaD specific	3,458,966	3,004
GsojaE specific	3,200,841	1,225
GsojaF specific	1,266,036	1,064
GsojaG specific	1,515,767	1,435

Supplementary Table 20. Variations in core and dispensable genes.

Variation type	Total		Frequency (sites/ per KB)		
	Core	Dispensable	Core	Dispensable	P-value (t-test)
SNP	1,538,464	271,301	2.67	4.12	<0.01
Indel	276,337	44,504	0.55	0.71	<0.01

Supplementary Table 21. Enrichment of core and dispensable genes in various GO categories.

Type	GO root	GO name	Frequency in core	Frequency in dispensable	FDR	p-value
Core	BP	biological regulation	0.215	0.192	0.00246	0.000783
Core	BP	cellular component organization or biogenesis	0.074	0.050	5.26E-06	4.98E-08
Core	BP	cellular process	0.688	0.610	0	2.35E-20
Core	BP	developmental process	0.091	0.039	0	3.33E-30
Core	BP	growth	0.019	0.006	0	6.55E-11
Core	BP	immune system process	0.012	0.006	0.000375	0.000109
Core	BP	multicellular organismal process	0.090	0.036	0	8.15E-33
Core	BP	multi-organism process	0.098	0.061	0	5.06E-14
Core	BP	negative regulation of biological process	0.011	0.006	0.00308	0.000982
Core	BP	regulation of biological process	0.204	0.185	0.009346	0.00366
Core	BP	reproduction	0.050	0.022	0	1.08E-16
Core	BP	reproductive process	0.049	0.021	0	6.91E-17
Core	BP	response to stimulus	0.250	0.174	0	8.58E-25
Core	BP	signaling	0.071	0.060	0.0153	0.00642
Core	BP	single-organism process	0.156	0.097	0	1.74E-22
Core	MF	binding	0.704	0.613	0	8.09E-36
Core	MF	enzyme regulator activity	0.017	0.012	0.013	0.00525
Core	MF	nucleic acid binding transcription factor	0.077	0.065	0.00664	0.00232

		activity				
Core	MF	transporter activity	0.071	0.060	0.00799	0.00288
Dispensable	MF	metabolic process	0.723	0.762	0	4.57E-07
Dispensable	MF	antioxidant activity	0.008	0.015	0.000189	1.12E-05
Dispensable	MF	nutrient reservoir activity	0.002	0.004	0.11	0.014
Dispensable	MF	receptor activity	0.009	0.014	0.0126	0.00101
Dispensable	MF	structural molecule activity	0.023	0.042	0	1.06E-13

Supplementary Table 22.Numbers of NBS-encoding genes predicted in seven *G. soja* and GmaxW82 genomes.

	GmaxW82	GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
NBS:TIR:LRRs	71	47	66	51	54	58	61	60
NBS:TIR:LRRs:WRKY	1	0	0	1	1	1	0	1
NBS:TIR:LRRs:MotA_activ	0	0	1	0	1	0	0	1
NBS:LRRs	116	76	91	81	101	91	84	74
NBS:LRRs:RPW8	4	4	4	5	4	5	4	4
NBS:LRRs:Peptidase_C48	3	1	2	1	1	2	1	2
NBS:LRRs:zf-BED	1	1	1	1	1	1	1	1
NBS:LRRs:Piptidase_C48:zf-BED	1	0	1	0	0	1	0	0
NBS:LRRs:Pkinase_Tyr	1	1	1	1	0	1	1	1
NBS:LRRs:PP2C	0	0	0	0	1	0	0	0
NBS:TIR	71	56	51	51	51	56	60	57
NBS:TIR:BSP	2	1	0	0	2	1	1	0
NBS:TIR:zf-RVT	1	1	1	1	1	1	1	1
NBS:TIR:RVT_3	1	0	1	0	0	0	1	0
NBS:TIR:RVT_3:zf-RVT	0	0	0	0	1	0	0	1
NBS:TIR:DYW_deaminase:PPR:PPR2	0	1	0	0	0	0	0	0
NBS:TIR:Glyco_hydro_20	0	0	0	0	0	0	0	1
NBS	183	140	151	152	158	148	155	149
NBS:RPW8	4	4	4	3	4	3	4	4
NBS:DUF616	0	1	1	0	0	0	0	0
NBS:RVT_3	0	0	0	1	1	0	0	0
NBS:BSP	0	0	0	1	0	1	0	0
NBS:Glyco_transf_8	0	0	0	0	0	1	0	0
NBS:RVP_2:Retrotrans_gag	0	0	0	0	0	1	0	0
NBS:P450	0	0	0	0	0	0	0	1
Total	460	334	376	350	382	372	374	358

Supplementary Table 23. R-genes with CNV identified among seven *G. soja* genomes and GmaxW82.

GeneID	Physical location	Type	No. of <i>G. soja</i> accessions with CNV	ID of homolog in Arabidopsis	Symbol of gene or QTL in soybean	Expression analysis	Related trait	Reference
Glyma01g04000	3,487,915-3,494,204	CNV-loss	3	AT5G17680.1		May work against broad range of pathogen attack		Kang et al. 2012, BMC Plant Biol
Glyma01g27440	36,685,736-36,695,527	CNV-loss	4	AT5G36930.2				
Glyma01g29495	39,818,463-39,819,244	CNV-loss	4	AT4G11170.1				
Glyma02g43630	48,381,734-48,387,902	CNV-loss	1	AT5G17680.1				
Glyma03g04030	3,991,872-3,996,475	CNV-gain	3	AT3G14460.1	RpsYD29		Phytophthora root rot (PRR)	Zhang et al. 2013, TheorAppl Genet
Glyma03g04080	4,035,138-4,041,387	CNV-gain	4	AT3G14460.1	RpsYD29		Phytophthora root rot (PRR)	Zhang et al. 2013, TheorAppl Genet
Glyma03g04100	4,075,016-4,079,699	CNV-gain	1	AT3G14460.1				
Glyma03g04121	4,099,246-4,103,215	CNV-loss	1	AT3G14470.1				
Glyma03g04140	4,154,349-4,159,039	CNV-loss	1	AT3G14460.1				
Glyma03g04180	4,176,524-4,180,146	CNV-loss	1	AT3G14470.1				
Glyma03g04200	4,250,529-4,257,697	CNV-loss	2	AT3G14460.1				
Glyma03g04560	4,732,588-4,737,220	CNV-loss	1	AT3G14460.1				
Glyma03g04610	4,784,580-4,789,034	CNV-loss	1	AT3G14460.1				
Glyma03g04780	4,981,933-4,986,668	CNV-gain	2	AT3G14460.1				
Glyma03g05290	5,567,902-5,571,304	CNV-gain	4	AT3G14460.1				
Glyma03g05400	5,646,458-5,650,182	CNV-gain /loss	2	AT3G14460.1				
Glyma03g05730	6,051,893-6,063,501	CNV-loss	1	AT5G41750.1				
Glyma03g05786	6,109,047-6,110,122	CNV-loss	5	AT3G14470.1				
Glyma03g06255	6,452,913-6,462,914	CNV-loss	2	AT5G17680.1				
Glyma03g06271	6,476,633-6,477,163	CNV-loss	1	AT4G12010.1				
Glyma03g06926	7,198,669-7,200,323	CNV-loss	1	AT5G36930.2				
Glyma03g06976	7,275,123-7,298,068	CNV-loss	2	AT1G27180.1				
Glyma03g07091	7,332,866-7,335,624	CNV-loss	2	AT5G36930.2				

Glyma03g07121	7,458,627-7,491,220	CNV-gain/loss	3	AT5G36930.2			
Glyma03g07181	7,546,780-7,575,160	CNV-loss	2	AT5G36930.2			
Glyma03g16224	20,614,092-20,615,227	CNV-loss	2	AT5G17680.1			
Glyma03g22120	28,100,313-28,104,905	CNV-loss	1	AT5G36930.2			
Glyma03g22151	28,137,514-28,142,464	CNV-loss	1	AT5G36930.2			
Glyma03g23185	29,772,156-29,773,961	CNV-loss	1	AT3G14470.1			
Glyma03g23215	29,798,439-29,800,412	CNV-loss	1	AT3G14470.1			
Glyma06g40690	43,857,936-43,861,836	CNV-gain	2	AT5G17680.1	Rpp?(Hyuuga)	Asian soybean rust (ASR)	Monteros et al. 2010, TheorAppl Genet
Glyma06g40710	43,882,233-43,886,906	CNV-gain	1	AT5G17680.1	Rpp?(Hyuuga)	Asian soybean rust (ASR)	Monteros et al. 2010, TheorAppl Genet
Glyma06g40740	43,913,600-43,918,380	CNV-loss	2	AT4G12010.1	Rpp?(Hyuuga)	Asian soybean rust (ASR)	Monteros et al. 2010, TheorAppl Genet
Glyma06g40780	43,967,782-43,972,339	CNV-gain	1	AT4G12010.1	Rpp?(Hyuuga)	Asian soybean rust (ASR)	Monteros et al. 2010, TheorAppl Genet
Glyma06g40980	44,268,825-44,273,222	CNV-gain	1	AT5G17680.1		May work against broad range of pathogen attack; highly expressed in the resistant lines	Bacterial leaf pustule (BLP) Kang et al. 2012, BMC Plant Biol Kim et al. 2011, DNA Res
Glyma06g41241	44,527,825-44,538,630	CNV-gain	1	AT5G17680.1			
Glyma06g41290	44,575,786-44,580,494	CNV-loss	1	AT5G17680.1			
Glyma06g41326	44,600,004-44,601,270	CNV-gain	3	AT5G17680.1			
Glyma06g41430	44,719,429-44,723,980	CNV-loss	1	AT5G17680.1			
Glyma06g41700	44,984,923-44,988,575	CNV-loss	4	AT5G17680.1			
Glyma06g41756	45,055,316-45,056,174	CNV-loss	5	AT5G36930.2			
Glyma06g41801	45,087,334-45,090,031	CNV-loss	5	AT5G17680.1			
Glyma06g41896	45,155,818-45,162,554	CNV-loss	1	AT3G25510.1			
Glyma06g46810	49,369,736-49,383,270	CNV-loss	1	AT3G07040.1	Homolog of RPM1	Significantly different expressed between resistant and susceptible NILs	Bacterial leaf pustule (BLP) Kang et al. 2012, BMC Plant Biol Kim et al. 2011, DNA Res

Glyma07g07100	5,717,035-5,737,954	CNV-gain	2	AT4G26090.1			
Glyma07g07110	5,756,107-5,779,579	CNV-gain	1	AT4G26090.1			
Glyma07g08214	6,824,048-6,827,468	CNV-loss	2	AT5G47260.1			
Glyma08g12560	9,199,265-9,203,838	CNV-loss	3	AT1G12280.1			
Glyma08g42930	42,872,278-42,875,132	CNV-loss	3	AT3G07040.1	Homolog of RPM1	Significantly different expressed between resistant and susceptible NILs	Kang et al. 2012, BMC Plant Biol
Glyma08g42971	42,890,484-42,894,906	CNV-loss	1	AT3G07040.1			
Glyma08g43020	42,908,118-42,917,722	CNV-loss	1	AT3G07040.1			
Glyma08g43530	43,336,391-43,349,138	CNV-gain	1	AT3G07040.1			
Glyma09g11883	12,222,577-12,223,212	CNV-loss	3	AT3G14470.1			
Glyma12g16450	15,730,002-15,734,533	CNV-loss	1	AT5G17680.1			
Glyma13g03770	3,846,578-3,852,431	CNV-loss	1	AT5G17680.1			
Glyma13g04230	4,576,019-4,581,410	CNV-gain	1	AT3G14470.1			
Glyma13g25780	29,007,441-29,013,112	CNV-gain	1	AT3G14470.1	Rsv1	Soybean Mosaic Virus (SMV)	Wen et al. 2013, Molecular plant-microbe interactions
Glyma14g37860	47,141,305-47,143,952	CNV-gain	3	AT3G46530.1	Rsv3	Soybean Mosaic Virus (SMV)	Suh et al. 2011, Plant Genome-US
Glyma15g37050	42,618,195-42,621,706	CNV-loss	4	AT3G14460.1			
Glyma15g37140	42,760,751-42,766,047	CNV-loss	1	AT3G14470.1			
Glyma15g37340	43,109,897-43,112,349	CNV-gain /loss	3	AT3G14470.1			
Glyma15g37390	43,164,159-43,168,071	CNV-gain /loss	3	AT3G14460.1			
Glyma15g39531	46,181,348-46,188,879	CNV-loss	5	AT4G27190.1			
Glyma15g39596	46,380,611-46,396,097	CNV-loss	1	AT4G27190.1			
Glyma16g10020	10,028,074-10,032,604	CNV-loss	2	AT5G36930.2			
Glyma16g10080	10,183,556-10,189,868	CNV-loss	1	AT5G36930.2			
Glyma16g17542	19,056,184-19,062,615	CNV-gain	7	AT5G41750.1			
Glyma16g23790	27,651,730-27,657,276	CNV-gain	3	AT5G36930.2			
Glyma16g23800	27,663,876-27,668,774	CNV-gain	3	AT5G17680.1			

Glyma16g25071	28,995,808-29,006,764	CNV-loss	2	AT5G17680.1
Glyma16g33616	36,479,474-36,484,843	CNV-gain	3	AT5G36930.2
Glyma16g33681	36,522,799-36,532,684	CNV-loss	1	AT5G36930.2
Glyma16g33780	36,560,125-36,565,634	CNV-gain	1	AT5G17680.1
Glyma16g33910	36,657,476-36,667,728	CNV-loss	1	AT5G36930.2
Glyma16g33930	36,671,048-36,674,931	CNV-gain /loss	3	AT5G36930.2
Glyma16g33941	36,676,886-36,680,457	CNV-gain /loss	3	AT5G36930.2
Glyma16g33950	36,685,930-36,690,094	CNV-gain	2	AT5G36930.2
Glyma16g33961	36,692,021-36,696,860	CNV-gain /loss	2	AT5G17680.1
Glyma16g33971	36,705,894-36,708,208	CNV-loss	1	AT5G17680.1
Glyma16g34000	36,720,280-36,724,095	CNV-loss	1	AT4G09430.1
Glyma16g34102	36,784,115-36,786,213	CNV-loss	1	AT5G17680.1
Glyma17g20995	20,051,606-20,056,795	CNV-loss	1	AT5G66900.1
Glyma17g21130	20,345,778-20,350,785	CNV-loss	1	AT5G66900.1
Glyma18g09130	7,950,238-7,954,223	CNV-loss	1	AT3G07040.1
Glyma18g09140	7,972,558-7,976,185	CNV-loss	3	AT3G07040.1
Glyma18g09170	7,993,323-7,997,136	CNV-loss	2	AT3G07040.1
Glyma18g09180	8,041,284-8,044,058	CNV-loss	4	AT3G07040.1
Glyma18g09220	8,096,590-8,100,424	CNV-loss	1	AT3G07040.1
Glyma18g09290	8,169,545-8,173,273	CNV-loss	1	AT3G07040.1
Glyma18g09314	8,202,191-8,203,672	CNV-loss	5	AT3G07040.1
Glyma18g09327	8,210,211-8,218,589	CNV-loss	2	AT3G07040.1
Glyma18g09340	8,223,483-8,226,314	CNV-loss	1	AT3G07040.1
Glyma18g09390	8,260,975-8,267,778	CNV-loss	6	AT3G07040.1
Glyma18g09410	8,278,329-8,282,216	CNV-loss	6	AT3G07040.1
Glyma18g09695	8,581,353-8,584,384	CNV-gain	1	AT3G07040.1
Glyma18g09720	8,614,114-8,617,241	CNV-gain /loss	2	AT3G07040.1
Glyma18g09744	8,623,657-8,625,341	CNV-loss	5	AT3G07040.1
Glyma18g09758	8,633,923-8,636,049	CNV-loss	5	AT3G50950.1

Rj2

Legume-rhizobia symbiosis

Yang et al. 2010,
ProcNatlAcidSci USA

Glyma18g09786	8,654,043-8,661,716	CNV-loss	5	AT3G07040.1
Glyma18g09800	8,671,218-8,675,034	CNV-loss	4	AT3G07040.1
Glyma18g09824	8,700,730-8,702,106	CNV-loss	4	AT3G07040.1
Glyma18g09850	8,712,094-8,715,692	CNV-loss	3	
Glyma18g09880	8,733,719-8,736,647	CNV-loss	3	AT3G07040.1
Glyma18g09901	8,764,841-8,766,616	CNV-loss	6	AT3G07040.1
Glyma18g10550	9,430,834-9,441,301	CNV-loss	1	AT3G07040.1
Glyma18g10670	9,526,324-9,528,288	CNV-loss	1	AT3G07040.1
Glyma18g46026	55,753,308-55,762,044	CNV-gain	2	AT4G27220.1
Glyma18g46050	55,780,756-55,823,919	CNV-gain	5	AT4G26090.1
Glyma18g46101	55,871,890-55,874,958	CNV-gain	5	AT4G27220.1
Glyma18g51765	60,543,561-60,551,568	CNV-loss	1	AT1G09730.1
Glyma18g51930	60,685,177-60,687,775	CNV-gain	1	AT3G46530.1
Glyma18g51950	60,693,444-60,696,291	CNV-gain	3	AT3G46530.1
Glyma18g51960	60,704,363-60,706,532	CNV-loss	1	AT3G46530.1
Glyma19g07700	9,229,498-9,234,274	CNV-loss	1	AT5G36930.2
Glyma19g31536	39,322,195-39,324,657	CNV-gain	1	AT3G14460.1
Glyma19g31674	39,473,386-39,476,650	CNV-loss	6	AT3G14470.1
Glyma19g31686	39,484,586-39,488,050	CNV-loss	7	AT3G14460.1
Glyma19g31698	39,505,790-39,509,002	CNV-gain	3	AT3G14460.1
Glyma19g31843	39,632,265-39,634,889	CNV-gain	5	AT3G14460.1
Glyma19g31856	39,638,873-39,639,934	CNV-gain	2	AT3G14470.1
Glyma20g08870	12,604,425-12,612,288	CNV-loss	5	AT3G14460.1
Glyma20g12720	18,088,175-18,095,180	CNV-loss	3	AT3G14460.1
Glyma20g12730	18,146,192-18,148,978	CNV-loss	2	AT3G14470.1
Glyma20g33740	42,256,803-42,260,407	CNV-loss	1	AT1G58602.1

Supplementary Table 24. Number of putative acyl lipid genes with genetic variation in *G. soja* accessions and GmaxW82.

Acyl lipid sub-pathways*	No. of putative acyl lipid genes in GmaxW82	No. of genes with CNV and/or large-effect SNP/indel and positively selected					P-value	
		Genes with CNV	Genes with large-effect SNP/indel	PSG#	Total	Percentage (%)	Whole genome ⁼	Acyl lipid pathway ^{&}
Cutin Synthesis & Transport 1	78	0	9	2	10	12.82	0.094	0.553
Cutin Synthesis & Transport 2	15	0	3	0	3	20	1.000	0.930
Eukaryotic Galactolipid&Sulfolipid Synthesis	60	0	6	0	6	10	0.049	0.277
Eukaryotic Phospholipid Synthesis & Editing	102	2	10	1	12	11.76	0.027	0.302
Fatty Acid Elongation & Wax Biosynthesis	298	12	40	2	49	16.44	0.052	0.816
Fatty Acid Elongation, Desaturation & Export From Plastid	26	0	2	0	2	7.69	0.148	0.380
Fatty Acid Synthesis	97	0	9	1	10	10.31	0.012	0.160
Lipid Trafficking	21	0	5	0	5	23.81	0.982	0.480
Mitochondrial Fatty Acid &Lipoic Acid Synthesis	23	1	4	0	4	17.39	0.846	1.000
Mitochondrial Lipopolysaccharide Synthesis	10	0	3	0	3	30	0.771	0.426
Mitochondrial Phospholipid Synthesis	30	0	4	0	4	13.33	0.404	0.899
Oxylipin Metabolism 1	174	6	34	2	37	21.26	1.000	0.047
Oxylipin Metabolism 2	135	6	26	2	29	21.48	1.000	0.077
Pathway, function or subcellular location uncertain	66	4	16	0	18	27.27	0.294	0.015
Phospholipid Signaling	181	0	24	3	25	13.81	0.019	0.487
Prokaryotic Galactolipid, Sulfolipid, & Phospholipid Synthesis 1	58	0	6	0	6	10.34	0.062	0.323
Prokaryotic Galactolipid, Sulfolipid, & Phospholipid Synthesis 2	33	0	3	0	3	9.09	0.136	0.404
Sphingolipid Biosynthesis 1	41	0	4	1	5	12.2	0.221	0.666
Sphingolipid Biosynthesis 2	24	0	1	0	1	4.17	0.073	0.194
Suberin Synthesis & Transport 1	63	0	7	2	9	14.29	0.233	0.865

Suberin Synthesis & Transport 2	24	0	3	1	4	16.67	0.767	1.000
Suberin Synthesis & Transport 3	23	0	3	0	3	13.04	0.481	0.934
Triacylglycerol & Fatty Acid Degradation	99	0	20	1	20	20.2	0.900	0.275
Triacylglycerol Biosynthesis	113	1	10	3	12	10.62	0.008	0.146

Notes:

* Referred from (Li-Beisson et al. 2013) (<http://aralip.plantbiology.msu.edu/pathways/pathways/>); Five sub-pathways with a significantly low frequency of variation were in bold.

PSG: positive selected genes;

= Compared the percentage of genes with CNV and/or large-effect SNP/indel and positively selected in each acyl lipid pathway with the whole genome (11,500/54,175=21.23%);

& Compared the percentage of genes with CNV and/or large-effect SNP/indel and positively selected in each acyl lipid sub-pathway with that acyl lipid pathway (211/1332=15.84%).

Supplementary Table 25. Large-effect mutations in *Glyma09g06950*, *Glyma09g07090*, *Glyma09g07206* and *Glyma09g07760*.

Position (Chromosome: Start Site - End Site)	Transcript ID	Mutated exon	Mutation Type	GmaxW82 Type	Mutated Type	Mutation status						
						GsojaA	GsojaB	GsojaC	GsojaD	GsojaE	GsojaF	GsojaG
9:5788213-5788213	Glyma09g06950.2	exon7	frameshift insertion	-	T	✓	✓	✓	✓	✓	✓	✓
9:5932689-5932689	Glyma09g07090.1	exon3	frameshift deletion	A	-	✓	✓	✓	✓	✓		✓
9:5932688-5932688	Glyma09g07090.1	exon3	frameshift deletion	A	-						✓	
9:6072822-6072823	Glyma09g07206.1	exon5	frameshift deletion	AT	-		✓	✓	✓	✓	✓	✓
9:6072841-6072843	Glyma09g07206.1	exon5	nonframeshift deletion	CAT	-		✓	✓	✓	✓	✓	✓
9:6072851-6072852	Glyma09g07206.1	exon5	frameshift deletion	AT	-		✓	✓	✓	✓	✓	✓
9:6072887	Glyma09g07206.1	exon5	stopgain SNP	C	T	✓					✓	
9:6695158-6695158	Glyma09g07760.2	exon2	nonframeshift insertion	-	GTG				✓			
9:6696072-6696072	Glyma09g07760.2	exon2	frameshift insertion	-	A	✓	✓	✓	✓	✓	✓	✓
9:6696083-6696083	Glyma09g07760.2	exon2	frameshift insertion	-	A	✓	✓	✓	✓	✓	✓	✓
9:6696085-6696085	Glyma09g07760.2	exon2	frameshift insertion	-	A	✓	✓	✓	✓	✓	✓	✓

Notes:

Glyma09g06950 encodes a nucleotide/sugar transporter family protein. Glyma09g07090 is a homolog of GATA21/GNC, encoding a GATA type zinc finger transcription factor family protein. Glyma09g07206 is a homolog of ATCKX2 (AT2G19500), encoding cytokinin oxidase 2. And Glyma09g07760 is involved in cell wall catabolism.