Supplementary Materials and Methods: Detailed materials and methods

Numbers of genes in the NBS and RLK families vary by more than four fold within a plant species and are regulated by multiple factors

Meiping Zhang[1,2], Yen-Hsuan Wu[1], Mi-Kyung Lee[1], Yun-Hua Liu[1], Ying Rong[1], Teofila S. Santos[1] , Chengcang Wu[1], Fangming Xie[3], Randall L. Nelson[4], and Hong-Bin Zhang[1,*]

[1] Department of Soil and Crop Sciences, Texas A&M University, College Station, Texas 77843-2474, USA; [2] College of Life Science, Jilin Agricultural University, Changchun, Jilin 130118, China; [3] International Rice Research Institute, DAPO BOX 7777, Metro Manila, Philippines; and [4] USDA-Agricultural Research Service, Soybean/Maize Germplasm, Pathology, and Genetics Research Unit, Dep. of Crop Sciences, 1101 W. Peabody Dr., University of Illinois, Urbana, IL 61801, USA

Meiping Zhang and Yen-Husan Wu contributed to this study equally.

*Corresponding author: H.-B. Zhang, Tel.: +1-979-862-2244; Fax: +1-979-845-0456;
E-mail: hbz7049@tamu.edu

Plant materials

A total of 187 lines or cultivars randomly selected from 57 species of *Oryza*, *Glycine* and *Gossypium* were used in this study (Supplementary Table S1). Of theses lines, 62 were from 12 *Oryza* species representing 8 diploids and 4 polyploids, or 2 cultivated species and 10 wild species, with 1 - 18 lines per species (Supplementary Table S1A); 28 from 10 *Glycine* species representing 1 cultivated species and 9 wild species, with 1 – 11 lines per species (Supplementary Table S1B); and 97 from 35 *Gossypium* species representing 30 diploids and 5 polyploids, or 4 cultivated species and 31 wild species, with 1 – 12 lines per species (Supplementary Table S1C).

DNA isolation and purification

The plants of each line were grown in a greenhouse and phenotypically verified during growth and development. Young leaves were collected from 1 - 5 plants of each line verified to be its representative. Nuclear DNA was isolated with a modified CTAB (cetytrimethylammonium bromide) method. Nuclei were first isolated in a buffer containing 350 mM sorbitol, 100 mM Tris, 5 mM EDTA and 0.38% (w/v) bisulfate, and then lysed to release nuclear DNA in the nuclei lysis buffer containing 0.2 M Tris.HCl, 50 mM EDTA, 2.0 M NaCl and 2% (w/v) CTAB. The DNA was purified with a chloroform/iso-amyl alcohol (24:1) mixture, collected by precipitation with isopropanol and dissolved in TE (10 mM Tris.HCl, 1 mM EDTA, pH 8.0). Furthermore, to minimize its effect on array preparation and gene copy number assay, RNA contaminated in the DNA was completely removed by treatment with RNase at 37^{o}C for 30 min, followed by extracting with phenol/chloroform/iso-amyl alcohol (25:24:1) mixture, precipitation with ethanol

and washing twice in 70% ethanol. DNA pellet was dried and dissolved in TE and the DNA concentration determined with a nanodrop spectrophotometer and a fluorometer, followed by verification on agarose gels.

Assay for numbers of genes in the NBS and RLK families

Methodology: Several methods have been used to assay the number of genes in a gene family in a genome. These include [1] whole-genome sequence blast analysis (WSBA) (1-3,6-10,14,15,19-21, 40), [2] membrane array (MA) (30), [3] microarray (M) (27), [4] random genomic clone sequencing (RGCS) (16,28), and [5] quantitative real-time PCR (qrtPCR) (29,31) (Supplementary Table S2). For the WSBA method, high-quality whole-genome sequence is generated, assembled into contigs and annotated. The number of genes in a gene family is assayed by iterative blast analysis of the sequence using the targeted gene nucleotide or amino acid sequence as inquiries. For the MA method, arrays are fabricated by printing total nuclear DNA or cDNA of target lines onto nylon membrane and probed with overgos designed from target genes. The hybridization signals or intensities of the samples are quantified using a PhosphorImager and the copy number of the genes in a genome is calculated using the hybridization signal data and the known copy number of the target genes. For the microarray method, arrays are fabricated by printing or synthesizing *in situ* gene-specific oligos or large-insert DNA clones (such as BAC) on a chemically-coated glass slide and probed with total genomic DNA or cDNA. The hybridization signals or intensities of the target elements are quantified using a microarray analyzer and the copy number of a gene or sequence in a genome is calculated as above for the MA method. For the RGCS method, shotgun or other types of DNA libraries having gene-length insert sizes are constructed and all clones or a sample of clones of the library are sequenced. The number

of genes in a gene family in a genome is calculated based on the number of clones sequenced, number of clones discovered to contain the target gene, clone insert size and genome coverage, and genome size. For the qrtPCR method, target gene-specific probes and/or primers are designed and used to amplify the DNA of a target genome and the copy number of a gene in the genome is determined based on fluoresce signal or intensity of the PCR product using the known copy number of the target element as a reference. In conclusion, these five methods are divided into two types. One is based on copy number counting, including the WSBA and RGCS methods, and the other is based on hybridization or fluorescent intensity, including the microarray, MA and qrtPCR methods.

We first evaluated the methods according to their sensitivity, reliability/reproducibility, and throughput/cost (Supplementary Table S2). It seems from previous studies (2,21,27-31,40) that all five methods have reasonable sensitivity, reproducibility and reliability that allow to determine the copy number change of a gene family in a genome; however, they have advantages and disadvantages in different aspects. The MA method is readily fabricated, simple and economical, and could be readily repeated. It is suitable for estimation of gene copy numbers with a range from a few to thousands of copies. The microarray and qrtPCR methods are often used to estimate the copy number of genes in a genomic region or that are smaller in the genome such as dozens of copies. Nevertheless, the results of the methods could be more significantly influenced by the nucleotide sequence homology among the members of genes and assay stringency, especially the qrtPCR method because it needs target-highly specific primers and/or probes for its reaction. Such primers or probes may make qrtPCR difficult to equally amplify all members of the targeted genes that may somehow differ at the sequence level in a gene family. The WSBA method has been widely used to estimate the number of genes in

a gene family in a genome (1-3,6-10, 14,15,19-21,40); however, its reliability is often subjected to the source sequence genome coverage, sequence assembly accuracy (particularly those of sequence-identical gene member assembly) and annotation accuracy (3,15,19-21). Therefore, this method often leads to underestimation of gene copy number of a gene family. It is also expensive to sequence a large number of genomes even though the new-generation high-throughput sequencing technology is used and it is rare to sequence a genome repeatedly for statistical analysis. The RGCS method reduces the sequencing cost and makes it feasible to sequence a large number of genomes, but the number of genes in a family is estimated based on sample sequencing and is more sensitive to original copy number counting because it is often subjected to the insufficient sequence read length and will be further modified to obtain the number of genes in the family in the entire genome (28).

Experimental design and pilot experiments: According to the above analysis, we chose the MA method as one candidate method and proposed another method named small-insert DNA library screening (SDLS) for this study. The MA method is similar to the microarray and qrtPCR methods in principle, all of which are hybridization or fluorescent intensity-based, but much simpler, more economical, and more readily repeatable, without having to sequence and annotate the entire genomes. The SDLS method takes the advantages of the sequence-based WSBA and RGCS methods in their countability of individual gene copy number change, but is much simpler and more economical. We conducted two pilot experiments to test the methods. The first pilot experiment was to test the MA and SDLS methods using 7 of the 187 lines and compared the results with those of rice cv. Nipponbare estimated by the WSBA method (15,19-21). *Oryza sativa* ssp. *japonica* cv. Nipponbare, *O. sativa* ssp. *indica* cv. Teqing and *O. rufipogon* acc.

PI590422 were selected for *Oryza* and *G. max* cv. Forrest for *Glycine*. This was because the numbers of genes in the NBS and RLK families were available for Nipponbare (15,19-21) and *O. sativa* ssp. *indica* cv. 93-11 (22) obtained by the WSBA method and we had the large-insert BAC library screening results for Nipponbare, Teqing and Forrest so that the results could be verified. PI590422 was a random sample. For *Gossypium*, *G. hirsutum*, *G. herbaceum* and *G. raimondii* were selected from its phylogenetic tree, each presenting a major lineage, the AD-genome, A+B+C+E+F+G+K-genome or D-genome lineage and the latter two being considered to be the donor diploid species of the former one.

For the MA method, 1,000 ng of the purified nuclear DNA for *Oryza*, 2,000 ng for *Glycine*, and 320, 640 and 960 ng for *Gossypium* D-genome, A-, B-, C-, E-, F-, G- and K-genome and AD-genome species, respectively, were printed per dot onto Hybond N+ membrane using an array blotting apparatus as described by the manufacturer (Bio-Rad, USA). To remove the potential noise background and determine the copy number of the target genes, two control groups were included in the arrays. The first group consisted of NBS and RLK family representative genes or degenerate overgos of known copy number as the positive controls and references to estimate the number of genes in the families in each genome. Furthermore, to minimize the potential effect of the plot plateau of hybridization intensity versus copy number on the assay, three levels of copy numbers for each target gene were applied on the array, with 50, 500 and 5000 copies per dot for NBS genes and 100, 1000 and 5,000 copies for RLK genes. Moreover, the hybridized arrays were exposed from 10 min to 5 hours to optimize the hybridization intensities of the samples for copy number estimation. The second group consisted of the printing buffer and non-homologous (salmon sperm) DNA as the negative controls to

remove the noise background. Six to eight sets of arrays for the experimental samples were fabricated independently for experimental replication.

To minimize the influence of gene member sequence homology on the gene number assay, for *Oryza* and *Glycine* we designed the NBS and RLK-specific degenerate overgos from the amino acid sequences of their conserved regions and verified them using 10 or more NBS or RLK genes cloned from the genera, followed by blast against GenBank. For *Gossypium*, we randomly chose 16 NBS genes cloned from *G. hirsutum* that represent all subfamilies of the cotton NBS family (53) (Supplementary Table S3). Subfamily-specific primers were used to amplify the genes using their clones as templates. The expected PCR products of the genes were purified on agarose gels. The *Oryza* and *Glycine* NBS and RLK degenerate overgos and the mixture of the 16 *Gossypium* NBS gene PCR products were used as the copy number positive control for array preparation as described above and as the probes for array hybridization and small-insert DNA library screening described below to estimate the number of genes in the families.

The MA hybridization experiment was conducted using the NBS and RLK degenerate overgos as probes for *Oryza* and *Glycine* and the combined PCR products of the 16 NBS genes as a probe for *Gossypium*. To minimize the potential influence of the hybridization stringencies on the copy number assay, we tested different hybridization stringencies, especially the washing stringencies, including high (0.2 x SSC), moderate (0.5 x SSC) and low (2 x SSC) stringencies. The PhosphoImager Bio-Imaging Analyzer BAS-1800II was used to quantify the hybridization signal or intensity of each line with the probes. The number of genes in the families in the genome of each line was estimated by comparing the hybridization intensity of each sample with that of the target gene positive controls of known copy number that were printed on the same array and had hybridization intensities closest to those of

the samples for *Oryza* and *Gossypium*. Since the hybridization of the copy number positive controls of target genes was unclear on some technical replicates for the *Glycine* arrays, the numbers of genes in the NBS and RLK families in the cultivar Forrest determined by the SDLS method (Supplementary Table S4C) were used as the positive controls and copy number reference for the estimation of numbers of genes in the families in other *Glycine* species lines. Comparative analysis showed that the numbers of the genes in the two families obtained from the three washing stringencies were identical or extremely similar; hence, the moderate washing stringency (0.5 x SSC) was used in this study. The MA experiment was replicated for 4 - 8 times so that the data could be analyzed by statistical tools (below).

For the SDLS method, two types of small-insert DNA libraries, shotgun-like and regular, were constructed from the nuclear DNA of the selected genotypes. The shotgun-like DNA libraries were constructed from the DNA partially digested with three 4-bp, blunt-ended enzymes, *Alu*I, *Hae*III and *Rsa*I, simultaneously, so that the resultant libraries would have a genome coverage as does a shotgun library constructed from DNA fragments physically sheared. The partially digested DNA was selected on an agarose gel, and the DNA fragments that best reflected the average size of target genes were selected and cloned in the *Eco*RV site of the pGEM5 vector. The regular DNA libraries were constructed from the DNA partially digested with one 4-bp enzyme, *Mbo*I, and size-selected on an agarose gel in the *Bam*HI site of the pUC18 vector (Supplementary Table S4). The titers, percentages of clones containing inserts and insert sizes of the libraries were determined by plating on agar selective medium and insert analysis of approximately 100 random clones of each library on agarose gels.

The libraries were blotted onto Hybond-N+ nylon membrane as described by the manufacturer (Amersham, USA) and screened by hybridization, as described above, using the same *Oryza* and *Glycine* NBS and RLK degenerate overgos or the 16 selected *Gossypium* NBS genes as probes, as those used in the MA hybridization. The actual number of positive clones were counted and used to calculate the number of genes in the gene families in the genomes of the genotypes (Supplementary Table S4). From this experiment we expected to get a second estimation of number of the genes in the selected genotypes, which were used to further verify the results of the MA hybridization, thus providing additional, independent references for estimation of the number of genes in the families.

Supplementary Table S5 shows the numbers of genes in the NBS and RLK families in the selected genotypes estimated by the MA and SDLS methods. The result showed that similar numbers of genes in the families were obtained for a line by the two methods ($r = 0.9656$, $p < 0.001$), with a different range from 0.51 – 13.6%. This percentage was comparable to the artificial variation (17.5%) in number of NBS genes in Nipponbare estimated by the WSBA method using its whole genome sequence (47) by different researchers (15,19-21). The numbers of NBS genes obtained between the MA (679) and WSBA (597) methods for Nipponbare differed by 13.7%. The difference was close to the above 13.6% or 17.5% range and probably due to the incomplete genome coverage of the rice sequence (47), improper assembly of the sequence-identical gene members and/or stringency and method of blast (3). Moreover, the results obtained from both MA and SDLS experiments agreed with the above large-insert BAC library screening results, all suggesting that Teqing have more NBS genes than Nipponbare. Therefore, we concluded that the MA, WSBA and SDLS methods have a similar sensitivity and reliability for our research purposes.

Furthermore, we conducted the second pilot experiment to further test the reliability and reproducibility of the MA method using the DNA of two rice cultivars, Nipponbare and Teqing, with five plants from each cultivar and 8 replicates for the entire pilot experiment as described above. Statistical analysis showed that no significant difference in number of genes in either NBS or RLK family was detected with the MA method among the replicates and individual plants of either cultivar ($P = 0.305 \sim 0.650$), confirming that the MA method is reliable and reproducible for our research purpose. In addition, among the methods used to date (Supplementary Table S2) the MA method is the most cost-efficiency and most readily repeated, thus being well suited for statistical analysis by which experimental errors, if any, could be excluded from the analysis. Hence, we decided to use the MA method to measure the numbers of genes in the NBS and RLK families in the genomes of all 187 lines collected. The experimental design was as described above and the experiment was technically duplicated 4 – 8 times. Biological replicates or multiple plants per line was not included because the pilot experiment in rice cv. Teqing and cv. Nipponbare showed that there was no significant variation in the number of genes in both NBS and RLK families among different plants of a cultivar.

Data analysis

Since the MA experiment was technically replicated for 4 – 8 times, statistical analyses could be conducted so that experimental errors, if any, resulting from the technical issues could be excluded from the comparative analysis between and among lines or species. ANOVA, Pearson's correlation and t-test were used at two-tailed significance levels. Since some data were some apart from the expected normal distribution, the measured numbers of genes in the families were transformed into $\log_{10}$-number of genes before

the analyses. Nevertheless, it should be pointed out here that identical or very similar results of significance levels were obtained when the numbers of genes or hybridization signals were directly used in the statistical analyses. Therefore, use of the hybridization signal data, actual numbers of genes and the $\log_{10}$-number of genes would not influence the conclusions of this study in the variation and evolution of the two gene families. The analyses were performed using the computer statistical program SPSS (Statistical Package for the Social Sciences).

Supplementary Table S1A. Numbers of genes in the NBS and RLK families in the genomes of different accessions or cultivars of *Oryza* species estimated by membrane array.

Species	Genome	Genome size (pg/1C) [a]	Accession/cultivar	Geographical origin	Mean number of genes in the NBS family (No./1C) [b]			Mean number of genes in the RLK family (No./1C) [b]		
					Acc. or cultivar	Species	Genome	Acc. or cultivar	Species	Genome
O. sativa	AA	0.46				664.91	617.74		1140.29	1106.42
O. sativa ssp. indica	AA	0.46	PI160476	Philippines	945.85	785.43		1392.02	1416.53	
			PI220255	Malaysia	462.18			1092.69		
			PI276487	India	570.34			857.83		
			PI281860	Sri Lanka	857.23			1271.64		
			PI402734	Philippines	845.84			1527.47		
			Teqing	China	1031.14			1604.73		
O. sativa ssp. japonica	AA	0.46	PI168945	Spain	328.32	575.20		775.58	965.96	
			PI224890	Japan	511.38			1104.02		
			PI224926	Japan	745.94			1057.65		
			PI202989	Japan	757.81			1183.33		
			Nipponbare	Japan	679.60			1120.04		
			Lemont	USA	428.14			747.70		
O. sativa ssp. javanica	AA	0.46	CIor3880	Philippines	356.20	634.10		863.55	1038.38	
			CIor7203	Indonesia	514.01			1020.62		
			PI184499	Indonesia	733.10			1205.17		

Supplementary Table S1A (continued)

Species	Genome	Genome size (pg/1C) [a]	Accession/cultivar	Geographical origin	Mean number of genes in the NBS family (No./1C) [b]			Mean number of genes in the RLK family (No./1C) [b]		
					Acc. or cultivar	Species	Genome	Acc. or cultivar	Species	Genome
			PI402749	Indonesia	445.38			1176.47		
			PI403074	Indonesia	635.83			730.51		
			PI403405	Indonesia	1120.06			1233.96		
O. rufipogon	AA	0.46	PI239671		687.31	629.16		1293.32	1003.96	
			PI590418		601.21			1032.42		
			PI590422		598.96			686.15		
O. barthii	AA	0.47	PI236393	Guinea	1000.28	615.12		2019.78	1232.18	
			PI237987	Sierra Leone	493.44			1172.05		
			PI590399	U.S.A.	623.77			915.73		
			PI590400	U.S.A.	342.97			821.16		
O. glaberrima	AA	0.37	PI231194		509.34	467.46		719.11	981.50	
			PI232853		588.06			1189.35		
			PI369480	Ghana	360.14			743.95		
			PI450221	Liberia	574.51			968.74		
			PI450291	Liberia	388.43			806.67		
			PI450333	Liberia	298.75			1151.56		
			PI369447	Liberia	552.96			1291.15		
O. punctata	BB	0.43	IRGC105986	Cameroon	196.50	238.07	238.07	410.75	491.60	491.60
			IRGC104154	Cameroon	328.03			710.86		

Supplementary Table S1A (continued)

Species	Genome	Genome size (pg/1C) [a]	Accession/cultivar	Geographical origin	Mean number of genes in the NBS family (No./1C) [b]			Mean number of genes in the RLK family (No./1C) [b]		
					Acc. or cultivar	Species	Genome	Acc. or cultivar	Species	Genome
			IRGC105981	Cameroon	218.69			540.35		
			IRGC105982	Cameroon	254.90			421.38		
			IRGC105984	Cameroon	192.24			374.66		
O. eichinggeri	CC	0.69	IRGC81803	Sri Lanka	185.32	175.67	159.23	747.10	514.56	403.52
			IRGC81804	Sri Lanka	189.94			547.76		
			IRGC99567	Tanzania	188.46			430.46		
			IRGC101425	Uganda	183.75			340.29		
			IRGC101442	Sri Lanka	130.87			507.17		
O. officinalis	CC	0.64	IRGC100947	India	141.12	209.37		253.65	433.66	
			IRGC105081	Myanmar	219.39			259.40		
			IRGC105088	Malaysia	277.29			870.27		
			IRGC105090	Malaysia	199.69			351.30		
O. rhizomatis	CC	0.96	IRGC103421	Sri Lanka	118.30	92.63		215.44	201.92	
			IRGC105448	Sri Lanka	81.62			107.82		
			IRGC105660	Sri Lanka	77.99			282.51		
O. punctata	BBCC	0.98	IRGC88824	Madagascar	183.13	133.92	149.95	237.22	249.86	345.61
			IRGC88826	Madagascar	138.83			185.40		
			IRGC88827	Madagascar	152.22			378.46		
			IRGC99571	Tanzania	74.77			132.49		

Supplementary Table S1A (continued)

Species	Genome	Genome size (pg/1C) [a]	Accession/cultivar	Geographical origin	Mean number of genes in the NBS family (No./1C) [b]			Mean number of genes in the RLK family (No./1C) [b]		
					Acc. or cultivar	Species	Genome	Acc. or cultivar	Species	Genome
			IRGC99572	Tanzania	120.64			315.72		
O. minuta	BBCC	0.90	IRGC101079	Philippines	153.54	165.98		551.40	441.36	
			IRGC101089	Philippines	128.35			161.17		
			IRGC101097	Philippines	180.98			645.14		
			IRGC101098	Philippines	212.67			582.73		
			IRGC101099	Philippines	154.37			266.37		
O. alta	CCDD	1.01	PI590398	USA	235.71	314.32	324.90	470.53	748.57	795.63
			PI590397		392.93			1026.60		
O. latifolia	CCDD	0.94	PI269727	Nondueas	335.49	335.49		842.69	842.69	

[a] The genome size data are from Miyabayashi et al. (45).

[b] The mean number of genes in each accession or cultivar was calculated from six replicates for the NBS family and four replicates for the RLK family.

Supplementary Table S1B. Numbers of genes in the NBS and RLK families in the genomes of different accessions or cultivars of *Glycine* species estimated by membrane array.

Species	Genome	Genome size (pg/1C) [a]	Accession/cultivar	Geographical origin/source	Mean number of genes in the NBS family (No./1C) [b]			Mean number of genes in the RLK family (No./1C) [b]		
					Acc. or cultivar	Species	Genome	Acc. or cultivar	Species	Genome
G. soja	GG	1.15	PI339871A	South Korea	2084.31	1319.10	1181.62	1988.17	1296.90	1202.11
			PI393551	Taiwan	1660.83			1687.34		
			PI407027	Japan	837.30			947.79		
			PI407140	Japan	1262.70			1273.14		
			PI407275	South Korea	1282.95			1259.24		
			PI458538	China	764.72			823.34		
			PI464935	China	1670.35			1600.96		
			PI483464A	China	1354.14			1233.24		
			PI483465	China	954.67			858.89		
G. max	GG	1.13	Forrest	USA	1427.00	1044.14		1704.00	1107.32	
			Williams 82	USA	501.42			596.97		
			PI291312	China	1032.14			950.98		
			PI567361	China	1087.81			1172.77		
			PI567452	China	858.46			1031.86		
			PI567503	China	616.66			811.50		
			PI594604	China	1179.40			1253.87		

Supplementary Table S1B (continued)

Species	Genome	Genome size (pg/1C) [a]	Accession/cultivar	Geographical origin/source	Mean number of genes in the NBS family (No./1C) [b]			Mean number of genes in the RLK family (No./1C) [b]		
					Acc. or cultivar	Species	Genome	Acc. or cultivar	Species	Genome
			PI605770	Vietnam	551.81			653.84		
			PI408342	South Korea	1801.10			1725.95		
			PI417194	Japan	1260.28			1218.72		
			PI605829	Vietnam	1169.50			1060.08		
G. latifolia	BB	1.08	PI321393	Australia	1252.56	1252.56	1180.41	1500.39	1500.39	1351.02
G. microphylla	BB	1.00	PI339664	Australia	1986.80	1986.80		2242.84	2242.84	
G. stenophita	BB	1.00	PI546986	Australia	301.87	301.87		309.85	309.85	
G. arenaria	HH	1.20	PI505204	Australia	695.04	695.04	923.36	906.58	906.58	953.46
G. pindanica	HH	1.00	PI595818	Australia	1151.67	1151.67		1000.34	1000.34	
G. cyrtoloba	CC	1.33	PI604471	Australia	852.30	852.30	852.30	976.04	976.04	976.04
G. falcata	FF	1.65	PI612234	Australia	502.23	502.23	502.23	617.15	617.15	617.15
G .canescens	AA	0.95	PI483192	Australia	1140.50	1140.50	1140.50	1191.41	1191.41	1191.41

[a] The genome size data are from Bennett and Leitch (43).

[b] The mean number of genes in each accession or cultivar was calculated from eight replicates.

Supplementary Table S1C. Numbers of genes in the NBS family in the genomes of different accessions or cultivars of *Gossypium* species estimated by membrane array.

Species	Genome	Genome size (pg/1C) [a]	Accession /cultivar	Geographical origin	Acc. or cultivar mean [b] (No. /1C)	Species mean (No. /1C)	Genome mean (No. /1C)
G. sturtianum	C1	2.060	C1-4	Australia	170.04	220.78	380.07
			C1-7	Australia	271.52		
G. nandewarense	C1-n		C1-n-5	Australia	330.77	539.37	
			C1-n-6	Australia	747.96		
G. costulatum	K		C5-3	Australia	406.77	362.96	459.53
			C5-4	Australia	319.14		
G. nobile	K	2.840	NWA35	Australia	397.82	397.82	
G. pulchellum	K		C8-1	Australia	693.48	693.48	
G. marchantii	K	2.675	NWA-6	Australia	383.87	383.87	
G. australe	G	1.875	C3-1	Australia	428.97	472.44	423.27
			C3-4	Australia	515.90		
G. nelsonii	G	1.795	C9-1	Australia	516.07	434.61	
			C9-2	Australia	353.15		
G. bickii	G1	1.795	G1-1	Australia	336.95	362.76	
			G1-3	Australia	388.57		
G. thurberi	D1	0.860	D1-1	Mexico	281.48	523.50	445.59
			D1-7	Mexico	765.52		
G. trilobum	D8	0.870	D8-7	Mexico	1167.19	1710.02	
			D8-8	Mexico	1999.34		
			D8-9	Mexico	1963.52		

Supplementary Table S1C (continued)

Species	Genome	Genome size (pg/1C) [a]	Accession /cultivar	Geographical origin	Acc. or cultivar mean [b] (No. /1C)	Species mean (No. /1C)	Genome mean (No. /1C)
G. davidsonii	D_{3d}	0.930	D_{3d}-1	Mexico	110.99	90.60	
			D_{3d}-2	Mexico	70.20		
G. klotzschianum	D_{3-k}	0.900	D_{3-k}-57	Ecuador	174.81	123.27	
			D_{3-k}-58	Ecuador	132.44		
			D_{3-k}-59	Ecuador	62.55		
G. armourianum	D_{2-1}	0.875	D_{2-1}-7	Mexico	187.42	319.78	
			D_{2-1}-9	Mexico	452.14		
G. harknessii	D_{2-2}	0.930	D_{2-2}-4	Mexico	816.26	816.26	
G. turneri	D10	0.930	D10-1	Mexico	1017.46	909.83	
			D10-2	Mexico	802.19		
G. aridum	D4	0.940	D4-5	Mexico	185.65	185.65	
G. lobatum	D7	0.955	D7-4	Mexico	159.49	136.08	
			0208082.07	Mexico	112.66		
G. laxum	D9	0.955	D9-3	Mexico	210.24	189.73	
			0208021.08	Mexico	169.22		
G. schwendimanii	D11		D11-1	Mexico	88.39	88.39	
G. gossypioides	D6	0.860	D6-6	Mexico	269.81	269.86	
			0208082.05	Mexico	269.91		
G. raimondii	D5	0.900	D5-3	Peru	412.10	429.72	
			D5-6	Peru	550.13		
			D5-8	Peru	326.94		

Supplementary Table S1C (continued)

Species	Genome	Genome size (pg/1C) [a]	Accession /cultivar	Geographical origin	Acc. or cultivar mean [b] (No. /1C)	Species mean (No. /1C)	Genome mean (No. /1C)
G. herbaceum	A1	1.705	A1-108		425.83	900.30	1026.49
			A1-111		1361.61		
			A1-120		1184.23		
			A1-127		786.31		
			A1-128		449.68		
			A1-129		268.56		
			A1-153		511.19		
			A1-154		1370.49		
			A1-172		1179.37		
			A1-180		1465.77		
G. arboreum	A2	1.749	0208083.10		862.06	1152.68	
			A2-142		1147.00		
			A2-47		881.51		
			A2-84		1720.13		
G. anomalum	B1	1.390	B1-1	Africa	279.41	267.22	348.95
			B1-7	Africa	255.02		
G. capits-virdis	B3	1.375	B3-1	Portugal	430.68	430.68	
G. longicakyx	F1	1.340	F1-1	Tanzania	224.70	190.23	190.23
			F1-4	Tanzania	155.75		
G. stocksii	E1	1.565	E1-3	Arabia	386.28	389.80	332.89

Supplementary Table S1C (continued)

Species	Genome	Genome size (pg/1C) [a]	Accession /cultivar	Geographical origin	Acc. or cultivar mean [b] (No. /1C)	Species mean (No. /1C)	Genome mean (No. /1C)
			E1-4	Arabia	393.31		
G. areysianum	E3	1.700	E3-1	Arabia	328.16	328.16	
G. incanum	E4		0208081.07		314.28	280.73	
			E4-4		247.17		
G. hirsutum	(AD)1	2.464	TM-1		1713.07	1347.71	973.04
			Wild Mexico Jack Jones		2160.79		
			Clevewilt 6		885.55		
			Auburn 56		1493.21		
			Stoneville 213		1479.36		
			Coker 201		758.24		
			Coker 310		1419.06		
			Deltapine 16		1182.92		
			Deltapine 61		1037.18		
G. barbadense	(AD)2	2.505	Pima S6		723.89	710.23	
			3-79		1005.91		
			(AD)2-81		573.70		
			(AD)2-372		659.08		
			K101		588.56		
G. tomentosum	(AD)3	2.435	(AD)3-15		549.74	803.42	
			(AD)3-16		620.50		

Supplementary Table S1C (continued)

Species	Genome	Genome size (pg/1C) [a]	Accession /cultivar	Geographical origin	Acc. or cultivar mean [b] (No. /1C)	Species mean (No. /1C)	Genome mean (No. /1C)
			(AD)3-17		718.36		
			(AD)3-25		741.97		
			0208081.05		762.63		
			(AD)3-26		628.80		
			(AD)3-1		933.96		
			(AD)3-3		793.65		
			(AD)3-4		942.22		
			(AD)3-5		974.52		
			(AD)3-7		934.01		
			(AD)3-11		1040.68		
G. mustelinum	(AD)4	2.425	0208082.04		1359.47	1353.05	
			(AD)4-9		1027.31		
			(AD)4-7		1672.38		
G. darwinii	(AD)5	2.415	(AD)5-3		562.42	650.77	
			(AD)5-7		739.11		

[a] The genome size data are from Hendrix and Stewart (44).

[b] The mean number of genes in each accession or cultivar was calculated from six replicates.

Supplementary Table S2. Comparison of the methods used for estimation of number of genes in a gene family.

Methods	Sensitivity	Reproducibility	Factors potentially affecting result accuracy	Cost	Reference
1. Whole-genome sequence blast analysis (WSBA)	Gene single-copy change	Impractical to repeat the experiment	Sequence genome coverage, sequence assembly accuracy, annotation, gene member sequence homology, and blast strategy and stringency	High	2
2. Membrane array (MA)	Gene single-copy change	Yes	Gene member sequence homology, and array hybridization stringency	Low	30
3. Microarray (M)	Gene single-copy change	Yes	Element genome coverage, gene member sequence homology, and array hybridization stringency	Moderate	27
4. Random genomic clone sequencing (RGCS)	Gene single-copy change	Yes	Clone insert size and genome coverage (sample size), sequence read length, sequence assembly accuracy, annotation, gene member sequence homology, and blast strategy and stringency	Moderate	28
5. qrtPCR	Gene single-copy change	Yes	Gene member sequence homology and PCR stringency	Fair	29
6. Small-insert DNA library screening (SDLS)	Gene single-copy change	Yes	Library insert size and genome coverage (sample size), gene member sequence homology, and library screening stringency	Fair	This study

Supplementary Table S3. Probes used in estimation of the number of genes in the NBS and RLK families

Oryza

1. NBS family:

NBS gene-specific Overgo: LSIVGMGGLGKTTL (NBS domain)

Degenerate nucleotide oligo: 5'-(T/C)TN(T/A)(C/G)NAT(T/C/A)GTNGGNATGGGNGGN(T/C)TNGGN AA(A/G)ACNACN(T/C)TN-3'

Overgo-A: 5'-(T/C)TN(T/A)(C/G)NAT(T/C/A)GTNGGNATGGGNGGN(T/C)-3'

Overgo-B: 5'-NA(G/T)NGTNGT(C/T)TTNCCNA(G/A)NCCNCCC-3'

PCR products of the NBS genes used as probes for BAC library screening:

r2 (AF032689), r3 (AF032690), r4 (AF032691), r5 (AF032692), r7 (AF032694), r8 (AF032695), r9 (AF032696), r12 (AF032699), r13 (AF032700), r16 (AF032703)

2. RLK family:

EKSDIYSFGVVLLE (Domain-IX)

Degenerate nucleotide oligo: 5'-GA(A/G)AA(A/G)(T/A)(G/C)NGA(T/C)AT(T/C/A)TA(T/C)(T/A)(G/C) NTT(T/C)GGNGTNGTN(T/C)TN(T/C)TNGA(A/G)-3'

Overgo-A: 5'- GA(A/G)AA(A/G)(T/A)(G/C)NGA(T/C)AT(T/C/A)TA(T/C)(T/A)(G/C)NTT(T/C)G-3'

Overgo-B: 5'-(C/T)TCNA(G/A)NA(G/A)NACNACNCC(G/A)AAN(G/C)(A/T)(G/A)-3'

Glycine

1. NBS family:

nTIR class: GMGGVGKTTLAQHV (NBS domain)

Degenerate nucleotide oligo: 5'-GGNATGGGNGGNGTNGGNAA(A/G)ACNACN(T/C)TNGCN CA(A/G)CA(T/C)GTN-3'

Overgo-A: 5'-GGNATGGGNGGNGTNGGNAA(A/G)ACNA-3'

Overgo-B: 5'-NAC(A/G)TG(T/C)TGNGCNA(A/G)NGTNGT(T/C)TTN -3'

TIR class: GMGGVGKTTLARAV (NBS domain)

Degenerate nucleotide oligo: 5'-GGNATGGGNGGNGTNGGNAA(A/G)ACNACN(T/C)TNGC N(C/A)GNGCNGTN-3'

Overgo-A: 5'-GGNATGGGNGGNGTNGGNAA(A/G)ACNA-3'

Overgo-B: 5'-NACNGCNC(G/T)NGCNA(A/G)NGTNGT(T/C)TTN -3'

PCR products of the NBS genes used as probes for BAC library screening: RGA3 or RLG3 (U55805)

2. RLK family:

EKSDVYSFGVVLLE (domain-IX)

Degenerate nucleotide oligo: 5'-GA(A/G)AA(A/G)(T/A)(C/G)NGA(T/C)GTNTA(T/C)(T/A)(C/G) NTT(T/C)GGNGTNGTN(T/C)TN(T/C)TNGA(A/G)-3'

Overgo-A: 5'- GA(A/G)AA(A/G)(T/A)(C/G)NGA(T/C)GTNTA(T/C)(T/A)(C/G)NTT(T/C)G-3'

Overgo-B: 5'-(T/C)TCNA(A/G)NA(A/G)NACNACNCC(A/G)AAN(C/G)(A/T)(A/G)-3'

Gossypium

NBS family:

2D17 (AY600405), 2B19 (AY600394), 2A21 (AY600391), 2B05 (AY600392), 2K15 (AY600423), 2O05 (AY600431), 2B08 (AY600382), 2D03 (AY600401), 2D14 (AY600383), 2B06 (AY600385), 2H01 (AY600413), 1C08 (AY600376), 2J21 (AY600419), 2K13 (AY600422), 2G13 (AY600410), 2F07 (AY600409)

Supplementary Table S4A. Number of genes in the NBS family in the genomes of three *Oryza* lines estimated by small-insert DNA library screening (SDLS) using the degenerate overgos designed from the conserved NBS domain amino acid sequences of the family as a probe (see Supplementary Table S3).

Description	Library [a]			
	O. sativa ssp. *japonica* (cv. Nipponbare)	*O. sativa* ssp. *indica* (cv. Teqing)		*O. rufipogon* (PI590422)
	pUC18	pGEM5	pUC18	pUC18
No. of clones screened	2,074	1,710	2,659	2,988
No. of clones with inserts	1,980 (95.45%)	1,496 (87.50%)	2,659 (100%)	2,750 (92.00%)
Average insert size (bp)	5,181	5,898	5,760	4,092
Genome coverage (Mb)	10.26	8.80	15.32	11.25
Genome coverage (%)	2.39	2.05	3.56	2.62
No. of positive clones	16	24	29	17
Total:				
Genome coverage (Mb)	10.26	24.12		11.25
Genome coverage (%)	2.39	5.61		2.62
No. of positive clones	16	53		17
No. of the genes in the genome	683.04	962.43		661.87

[a] The pUC18 libraries were constructed from nuclear DNA partially digested with *Mbo*I and size-selected on agarose gels, and the pGEM5 libraries from nuclear DNA partially digested with an enzyme mixture of *Hae*III, *Alu*I and *Rsa*I and size-selected on agarose gels.

Supplementary Table S4B. Number of genes in the RLK family in the genomes of three *Oryza* lines estimated by small-insert DNA library screening (SDLS) using the degenerate overgos designed from the conserved domain amino acid sequences of the family as a probe (see Supplementary Table S3).

Description	Library *O. sativa* ssp. *japonica* (cv. Nipponbare) pUC18	*O. sativa* ssp. *indica* (cv. Teqing) pUC18	*O. rufipogon* (PI590422) pUC18
No. of clones screened	2,074	4,848	1992
No. of clones with inserts	1,980 (95.45%)	4,605 (95.00%)	1833 (92.00%)
Average insert size (bp)	5,181	5,760	4,092
Genome coverage (Mb)	10.26	26.52	7.5
Genome coverage (%)	2.39	6.50	1.74
No. of positive clones	23	86	11
No. of the genes in the genome	986.35	1426.84	642.40

Supplementary Table S4C. Number of genes in the NBS and RLK families in the genome of *Glycine max* cv. Forrest estimated by small-insert DNA library (SDLS) screening using the degenerate overgos designed from the conserved domain amino acid sequences of the families as probes, respectively (see Supplementary Table S3).

Description	Library: pUC18	
	NBS	RLK
No. of clones screened	3,189	6,378
No. of clones with inserts	3,044 (95.45%)	6,088 (95.45%)
Average insert size (bp)	4,518	4,518
Genome coverage (Mb)	13.75	27.50
Genome coverage (%)	1.26	2.52
No. of positive clones	18	43
No. of the genes in the genome	1,426.91	1,704.36

Supplementary Table S4D. Number of genes in the NBS family in the genomes of three *Gossypium* lines estimated by small-insert DNA library screening (SDLS) using purified PCR products of the 16 NBS genes representing the family (see Supplementary Table S3).

Description	Library				
	G. herbaceum (A1-120)		*G. raimondii* (D5-8)	*G. hirsutum* (TM-1)	
	pGEM5	pUC18	pGEM5	pGEM5	pUC18
No. of clones screened	3,600	2,760	5,503	3,771	6,405
No. of clones with inserts	3,240 (90.00%)	1,756 (63.63%)	5,273 (95.83%)	3,186 (84.00%)	5,124 (80.00%)
Average insert size (bp)	4,748	2,894	6,089	6,840	4,067
Genome coverage (Mb)	15.38	5.08	32.11	21.80	20.84
Genome coverage (%)	0.932	0.299	3.65	0.899	0.859
No. of positive clones	14	2	13	11	16
Total:					
Genome coverage (Mb)	20.47		32.11	42.63	
Genome coverage (%)	1.231		3.65	1.758	
No. of positive clones	16		13	27	
No. of the genes in the genome	1,326.66		356.26	1,535.74	

Supplementary Table S5. Comparison in the numbers of genes in the NBS and RLK families estimated with different methods.

Method [a]	*Oryza*						*Gossypium*		
	O. sativa [b] cv. Nipponbare		*O. sativa* cv. Teqing		*O. rufipogon* acc. PI590422		*G. herbaceum* acc. A1-120	*G. raimondii* acc. D5-8	*G. hirsutum* cv. TM-1
	NBS	RLK	NBS	RLK	NBS	RLK	NBS	NBS	NBS
Membrane array hybridization	679.60	1120.04	1031.14	1604.73	598.96	686.15	1184.23	326.94	1713.71
Library screening	683.04	986.35	962.43	1426.84	661.87	642.40	1326.66	356.26	1535.74
Difference between two methods (%)	0.51	13.55	7.14	12.47	10.50	6.81	12.03	8.97	11.50

[a] The numbers of genes measured with the membrane array (MA) method correlated with those measured by the small-insert DNA library screening (SDLS) method ($r = 0.966$, $P < 0.001$).

[b] The number of NBS genes in the Nipponbare genome estimated by the WSBA method ranged from 508 – 597 by different researchers (15,19-21) (also see Supplementary Table S2). The number of NBS genes in Nipponbare estimated by the MA method differed by 13.7% [(679-597)/597 x 100] from that (957) estimated by the WSBA method, but the percentage was close to the 17.5% [(597-508)/508 x 100] artificial variation of the result estimated with WSBA by different researchers (15,19-21).

Supplementary Table S6. Variation in the number of genes in the NBS family between two rice cultivars, Nipponbare and Teqing, and between two soybean cultivars, Williams 82 and Forrest, estimated by screening BAC libraries using its subfamily-specific probes.

Probe [a]	BACs hybridized with a single probe	BACs cross-hybridized with ≥ 2 probes	Total	BACs hybridized with a single probe	BACs cross-hybridized with ≥ 2 probes	Total
	O. sativa ssp. *japonica* cv. Nipponbare BAC libraries (8.3x) [b]			*O. sativa* ssp. *indica* cv. Teqing BAC libraries (8.1x) [b]		
r2	12	4	16	40	39	79
r3	34	1	35	125	24	149
r4	7	0	7	102	52	154
r5	5	5	10	82	20	102
r7	3	0	3	290	75	365
r8	62	17	79	244	102	346
r9	94	8	102	165	120	285
r12	7	2	9	22	27	49
r13	9	1	10	27	45	72
r16	48	14	62	116	129	245
Total	281	52	333	1213	633	1846
	G. max cv. Williams 82 BAC library (5.5x)			*G. max* cv. Forrest BAC library (5.5x)		
RGA3			11			353

[a] See Leister et al. (32) for the detail of rice probes and Kanazin et al. (38) for the detail of soybean probe.

[b] Both Nipponbare (35) and Teqing (33,34) BAC libraries were constructed with *Hin*dIII, *Bam*HI and *Eco*RI, respectively. The Nipponbare BAC libraries have an average insert size of 151 kb and the Teqing BAC libraries have an average insert size of 133 kb. A total of 23,040 (8.3x) Nipponbare BACs and 26,112 (8.1x) Teqing BACs were screened. Both Williams 28 (36) and Forrest (37) BAC libraries were constructed with *Eco*RI. The Williams 82 library has an average insert size of 150 kb and the Forrest BAC library has an average insert size of 157 kb. A total of 40,000 (5.5x) Williams 82 BACs and 38,400 (5.5x) Forrest BACs were screened.

Supplementary Table S7. Variation correlation between species phylogenetic distance and $\log_{10}$-transformed number of genes in the NBS and RLK families

Species analyzed	Phylogenetic distance		N	NBS [a]		RLK [a]	
	Source	Reference		Coefficient (*r*)	*P* (2-tailed)	Coefficient (*r*)	*P* (2-tailed)
Oryza	142 nuclear genes	56	21	0.792***	0.000	0.727***	0.000
Oryza	20 chloroplast genes	57	15	0.692**	0.004	0.777***	0.001
Glycine	rDNA ITS-I	55	15	- 0.446	0.096	- 0.526*	0.044
Gossypium:							
Diploids	22 repeated sequences	52	1485	0.241***	0.000	Not studied	
Polyploids	22 repeated sequences	52	227	0.090	0.177	Not studied	

[a] "*", "**" and "***" indicate that the variation is significant in two tails at $P \leq 0.05$, 0.01 and 0.001, respectively.