Theor Appl Genet (2004) 108:903–913
DOI 10.1007/s00122-003-1499-2

ORIGINAL PAPER

Ai-Guo Tian · Jun Wang · Peng Cui · Yu-Jun Han · Hao Xu · Li-Juan Cong · Xian-Gang Huang · Xiao-Ling Wang · Yong-Zhi Jiao · Bang-Jun Wang · Yong-Jun Wang · Jin-Song Zhang · Shou-Yi Chen

Characterization of soybean genomic features by analysis of its expressed sequence tags

Received: 20 August 2003 / Accepted: 9 October 2003 / Published online: 18 November 2003

Abstract We analyzed 314,254 soybean expressed sequence tags (ESTs), including 29,540 from our laboratory and 284,714 from GenBank. These ESTs were assembled into 56,147 unigenes. About 76.92% of the unigenes were homologous to genes from *Arabidopsis thaliana* (*Arabidopsis*). The putative products of these unigenes were annotated according to their homology with the categorized proteins of *Arabidopsis*. Genes corresponding to cell growth and/or maintenance, enzymes and cell communication belonged to the slow-evolving class, whereas genes related to transcription regulation, cell, binding and death appeared to be fast-evolving. Soybean unigenes with no match to genes within the *Arabidopsis* genome were identified as soybean-specific genes. These genes were mainly involved in nodule development and the synthesis of seed storage proteins. In addition, we also identified 61 genes regulated by salicylic acid, 1,322 transcription factor genes and 326 disease resistance-like genes from soybean unigenes. SSR analysis showed that the soybean genome was more complex than the *Arabidopsis* and the *Medicago truncatula* genomes. GC content in soybean unigene sequences is similar to that in *Arabidopsis* and *M. truncatula*. Furthermore, the combined analysis of the EST database and the BAC-contig sequences revealed that the total gene number in the soybean genome is about 63,501.

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00122-003-1499-2

A.-G. Tian and J. Wang contributed equally to this work.

Communicated by H.F. Linskens

A.-G. Tian · B.-J. Wang · Y.-J. Wang · J.-S. Zhang (✉) · S.-Y. Chen (✉)
Plant Biotechnology Laboratory,
Institute of Genetics and Developmental Biology,
Chinese Academy of Sciences,
Datun Road, 100101 Beijing, China
e-mail: jszhang@genetics.ac.cn
e-mail: sychen@genetics.ac.cn
Tel.: +86-10-64886859
Fax: +86-10-64873428

J. Wang · P. Cui · Y.-J. Han · H. Xu · L.-J. Cong · X.-G. Huang · X.-L. Wang · Y.-Z. Jiao
Beijing Genomics Institute,
Chinese Academy of Sciences,
101300 Beijing, China

Introduction

For the past two decades, random sequencing of cDNA has been a simple and efficient method for the identification of many genes in an organism (Adams et al. 1991; Okubo et al. 1991). Recently, multi-tissue expressed sequence tag (EST) projects have been reported for *Arabidopsis* (Delseny et al. 1997), rice (Ewing et al. 1999), maize (Fernandes et al. 2002) and soybean (Shoemaker et al. 2002). Large-scale EST/cDNA discovery and analysis are widely used for the study of gene expression, the identification of candidate genes for biological processes and phenotypes (Hatey et al. 1998) and also provide a means with which to study gene evolution (Van der Hoeven et al. 2002). Isolated and partially characterized cDNA clones have been used not only as expressed sequence tags on RFLP linkage maps (Kurata et al. 1994), but also as resources for simple sequence repeats (SSRs) (Cardle et al. 2000) and single nucleotide polymorphisms (SNPs) (Garg et al. 1999).

Soybean (*Glycine max* L. Merr.) is one of the most economically important crop species in the world. It was domesticated and cultivated in China for thousands of years (Xu et al. 1986). Today, soybean is a model crop system because of its densely saturated genetic map (Cregan et al. 1999) and a well-developed genetic transformation system (Zhang et al. 1999). In our laboratory, a genetic map has been constructed and several soybean disease resistance analogs and defense genes have been isolated and characterized (Zhang et al. 1997; Liu et al. 2000; Wu et al. 2001; He et al. 2001, 2002, 2003). To further understand the overall genomic

features of soybean, it would be beneficial to undertake sequencing of the entire genome as in the case of *Arabidopsis* and rice. Because soybean has a large genome of 1,115 Mb (Arumuganathan and Earle 1991) compared to *Arabidopsis* (145 Mb) and rice (420 Mb) and approximately 40–60% of the genome sequence is repetitive sequence and heterochromatic (Goldberg 1978; Gurley et al. 1979), it would be difficult to undertake sequencing of the entire genome at present. However, we can still gain some information about the soybean genome by single pass sequencing of cDNA clones to generate ESTs and the analysis of these ESTs. Here we sequenced cDNA clones to generate 29,540 ESTs from soybean. The 29,540 soybean ESTs from our laboratory and 284,714 from GenBank were analyzed and comparisons with *Arabidopsis* and *Medicago truncatula* were made. In addition, several soybean BAC sequences from GenBank were analyzed. As a result, we identified some soybean-specific genes, disclosed some of the main features of the soybean genome and prepared for sequencing of the soybean genome in the future.

Materials and methods

cDNA library, sequencing and data set preparation

One part of our soybean EST data set was derived from our own database, which was constructed from 2-week old seedlings of soybean cultivar Kefeng 1 cultivated in pots and treated by spreading 2.0 mmol/l salicylic acid (SA) until run-off for 24, 36, 48 and 72 h (He et al. 2003). The primary phage cDNA library was constructed and excised into bacterial cultures as phagemids according to the manufacture's instructions (Stratagene; http://www.stratagene.com). Plasmids were isolated according to a standard alkaline lysis protocol and used for capillary sequencing (MegaBACE 1000). Raw data as chromatogram files were processed for base-calling and quality assessment using the Phred software program (Phred-Phrap-Consed package; Ewing and Green 1998; Ewing et al. 1998). Vector sequence was masked with CROSS_MATCH (version 0.990319, Phil Green). All sequences were screened for homology to mitochondrial sequences, rRNA and viral sequences using BLASTN (Schaffer et al. 1999). Finally, sequences whose length was less than 80 bp were excluded from the following analysis. The EST sequences were all deposited in the GenBank dbEST database (under accession numbers: CD390295-CD418706 and CD486719-CD487846).

The soybean genome sequences data set, *M. truncatula* ESTs and genome sequence data sets, *Arabidopsis* genome sequence and unigene data sets were all downloaded from GenBank on 19 October 2002 (http://www.ncbi.nlm.nih.gov). The proteome of *A. thaliana* was downloaded from the *Arabidopsis* database (http://www.tigr.org) in November 2002.

EST assembly

In this study, we used a new method to assemble the large EST data set. First, we used software to mask the mathematically defined repeats (MDRs) (Wang et al. 2002; Yu et al. 2002). These masked EST sequences were then clustered by the D2-cluster program (Burke et al. 1999) using the default parameters. After clustering was complete, the EST sequences in each cluster that were chosen from the primary EST data set were assembled into contigs by Phrap software (P. Green, http://www.phrap.com) using the default parameters.

Dot-blot analysis

Purified plasmid DNA from 103 EST clones of interest were immobilized onto two sets of Hybond N^+ nylon membranes and hybridized with cDNA probes from control and 2.0 mmol/l SA-treated (24 h) leaves, respectively. The cDNA probes were synthesized by reverse transcription in the presence of ^{32}P-dCTP. Hybridization was carried out for 16 h at 65°C. The filters were washed successively with 2×, 1× and 0.5×SSC containing 0.1% SDS for 15 min at 65°C and exposed to Fuji Medical X-ray film at −70°C.

SSRs and nucleotide composition analysis

To find SSRs in the data sets, the Sputnik program (a c-program written by Chris Abajian, Washington University; http://www.abajian.com/sputnik) was used, which detects SSR units between 2 and 5 bases. The program can detect imperfections in the pattern. In the analysis, only those SSRs with perfect repeat patterns were included in the following analysis. Dinucleotide, trinucleotide, tetranucleotide, and pentanucleotide repeats with lengths of 14 bp, 15 bp 16 bp and 20 bp, respectively, were categorized as SSRs, as similarly defined in other studies (e.g., Cardle et al. 2000). Several Perl scripts were then written to summarize the SSR data from the Sputnik program (e.g., the frequency of occurrence of a particular SSR motif). The average GC content within genes was calculated in a 500-bp window using a Perl script.

Functional annotation of the EST sequences

The deduced peptide sequences corresponding to the cDNA sequences were compared with the proteome of *Arabidopsis* whose functional categories were assigned by TAIR (http://www.arabidopsis.org/) and TIGR (http://www.tigr.org/) with terms from the Gene Ontology Consortium controlled vocabularies (http://www.geneontology.org) using the BLASTX program. In the annotation, the following criteria were used. When multiple hits were found, the one with the longest extended homology was selected, where, at the same time, at least 25% of the protein length was matched (Yu et al. 2002).

Analysis of the BAC-contig sequences

All the BAC-contig sequences were downloaded from GenBank. The gene prediction of these BAC-contig sequences was based on the gene-prediction program FGENSH (*Arabidopsis* match/FGENESH prediction; http://www.softberry.com/berry.phtml). The predicted proteins were searched against the soybean unigene data set using the TBLASTN program and the GenBank non-redundant nucleotide database and protein database for functional annotation using the BLASTN and BLASTX programs.

Results

EST assembly and establishment of the unigene sets

We obtained 29,540 ESTs by sequencing a cDNA library constructed from SA-treated soybean seedlings. These ESTs, together with 284,714 soybean ESTs from the GenBank database, were analyzed. After clustering and assembly, 56,147 unigenes were obtained, including 32,278 contigs and 23,869 singletons. The sequence length of these unigenes ranged from 0.1 kb to 4.1 kb, and the average length was 360 bp and 816 bp for singletons and contigs, respectively.

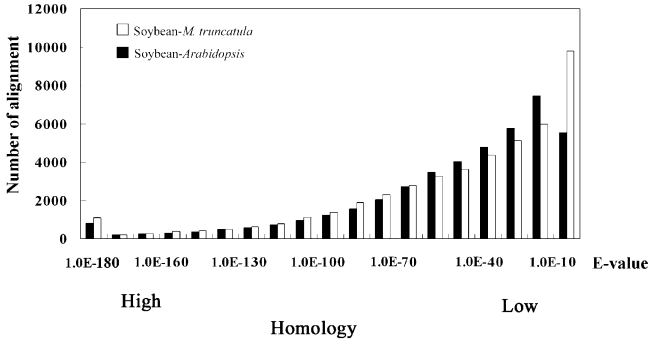

Fig. 1 Homology analysis of soybean unigene-encoded proteins compared with those from *Arabidopsis* and *Medicago truncatula*. Soybean unigene-encoded proteins were searched against proteins from *Arabidopsis* and *M. truncatula* unigene translations with BLASX or TBLASTX (E-value≤1.0E−1). The *filled* columns represent number of genes that showed homology in amino acid sequence at a given E-value between soybean and *Arabidopsis*. The *open* columns repesent that between soybean and *M. truncatula*

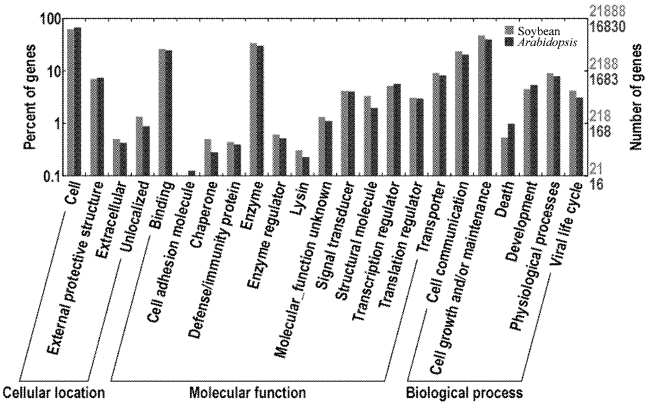

Fig. 2 Functional classification of soybean unigenes according to the Gene Ontology Consortium. The classification was assigned by homology with categorized *Arabidopsis* genes. Only 61.7% of 27,290 predicted genes for *Arabidopsis* were classified. For soybean, 39.0% of 56,147 unigenes were classified

To compare the data from soybean with that from another legume species, *M. truncatula*, the 170,500 ESTs from *M. truncatula* in the GenBank database were also assembled and 34,262 unigenes resulted, including 17,947 contigs and 16,315 singletons. The unigene sequence length ranged from 0.1 kb to 4.2 kb, and the average length was 497 bp and 853 bp for singletons and contigs, respectively.

Functional annotation of the soybean unigene set

The soybean unigenes were searched against the proteome of *Arabidopsis* by BLASTX and against *M. truncatula* unigenes by TBLASTX for homology analysis. The results in Fig. 1 showed that, when the expectation value (E-value) was set to be less than 1.0E−1, 76.92% of the soybean unigenes were homologous to genes from *Arabidopsis*, and 81.56% of them were homologous to the unigenes from *M. truncatula*. When the E-value was lower, the number of homologous sequences decreased in both comparisons (Fig. 1). When the E-value was set to be less than 1.0E−180 and very high level of homology was concerned, only 1.80% of the soybean unigenes showed homology to the proteome of *Arabidopsis*, and 2.33% of the soybean unigenes showed homology to the *M. truncatula* unigenes.

The soybean unigenes were further annotated on the basis of the existing annotations for the proteome of *Arabidopsis,* which were assigned by the Gene Ontology Consortium (http://www.geneontology.org). During the annotation, two criteria were used. When multiple hits were found, the one with the longest extended homology was selected, where, at the same time, at least 25% of the protein length was matched (Yu et al. 2002).

Of the 43,187 soybean unigenes (76.92% of the total soybean unigenes) whose products showed homology (E-value≤1.0E−1) to the *Arabidopsis* proteome, 50.7% (21,888 of 43,187) had a significant match with the categorized proteins of *Arabidopsis* (16,830) based on the two criteria and were assigned putative functions (Fig. 2). However, 49.3% (21,299) of the soybean unigenes had no significant match with the categorized proteins of *Arabidopsis* and could not be classified. The largest proportion of the functionally assigned unigenes fell into five categories: cell, cell growth or maintenance, enzymes, cell communication and binding (Fig. 2). These categories were also the largest found for the *Arabidopsis* proteome and *M. truncatula* unigenes (data not shown).

Identification of soybean-specific genes

The encoded products of the soybean unigenes were compared with the proteome of *Arabidopsis* using BLASTX. The corresponding unigenes that had no homology at E-value≤1.0E−1 were further searched against the *Arabidopsis* genomic sequences using TBLASTX. The remaining unigenes that exhibited no detectable match at E-value<1.0E−1 to *Arabidopsis* genomic sequences were regarded as soybean-specific unigenes. As a result, a total of 12,927 (23.0%) soybean-specific unigenes were identified. These unigenes were further searched against the GenBank non-redundant nucleotide database and protein database to identify their putative functions. We found that only 331 of these unigenes showed significant similarity (E-value <1.0E−10) to GenBank entries. Among these, 111 unigenes can be classified into groups of signal transduction (21), nodule (16), enzymes (44), cell wall (11), proteinase inhibitor (7), seed proteins (12) (see Electronic Supplementary Material, Supplement 1), and 189 encoded hypothetical proteins. The remaining 31 sequences represented contaminating sequences. In the nodule class, all the genes are involved in nodule development, nitrogen fixation or regulation of symbiosis. These genes may have

been lost in other non-legume plants. In the seed protein class, genes encoding soybean seed storage proteins were included. These products are of major economic importance as a nutritive source. In the enzyme class, many polyphenol oxidase genes were identified. The polyphenol oxidases are usually involved in resistance against herbivory, and also present in tomato (*Lycopersicon esculentum*) and many other plants (Van der Hoeven et al. 2002). Why this kind of gene is not present in the *Arabidopsis* genome remains unknown.

Comparisons of soybean unigenes with those in *A. thaliana* and *M. truncatula*

To study the extent of gene conservation between soybean and *Arabidopsis* or between *M. truncatula* and *Arabidopsis*, the encoded amino acid sequences of soybean and *M. truncatula* unigenes were compared with the categorized proteins of *Arabidopsis*. The E-value of a BLAST similarity search was used as an estimation of sequence conservation (Van der Hoeven et al. 2002). A sequence from soybean or *M. truncatula* that had a homologous sequence in *Arabidopsis* with an E-value less than 1.0E−50, from 1.0E−15 to 1.0E−50, and from 1.0E−1 to 1.0E−15 was regarded as a slow-evolving, intermediate-evolving or fast-evolving sequence, respectively. The unigenes in each evolving class were also assigned to functional categories by analyzing their homology with categorized proteins of *Arabidopsis* in order to examine whether certain functional classes of genes evolved more rapidly between soybean and *Arabidopsis* or between *M. truncatula* and *Arabidopsis*. The summary of this analysis is presented in Electronic Supplementary Material, Supplement 2. Of the 56,147 soybean unigenes, 29,521 (52.6%) showed similarity to the categorized proteins of *Arabidopsis* (E-value<1.0E−1), and fast-evolving, intermediate-evolving, and slow-evolving sequences accounted for 26.0%, 38.2% and 35.8%, respectively (Electronic Supplementary Material, Supplement 2A–C).

In the slow-evolving category (Electronic Supplementary Material, Supplement 2C), the percentage of genes was 50.6%, 37.1%, and 25.2% for cell growth and/or maintenance, enzymes, and cell communication, respectively. In the intermediate-evolving category, the percentage of the three sections decreased to 43.7%, 31.7%, and 22.6%, respectively. In the fast-evolving category, the proportions of these sections only accounted for 43.9%, 31.6%, and 20.6%, respectively. The decline of these percentages from slow-evolving to intermediate-evolving and/or fast-evolving categories may indicate the relative house-keeping functions of the corresponding proteins. In contrast, the genes for transcription regulation, cell, binding and death appeared to have evolved quickly (Electronic Supplementary Material, Supplement 2). For transcription regulators, the percentage changed from 4.6% in the slow-evolving category to 6.0% in the intermediate-evolving category and 6.7% in the fast-evolving category. For cell, the percentage was 59.4% in the slow-evolving category, but 65.1% and 64.9% in the intermediate-evolving category and the fast-evolving category, respectively. For binding, the percentage was 26.7%, 26.2% and 28.3% for the categories of slow-evolving, intermediate-evolving, and fast-evolving, respectively. The functional category for death also evolved quickly because the percentage was 0.16% in the slow-evolving category, but 0.99% and 1.71% in the intermediate-evolving and fast-evolving categories, respectively. For the evolutionary analysis between *M. truncatula* and *Arabidopsis,* similar patterns were observed (Electronic Supplementary Material, Supplement 2).

The genes regulated by SA

The soybean ESTs from Genbank were derived from over 50 cDNA libraries and the ESTs from our laboratory were obtained from SA-treated soybean seedlings. Considering that the other libraries were constructed from all major plant organs (roots, leaves, flower, pods and seeds), but not from SA-treated tissue, we identified the transcripts specific to our SA-treated library by searching the assembled soybean database. The contigs that were composed only of ESTs derived from our SA-treated cDNA library, but not from any other libraries were chosen. In this way, 49 entries were identified as SA-specific contigs. These contigs were further compared, using the TBLASTN program, with sequences from the GenBank database and the proteome of *Arabidopsis*. Twenty-nine contigs showed strong similarity to accessions in the GenBank database or the proteome of *Arabidopsis* (Electronic Supplementary Material, Supplement 3). Of these contigs, 19 encoded unknown proteins. The rest included two contigs for binding, one for cell, two for cell growth and/or maintenance, four for enzymes and metabolism and one for transcription.

The above ESTs represent genes whose expression was induced by SA where no expression occurred under normal conditions. However, there were still many genes whose expression was significant under normal conditions and induced or inhibited by SA. To identify these genes, 103 ESTs were selected for a more detailed examination of their expression in response to SA by dot-blot analysis. These ESTs represent genes involved in signal transduction, disease resistance, defense response, stress response, RNA metabolism and protein synthesis and processing. The hybridization results revealed that eight ESTs were up-regulated whereas four ESTs were down-regulated. The up-regulated genes encoded the following products: Pti1, which is involved in the hypersensitive response (Zhou et al. 1995); poly(A)-binding protein, which can interact with zucchini yellow mosaic potyvirus (Wang et al. 2000); a transcription factor which can bind to auxin response elements (Ulmasov et al. 1997); 9-*cis*-epoxycarotenoid dioxygenase, which is involved in abscisic acid biosynthesis under water stress (Iuchi et al. 2000); snakin-1, which is active against plant pathologens; a water channel protein; translation initiation factor

Table 1 Comparison of transcription factor (TF) families in soybean, *Arabidopsis* and rice. The number of the selected TF families in soybean were based on homology with TFs in *Arabidopsis* and the GenBank database

TF family	Number in soybean unigenes	Number in *Arabidopsis*	Number in rice
MYB superfamily	206	190	156
Ap2/EREBP	290	144	143
CH zinc finger	61	105	125
MADS box	34	82	71
WRKY	125	72	83
Dof	39	36	21
Trihelix	13	28	8
BZip	165	81	75
YABBY	16	6	5

SUI1and a translation initiation factor (TIF). The down-regulated genes encoded the following products: protein synthesis factor eIF-4C (Dever et al. 1994); PP7 (Andreeva et al. 1998); leucine zipper proteins (Tang et al. 2001) and an abscisic acid-, stress-, ripening-induced (ASR)-like protein. Further analysis of these genes should reveal their functions in SA-mediated responses.

Transcription factors in the soybean unigenes

To analyze transcription factors (TFs) in soybean, the TFs from the *Arabidopsis* protein database were used as queries to search against soybean unigenes. The soybean unigenes were also used as queries to search against the sequences in the GenBank database for TFs. The two searches were incorporated and 1,322 TF genes were identified (Table 1). This number was similar to the 1,533 TFs in *Arabidopsis* (Riechmann et al. 2000) and 1,306 TFs in rice (Goff et al. 2002). The TFs were divided into several classes according to their structural features (Table 1). The number of members of the MYB family in soybean was similar to that in *Arabidopsis* but more than that in rice. Both the Ap2 and BZip families in soybean appeared to have two-fold more members than those in *Arabidopsis* and rice. The WRKY family members were also more numerous than those in *Arabidopsis* and rice. However, the C2H2 zinc finger and the MADS box families in soybean had fewer members when compared with those in *Arabidopsis* and rice. Soybean also had 41 members identified as belonging to the GATA/CO class, less than 61 in *Arabidopsis* and more than 36 in rice.

Disease resistance genes in soybean unigenes and in a *M. truncatula* BAC-contig sequence

Disease resistance genes (*R* genes) are responsible for early and specific recognition of pathogen attack and initiation of signal transduction leading to deployment of defense mechanisms. *R* genes fall into two major and three minor structural classes. The largest class of known *R* gene products contain nucleotide binding sites (NBS), leucine-rich repeats (LRRs) and an apoptosis-resistance-conserved (ARC) domain. By comparing these classified *R* genes from *Arabidopsis* (http://www.tigr.org/tdb/e2k1/ath1/disRgenes.shtml) with the present soybean unigenes and by comparing soybean unigenes with the sequences from the GenBank database, 326 *R* gene candidates were identified. Among these, 79.8% (260 of 326) contained Toll/Il-1 receptor (TIR), NBS, or LRR domains, more than the 193 identified in *Arabidopsis* (http://www.tigr.org/tdb/e2k1/ath1/disRgenes.shtml).

Plant *R* genes have been identified in many plants (Hammond-Kosack and Jones 1997), and they frequently occur in tightly linked clusters (Michelmore and Meyers 1998). Kanazin et al. (1996) found that *R* genes in soybean were clustered and Graham et al. (2002) identified 16 different resistance-like gene sequences in a BAC sequence. We examined whether *R* genes in another legume, *M. truncatula,* were clustered. The six BAC-contig sequences of *M. truncatula* were analyzed and the results based on the gene prediction program FGENSH (*Arabidopsis* match/FGENESH prediction) (http://www.softberry.com/berry.phtml). The predicted proteins were compared to the sequences from the GenBank database using the BLASTP program. The results identified eight predicted proteins in a BAC-contig sequence (Table 2). These products exhibited high homology to soybean *R* gene product KR1 (He et al. 2003) and were named RLG1–RLG8, respectively. This analysis indicated that the resistance gene homologs in *M. truncatula* are also clustered.

Identification of SSRs

SSRs are arranged repeats of short DNA motifs (1–6 bp length) that frequently exhibit variation in the number of repeats at a locus. They are believed to vary through DNA replication slippage and are related to genetic instability. In the present study, 56,147 unigenes from soybean and 7,559,020 bp of soybean genomic sequence were analyzed. There were 3,383 and 1,908 SSRs in soybean unigenes and genomic sequences, respectively (Table 3). Furthermore, 34,262 unigenes and 26,291,568 bp of genomic sequences from *M. truncatula*, and 27,159 unigenes and the whole genome from *Arabidopsis* were also analyzed. There were 3,962 SSRs in *Arabidopsis* unigenes, and 14,229 SSRs in *Arabidopsis* genomic sequences. In *M. truncatula*, there were 2,479 SSRs in the unigenes, and 3,175 SSRs in genomic sequences (Table 3).

Table 2 Prediction of eight disease resistance-like proteins in a BAC-contig sequences from *Medicago truncatula*

Name of predicted protein	Number of exons	Length of predicted protein	Site in BAC-contig sequence	Annotation	E-Value
RGL1	2	673	75,512–78,088	Functional candidate resistance protein KR1	6.00E−90
RGL2	3	826	80,814–83,949	Functional candidate resistance protein KR1	1.00E−123
RGL3	2	553	85,026–86,924	Functional candidate resistance protein KR1	1.00E−177
RGL4	6	647	93,357–101,390	Functional candidate resistance protein KR1	1.00E−114
RGL5	4	738	155,810–160,526	Functional candidate resistance protein KR1	1.00E−173
RGL6	6	1,066	162,096–171,940	Functional candidate resistance protein KR1	1.00E−180
RGL7	2	551	173,788–175,680	Functional candidate resistance protein KR1	1.00E−161
RGL8	4	612	185,535–191,288	Functional candidate resistance protein KR1	1.00E−168

Table 3 SSR survey in unigenes and genomic sequences from soybean, *Arabidopsis* and *M. truncatula*

Source	Soybean		*M. truncatula*		*Arabidopsis*	
	Unigene	Genome	Unigene	Genome	Unigene	Genome
Dinucleotide	1,319	769	867	1,575	593	6,630
Trinucleotide	1,617	927	1,205	950	3,252	6,103
Tetranucleotide	298	181	255	434	89	1,021
Pentanucleotide	149	31	152	216	28	475
Total SSR	3,383	1,908	2,479	3,175	3,962	14,229
Total length (Mb)	34.95	7.56	23.43	26.3	37.15	115.4
Average distance (kb)	10.33	3.96	9.45	8.28	9.38	8.11

The average distance between SSRs was different between different species and between unigenes and genomic sequences. In the three species, the average distance between two SSRs in unigenes was very similar, i.e., 10.33 kb in soybean, 9.45 kb in *M. truncatula* and 9.38 kb in *Arabidopsis*. The average distance between two SSRs in the genomic sequences of soybean was only half of that in *M. truncatula* and *Arabidopsis* (Table 3).

The frequency of occurrence of SSRs according to repeat motif length (di-, tri-, tetra-, and penta-) was different (Fig. 3). Of all the SSRs found in soybean unigenes, dinucleotide, trinucleotide, tetranucleotide and pentanucleotide repeats accounted for 39.0%, 47.8%, 8.8% and 4.4% respectively, which was similar to 40.3%, 48.6%, 9.5% and 1.6%, respectively, in soybean genomic sequences. In *Arabidopsis* unigenes, dinucleotide, trinucleotide, tetranucleotide and pentanucleotide repeats accounted for 15%, 82.1%, 2.2% and 0.7%, respectively, which was different from 46.6%, 42.9%, 7.2% and 3.3% in genomic sequences. In *M. truncatula* unigenes, dinucleotide, trinucleotide, tetranucleotide and pentanucleotide repeats accounted for 35.0%, 48.6%, 10.3% and 6.1% respectively, which was different from 49.6%, 29.9%, 13.7% and 6.8% in genomic sequences. In short, the most common type of repeat motif in unigenes was trinucleotide SSRs in all the three species. In genomic sequences, the most common type in soybean was trinucleotide repeats, whereas in *Arabidopsis* and *M. truncatula*, the most common type was dinucleotide repeats.

The proportion of SSR repeat units in different species and between unigene and genomic sequences was different (Fig. 3). In unigenes, the most common repeat unit was AG repeats for soybean and *M. truncatula*, and GAA repeats for *Arabidopsis*. But in genomic sequences, the most common repeat unit was AT repeats for all three species.

GC content in genomic and cDNA sequences

The average genomic GC content for prokaryotes and eukaryotes varies widely. It ranges from less than 22% in the human malaria parasite, to more than 68% in the large amplicon of *Halobactrium* sp. NRCI. In order to elucidate the difference in GC content between soybean, *Arabidopsis* and *M. truncatula*, we used the 500-bp window size for analysis of unigenes and genomic sequences. The mean GC content was 0.43, 0.44 and 0.40 for unigene sequences in soybean, *Arabidopsis* and *M. truncatula*, respecively (Fig. 4). Furthermore, the GC content distribution in unigenes was similar in all three species. In

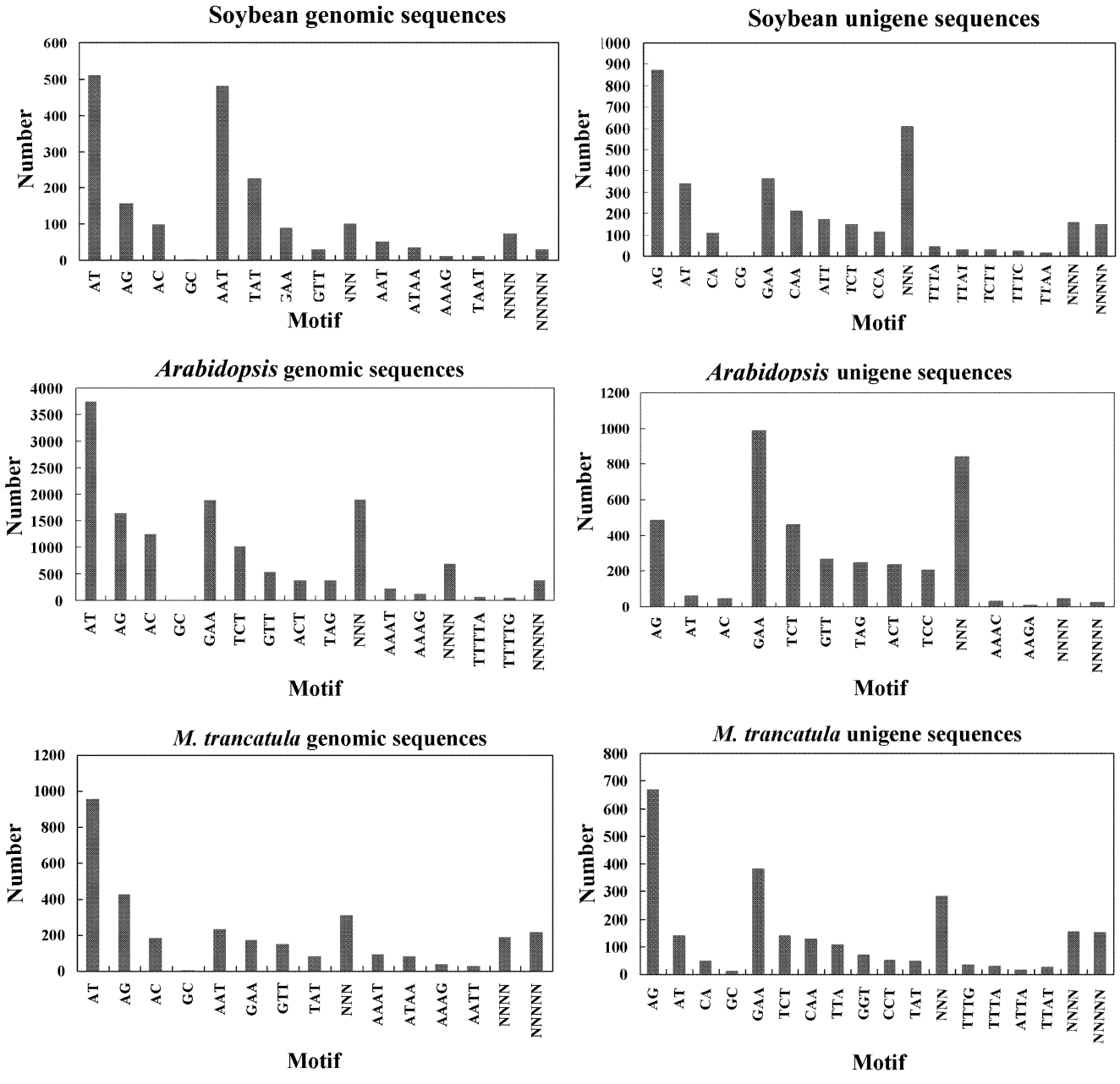

Fig. 3 Frequency of different repeat types in unigenes and genomic sequences from soybean, *Medicago truncatula* and *Arabidopsis thaliana*

genomic sequences, the mean GC content was 0.35, 0.36, and 0.34 for soybean, *Arabidopsis* and *M. trancatula,* respectively. The GC content distribution in the *Arabidopsis* genomic sequence displayed a shoulder on the AT-rich side as reported by Yu et al. (2002). However, GC content distribution in soybean and *M. truncatula* did not display this shoulder.

Gene number estimation based on BAC sequences and the EST database

The four BAC-contig sequences from soybean were analyzed and the prediction was based on the gene-prediction program FGENSH (*Arabidopsis* match/FGENESH prediction; http://www.softberry.com/berry.phtml). The predicted proteins were compared with the soybean unigene data set using the TBLASTN program.

A total of 169 putative genes were identified computationally on the four BAC-contig sequences, and 97 (57.4%) had perfect matches in the EST-derived unigene

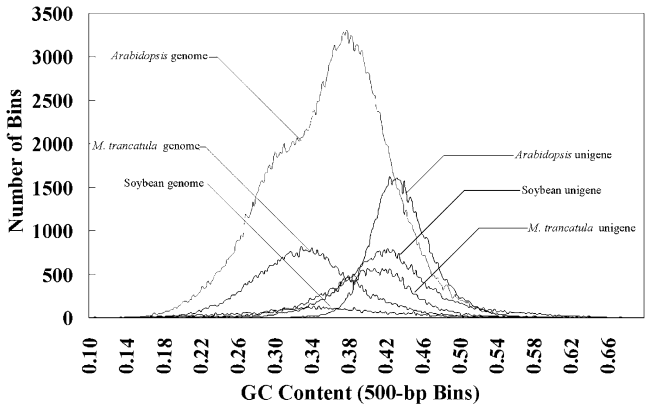

Fig. 4 GC content distribution was computed over a window of 500 bp in the unigenes from soybean, *Arabidopsis* and *Medicago trancatula*, and in the genomic sequences of soybean, *Arabidopsis* and *M. trancatula*. The mean GC content in the unigenes was 0.43, 0.44 and 0.40 for soybean, *Arabidopsis* and *M. trancatula,* respectively. The mean GC content in the genomic sequences was 0.35, 0.36 and 0.34 for soybean, *Arabidopsis* and *M. trancatula,* respectively

set (Table 4). For another 72 genes, imperfect or no matches were found. The predicted gene density of the four BACs varied from 5.3 kb/gene to 6.9 kb/gene and the average gene density was 6.074 kb/gene. The total gene number was therefore estimated to be over 183,569, according to the average gene density of these four BACs and the size of the entire soybean genome (1,115 Mb). However, this number is probably an overestimate due to the possibility that the average gene density of the sequenced BACs was not representative of the entire soybean genome and only gene-rich genomic regions were analyzed.

A more accurate estimation of the total gene number of soybean may be made by comparing the number of the EST-derived unigenes and the percentage of predicted genes in genomic sequence (e.g., BAC sequences) that showed a perfect match with the EST-derived unigenes, as described for tomato by Van der Hoeven et al. (2002). A 35% overestimate of the actual number of genes in *Arabidopsis* has been revealed by genomic sequencing (*Arabidopsis* Genome Initiative; http://www.arabidopsis.org/) and this percentage has been used to estimate the gene content of tomato (Van der Hoeven et al. 2002). A similar situation may be true for soybean. Assuming a 35% overestimation, the soybean EST-derived unigene set was calculated to contain only 36,495 (56,174×65%) unique genes instead of 56,174. Considering that these unigenes contained perfect matches for 57.4% of the predicted genes in the four sequenced BACs, we estimated that the total gene number of soybean was 63,501 (36,495/57.4%). This number is about two-fold higher than that of *Arabidopsis*. However, this increase in gene number is not proportional to the nine-fold larger size of the soybean genome (1,115 Mb for soybean versus 125 Mb for *Arabidopsis*). Using the same method, we estimated that the gene number for *M. truncatula* is about 64,000 (data not shown), similar to that of soybean.

Discussion

Characteristics of soybean unigenes

The analysis of the large EST database showed that 314,254 soybean EST sequences can be assembled into 56,147 unigenes, including 23,869 singletons and 32,278 contigs. In all these soybean unigenes, 76.92% showed a match (E-value<1.0E−1) to *Arabidopsis* genes and 81.56% showed a match (E-value<1.0E−1) to *M. truncatula* unigenes. Among those sequences that matched *Arabidopsis* genes, the categories of cell growth and/or maintenance, enzymes and cell communication appeared to evolve more slowly, whereas the genes whose functions belonged to the TF and death categories appeared to diverge more rapidly between soybean and *Arabidopsis*. A similar observation has been made in a comparison between *M. truncatula* and *Arabidopsis*.

The soybean unigenes with no match to the proteome of *Arabidopsis* represented soybean-specific genes. These genes belonged to functional groups of signal transduction, nodule, enzymes, cell wall, proteinase inhibitor, and

Table 4 The gene prediction in four BAC-contig sequences from soybean and comparison with soybean unigenes

GenBank accession number	Predicted gene number	Number of predicted genes with perfect unigene matches	Number of predicted genes with imperfect or no unigene matches	Sequence length (bp)	Gene density (kb/gene)	Reference[a]
AX196294	21	13	8	127,197	6.057	Hauge BM, Wang ML, Parsons JD, Parnell LD
AX196296	48	26	22	335,913	6.998	Hauge BM, Wang ML, Parsons JD, Parnell LD
AX196297	60	38	22	349,954	5.832	Hauge BM, Wang ML, Parsons JD, Parnell LD
AX223856	40	20	20	213,535	5.338	Hauge BM, Wang ML, Parsons JD, Parnell LD
Total	169	97	72	1,026,599	6.074	Hauge BM, Wang ML, Parsons JD, Parnell LD

[a] The sequences that were downloaded from the NCBI were submitted by Hauge et al.

seed proteins, and may represent the fast-evolving genes between soybean and *Arabidopsis*. Most of these genes have new functions in soybean, including nodule development, nitrogen-fixation, storage protein, etc.

SA-regulated genes, disease resistance genes, and TFs

SA plays an important role in the defense responses of plants, mediates the oxidative burst that leads to cell death in the hypersensitive response and acts as a signal for the hypersensitive response and the development of systemic acquired resistance (Shirasu et al. 1997). SA is also involved in the plant response to adverse environmental conditions, such as salt and osmotic stress (Borsani et al. 2001). In the present study, 49 genes related to SA were identified using the soybean EST database from GenBank and our own EST database from a SA-treated cDNA library. Among these, 29 unigenes showed strong similarity to the sequences from the GenBank database and were classified into binding, cell, cell growth and/or maintenance, enzymes and metabolism, transcription and unclassified functional categories. By dot-blot analysis of an additional 103 ESTs, 12 ESTs were found to be regulated by SA. Further analysis of these genes may provide an explanation of the mechanism of SA function in plants.

Disease resistance genes (*R* genes) are involved in the recognition of pathogen attack and initiation of the defense response. In the present study, we identified 326 soybean unigenes that encoded proteins similar to known *R* gene products using homologous comparison. A number of these *R* genes were reported to be clustered in the soybean genome (Kanazin et al. 1996; Graham et al. 2002). In *M. truncatula*, eight resistance-like genes in a contig sequence were identified, and their encoded proteins all showed strong similarity to soybean *R* gene product KR1 (He et al. 2003), which had several homologs in the soybean genome. It is thus possible that KR1 and its homologs are also clustered in the soybean genome.

TFs are important for the regulation of eukaryotic gene expression. 1,322 TFs were identified in the present work by their significant similarity to known TFs. The number of members in the MYB family, AP2 family, WRKY family, YABBY and single zinc-finger Dof family was more than that in *Arabidopsis* and rice. But the number in the MADS box family and the C2H2 zinc finger class in soybean was less than that in *Arabidopsis* and rice. The difference in the gene numbers may reflect different requirements in various plant species. A given portion of TFs may play important roles in certain stages of plant growth and development, hence determining the differences observed in diverse plant species.

SSR analysis, GC content and gene content of soybean

Since the soybean genome contains 40–60% repetitive and heterochromatic sequences, we analyzed the SSRs in both genomic sequences and unigenes. In the genomic sequences, the average distance between two SSRs in soybean was 3.96 kb, which is half that observed in *M. truncatula* and *Arabidopsis* and consistent with the fact that soybean has more repeat sequences. However, in unigenes, the average distance between two SSRs was about 10 kb in all the three species, indicating that unigenes contain fewer repeated sequences. Our analysis showed that the most common type of repeat motif in unigenes was trinucleotide SSRs in all three species. In genomic sequences, the most common type in soybean was trinucleotide, which was different from dinucleotide in *Arabidopsis* and *M. truncatula*. Furthermore, the proportion of SSR repeat units in unigenes was different. In unigenes, the most common repeat unit was AG repeats for both soybean and *M. truncatula*; however, the common repeat unit for *Arabidopsis* was GAA. In genomic sequences, the most common repeat unit was AT repeats for all three species. The SSRs obtained in the present study can be used as molecular markers for the fine mapping of agronomic traits. This analysis also showed that the soybean genomic sequences are more complex than those in *M. truncatula* and *Arabidopsis*, and the assembly of sequences will be more difficult for soybean genome sequencing.

The compositional gradients in soybean were analyzed using a 500-bp window size. The results showed that the mean GC content in unigene sequences of the three plant species was higher than that in genomic sequences, and the mean GC content was similar in soybean, *M. truncatula* and *Arabidopsis*. By comparison, the GC content in rice is 0.51 for exons and 0.43 for genomic sequences (Yu et al. 2002). Both values are higher than those in soybean, which are 0.43 and 0.35 for unigene and genomic sequences, respectively. The GC content may reflect the relative stability of a given plant genome.

How many genes does soybean have? The present analysis indicates that there are probably 63,501 genes in the soybean genome. This number is higher than that of rice and tomato (Yu et al. 2002; Van der Hoeven et al. 2002), but about two-fold higher than that of *Arabidopsis*. From this estimation, we can see that the gene number in soybean is not correlated with its genome size when compared to the situation in rice and *Arabidopsis*. The reason for this remains unclear. The exact gene number will be revealed when full genome sequencing is performed.

Acknowledgements This research was supported by National Key Basic Research Special Funds, P.R. China (G1998010209, 2002CB111301).

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Andreeva AV, Evans DE, Hawes CR, Bennett N, Kutuzov MA (1998) PP7, a plant phosphatase representing a novel evolutionary branch of eukaryotic protein Ser/Thr phosphatases. Biochem Mol Biol Int 44:703–715

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–219

Borsani O, Valpuesta V, Botella MA (2001) Evidence for a role of salicylic acid in the oxidative damage generated by NaCl and osmotic stress in *Arabidopsis* seedlings. Plant Physiol 126:1024–1030

Burke J, Davison D, Hide W (1999) D2-cluster: a validated method for clustering EST and full-length cDNAsequences. Genome Res 9:1135–1142

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156:847–854

Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, VanToai TT, Lohnes DG, Chung J, Specht JE (1999) An integrated genetic linkage map of the soybean genome. Crop Sci 39:1464–1490

Delseny M, Cooke R, Raynal M, Grellet F (1997) The *Arabidopsis thaliana* cDNA sequencing projects. FEBS Lett 403:221–224

Dever TE, Wei CL, Benkowski LA, Browning K, Merrick WC, Hershey JW (1994) Determination of the amino acid sequence of rabbit, human, and wheat germ protein synthesis factor eIF-4C by cloning and chemical sequencing. J Biol Chem 269:3212–3218

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie JM (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res 9:950–959

Fernandes J, Brendel V, Gai X, Lal S, Chandler VL, Elumalai RP, Galbraith DW, Pierson EA, Walbot V (2002) Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and micro-array hybridization. Plant Physiol 128:896–910

Garg K, Green P, Nickerson DA (1999) Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Res 9:1087–1092

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Goldberg RB (1978) DNA sequence organization in the soybean plant. Biochem Genet 16:45–68

Graham MA, Marek LF, Shoemaker RC (2002) Organization, expression and evolution of a disease resistance gene cluster in soybean. Genetics 162:1961–1977

Gurley WB, Hepburn AG, Key JL (1979) Sequence organization of the soybean genome. Biochem Biophys Acta 561:167–183

Hammond-Kosack KE, Jones JDG (1997) Plant disease resistance genes. Annu Rev Plant Physiol Plant Mol Biol 48:575–607

Hatey F, Tosser-Klopp G, Clouscard-Martinato C, Mulsant P, Gasser F (1998) Expressed sequence tags for genes: a review. Genet Sel Evol 30:521–541

He C-Y, Zhang Z-Y, Chen S-Y (2001) Isolation and characterization of soybean NBS analogs. Chin Sci Bull 46:1984–1988

He C-Y, Zhang J-S, Chen S-Y (2002) A soybean gene encoding a proline-rich protein is regulated by salicylic acid, an endogenous circadian rhythm and by various stresses. Theor Appl Genet 104:1125–1131

He C-Y, Tian A-G, Zhang J-S, Zhang Z-Y, Gai J-Y, Chen S-Y (2003) Isolation and characterization of a full-length resistance gene homolog from soybean. Theor Appl Genet 106:786–793

Hoeven R van der , Ronning C, Giovannoni J, Martin G, Tanksley S (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 14:1441–1456

Iuchi S, Kobayashi M, Yamaguchi-Shinozaki K, Shinozaki K (2000) A stress-inducible gene for 9-cis-epoxycarotenoid dioxygenase involved in abscisic acid biosynthesis under water stress in drought-tolerant cowpea. Plant Physiol 123:553–562

Kanazin V, Marek LF, Shoemaker RC (1996) Resistance gene analogs are conserved and clustered in soybean. Proc Natl Acad Sci USA 93:11746–11750

Kurata N, Nagamura Y, Yamamoto K, Harushima Y, Sue N, Wu J, Antonio BA, Shomura A, Shimizu T, Lin SY, Inoue T, Fukuda A, Shimano T, Kuboki Y, Toyama T, Miyamoto Y, Kirihara T, Hayasaka K, Miyao A, Monna L, Zhong HS, Tamura Y, Wang ZX, Momma T, Umehara Y, Yano M, Sasaki T, Minobe Y (1994) A 300 kilobase interval genetic map of rice including 883 expressed sequences. Nat Genet 8:365–372

Liu F, Zhuang BC, Zhang JS, Chan SY (2000) Construction and Analysis of Soybean Genetic Map. Acta Genet Sin 27:1018–1026

Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res 8:1113–1130

Okubo K, Hori N, Matoba R, Niiyama T, Matsubara K (1991) A novel system for large-scale sequencing of cDNA by PCR amplification. DNA Seq 2:137–144

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. Science 290:2105–2110

Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15:1000–1011

Shirasu K, Nakajima H, Rajasekhar VK, Dixon RA, Lamb C (1997) Salicylic acid potentiates an agonist-dependent gain control that amplifies pathogen signals in the activation of defense mechanisms. Plant Cell 9:261–270

Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton SW, Waterston R, Smoller D, Coryell V, Khanna A, Erpelding J, Gai X, Brendel V, Raph-Schmidt C, Shoop EG, Vielweber CJ, Schmatz M, Pape D, Bowers Y, Theising B, Martin J, Dante M, Wylie T, Granger C (2002) A compilation of soybean ESTs: generation and analysis. Genome 45:329–338

Tang Z, Sadka A, Morishige DT, Mullet JE (2001) Homeodomain leucine zipper proteins bind to the phosphate response domain of the soybean vspb tripartite promoter. Plant Physiol 125:797–809

Ulmasov T, Hagen G, Guilfoyle TJ (1997) ARF1, a transcription factor that binds to auxin response elements. Science 276:1865–1868

Wang J, Wong GK, Ni P, Han Y, Huang X, Zhang J, Ye C, Zhang Y, Hu J, Zhang K, Xu X, Cong L, Lu H, Ren X, He J, Tao L, Passey DA, Yang H, Yu J, Li S (2002) RePS: a sequence

assembler that masks exact repeats identified from the shotgun data. Genome Res 12:824–831

Wang X, Ullah Z, Grumet R (2000) Interaction between zucchini yellow mosaic potyvirus RNA-dependent RNA polymerase and host poly-(A) binding protein. Virology 275:433–443

Wu X-L, He C-Y, Wang Y-J, Zhang Z-Y, Dong F-Y, Zhang J-S, Chen S-Y, Gai J-Y (2001) Construction and analysis of a genetic linkage map of soybean. Acta Genet Sin 28:1051–1061

Xu B, Zhen HY, Lu QH, Zhao SW, Zhou SH, Hu ZA (1986) Three new evidences of the original area of soybean. Soybean Sci 5:123–130

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Li J, Liu Z, Qi Q, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Zhao W, Li P, Chen W, Zhang Y, Hu J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Tao M, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92

Zhang DS, Dong W, Hui DW, Chen SY, Zhuang B-C (1997) Construction of a soybean linkage map using an F2 hybrid population from a cultivated variety and a semi-wild soybean. Chin Sci Bull 42:1326–1330

Zhang ZA, Xing P, Staswick T, Clemente T (1999) The use of glufosinate as a selective agent in Agrobacterium-mediated transformation of soybean. Plant Cell Tissue Organ Cult 556:37–46

Zhou JM, Loh YT, Bressan RA, Martin GB (1995) The Tomato Gene Ptil encodes a Serine/Threonine kinase that is phosphorylated by Pto and is involved in the hypersensitive response. Cell 83:925–935