Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Abundance and diversity of resistance genes in the sugarcane transcriptome revealed by *in silico* analysis

A.C. Wanderley-Nogueira, N.M. Soares-Cavalcanti, D.A.L. Morais, L.C. Belarmino, A. Barbosa-Silva and A.M. Benko-Iseppon

Departamento de Genética, Laboratório de Genética e Biotecnologia Vegetal, Centro de Ciências Biológicas,
Universidade Federal de Pernambuco, Recife, PE, Brasil
Corresponding author: A.M. Benko-Iseppon
E-mail: ana.benko.iseppon@pesquisador.cnpq.br

Genet. Mol. Res. 6 (4): 866-889 (2007)
Received August 03, 2007
Accepted September 25, 2007
Published October 05, 2007

ABSTRACT. Resistance genes (*R*-genes) are responsible for the first interaction of the plant with pathogens being responsible for the activation (or not) of the defense response. Despite their importance and abundance, no tools for their automatic annotation are available yet. The present study analyzed *R*-genes in the sugarcane expressed sequence tags database which includes 26 libraries of different tissues and development stages comprising 237,954 expressed sequence tags. A new annotation routine was used in order to avoid redundancies and overestimation of *R*-gene number, common mistakes in previous evaluations. After *in silico* screening, 280 *R*-genes were identified, with 196 bearing the complete domains expected. Regarding the alignments, most of the sugarcane's clusters yielded best matches with proteins from *Oryza sativa*, probably due to the prevalence of sequences of this monocot in

data banks. All *R*-gene classes were found except the subclass LRR-NBS-TIR (leucine-rich repeats, nucleotide-binding site, including Toll interleukin-1 receptors), with prevalence of the kinase (*Pto*-like) class. *R*-genes were expressed in all libraries, but flowers, transition root to shoot, and roots were the most representative, suggesting that in sugarcane the expression of *R*-genes in non-induced conditions prevails in these tissues. In leaves, only low level of expression was found for some gene classes, while others were completely absent. A high allelic diversity was found in all classes of *R*-genes, sometimes showing best alignments with dicotyledons, despite the great number of genes from rice, maize and other grasses deposited in data banks. The results and future possibilities regarding *R*-genes in sugarcane research and breeding are further discussed.

Key words: *Saccharum*, Defense, Pathogens, NBS-LRR, Kinase, Expression profile

INTRODUCTION

A major concern regarding plant genome research is to recognize genes responsible for important traits, including defense genes against infection by pathogens. Because plants are sessile, they cannot move to avoid biotic attack or abiotic stress, or to find mating partners. Thus, they depend heavily on chemical signals. Large-scale sequencing revealed that plants possess many more genes than do animals, mainly due to polyploidy or large-scale duplication (Borevitz and Ecker, 2004).

Different defense mechanisms are responsible for the protection of plants and animals against their biotic environment. No major histocompatibility complex genes or antibody-like genes have been identified in plants; however, plant resistance genes (*R*-genes) are abundant and can be grouped into subfamilies (Jones, 2001; Meyers et al., 2003).

Sugarcane is one of the most important sources of sugar and alcohol in the world and is cultivated in tropical and subtropical areas in more than 80 countries around the globe. In 2004/2005, 2.7 x 10^7 tons of sugarcane were produced in Brazil alone, in an area estimated at one million hectares, and were used mainly for sugar consumption or as energy source (ethanol), contributing to 25% of the world's production (UDOP, 2007). The cultivated sugarcane varieties are the result of interspecific hybridization involving *Saccharum officiarum*, *S. barberi*, *S. sinense*, and the two wild species *S. spontaneum* and *S. robustum*. It is thought that *S. officinarum* was originally selected by humans in Papua New Guinea, perhaps from *S. robustum* germplasm. Because of its multispecific origin, sugarcane is thought to have one of the most complex plant genomes, carrying also variable chromosome numbers (generally 2n = 70-120) with a commensurately large DNA content (Lu et al., 1994).

A large-scale sequencing of SUCEST (sugarcane expressed sequence tags, ESTs) was carried out as a first step in depicting the genome of this important tropical crop. Twenty-six

unidirectional cDNA libraries were constructed from a variety of tissues sampled from thirteen different sugarcane cultivars. A total of 291,689 cDNA clones were sequenced in their 5' and 3' end regions. After trimming low-quality sequences and removing vector and ribosomal RNA sequences, 237,954 ESTs potentially derived from protein-encoding messenger RNA remained. The average insert size in all libraries was estimated to be 1,250 bp with the insert length varying from 500 to 5,000 bp. Clustering the 237,954 sugarcane ESTs resulted in 43,141 clusters (Vettore et al., 2001). No general evaluation of *R*-genes is yet available for the sugarcane transcriptome.

Despite the importance of such genes for breeding purposes, no automatic annotation tools are available yet. This may be explained by the nature of *R*-genes that combine a limited number of related functional domains also regarding different gene classes (Ellis et al., 1999; Ellis and Jones, 2000). Previous grouping using domains as primary seed sequence resulted in overestimation of gene number and misclassification. This can be explained by evidence that known *R*-genes combine a limited number of related functional domains (Ellis et al., 1999; Ellis and Jones, 2000). Thus, a better understanding of the nature of these genes is necessary in order to understand the difficulties and potentialities regarding automatic annotation, especially in complex genomes such as that of sugarcane.

Plant *R*-genes are responsible for the specific defense response and are the most important group of genes used by breeders for disease control (Rommens and Kishore, 2000). These genes evolve rapidly, since they undergo constant selection pressure by pathogen evolution. For each *R*-gene, there is a corresponding gene in the pathogen, called avirulence (*avr*) gene, which determines pathogenicity. Plants will be resistant and the growth of the pathogen will be arrested only when both genes, *R* and *avr*, are present and compatible (Ellis and Jones, 2000). Thus, for each *R*-gene, there is a corresponding *avr* gene; this is the basis of the gene-for-gene concept, suggested by Flor (1956, 1971). This gene-for-gene interaction is very specific (Meyers et al., 2005; Salvaudon et al., 2005; Ohtsuki and Sasaki, 2006). The *avr* genes determine the inability of a given pathogenic strain to infect a plant that carries the corresponding *R*-gene triggering the hypersensitive reaction (Bonas and Van den Anckerveken, 1999). This relationship that is hypersensitive, race-specific, and governed by interactions between avirulence genes in pathogens and resistance genes in hosts is called qualitative resistance (Nelson, 1972).

In contrast to *R*-genes, *avr* gene products described to date do not comprise a defined family of related proteins, since no sharing of similar motifs or domains has been identified (Richter and Ronald, 2000). Resistance genes are members of a very large multigene family, are highly polymorphic and have diverse recognition specificities (Pryor and Ellis, 1993). The cloned resistance genes were grouped into five classes, based on the predicted protein structure (Song et al., 1997).

The first class includes the tomato gene *Pto*, which confers resistance to *Pseudomonas syringae* pv. The gene *tomato* encodes an active serine/threonine kinase that plays a direct role in both signaling processes and pathogen effectors (Tang et al., 1999; Anderson et al., 2006).

Regarding the second class, the common feature is the presence of leucine-rich repeats (LRRs) which play a direct role in protein-protein-specific recognition events; a nucleotide-binding site (NBS) that is usually involved with signaling molecules in programmed cell death and a leucine zipper or a coiled-coil (CC) sequence, involved in signal transduction during various cell processes. *R*-genes of this class could be found in many plants as

Arabidopsis thaliana (*Rps2*, *RPP8*, *RPP13*, and *Rpm1*), rice (*Pib*, *Pi-ta* and *Xa1)*, tomato (*Prf*, *I2*, Mi, and *Sw5*), and potato (*Hero*) *R*-genes (Mindrinos et al., 1994; Whitham et al., 1994; Grant et al., 1995; Lawrence et al., 1995; Bent, 1996; Salmeron et al., 1996; Ori et al., 1997; McDowell et al., 1998; Milligan et al., 1998; Yoshimura et al., 1998; Wang et al., 1999; Bittner-Eddy et al., 2000; Bryan et al., 2000; Brommonschenkel et al., 2000; Ernst et al., 2002; Rehmany et al., 2005).

The third class includes proteins similar to those described for the second class (often both types are classified together in a single gene class), but instead of a CC sequence at the amino terminal region (Meyers et al., 1999) these proteins have a TIR (Toll interleukin-1 receptor domain), including the genes *L* (Lawrence et al., 1995), and *P* (Dodds et al., 2001) of flax; *RPP1* (Botella et al., 1998), *RPP4* (van der Biezen et al., 2002), *RPP5* (Parker et al., 1997), and *RPS4* (Gassmann et al., 1999) of *A. thaliana*, and *N* (Whitham et al., 1996; Mestre and Baulcombe, 2006) of tobacco. The TIR domain is also present in animals and is believed to be absent in monocotyledonous plants (Ellis and Jones, 2000), but it has been shown to be present in all dicotyledonous taxa studied to date.

A fourth class of resistance genes is represented by the tomato *Cf g*ene family (*Cf2*, C*f4*, *Cf5*, and *Cf9*), which mediates resistance to the fungal pathogen *Cladosporium fulvum* (Jones et al., 1994; Dixon et al., 1996; Kruijt et al., 2005). This gene encodes a putative membrane-anchored protein (TM, transmembrane domain) with the LRR motif in the presumed extracellular domain and a short C-terminal tail in the intracellular domain.

The fifth class is represented by the rice gene *Xa21* (Song et al., 1995; Wang et al., 1996) which encodes an extracellular receptor-like kinase, including also a TM, LRRs and an intracellular serine/threonine kinase domain. Thus, the structure of *Xa21* indicates an evolutionary link between different classes (I and IV) of plant disease resistance genes (Song et al., 1997; Xu et al., 2006).

There is yet a sixth class, which encodes genes of the reductase group with no conserved domains as cited above. This class is represented by the maize *Hm1* gene, which confers resistance against the toxin produced by the fungus *Cochliobolus carbonum* (Johal and Briggs, 1992) and *Mlo* from barley, a putative regulator of defense against *Blumenaria graminis* (Piffanelli et al., 2002).

Previous evaluations have shown that the automatic annotation of *R*-genes may lead to redundancy and wrong classification of *R*-genes (Meyers et al., 1999). In the present paper, we used 30 complete sequences of previously described *R*-genes as template and propose a new approach for unambiguous identification and classification of *R*-gene candidates in plants.

Further important questions regarding the present study include: How many *R*-genes can be identified in the SUCEST database? Do they correspond to the known *R*-gene classes with the same combinations of conserved domains? Are they preferably similar to other Poaceae (e.g., rice, wheat, maize) available sequences in databases? In which tissues are they expressed in non-induced conditions? Considering the allopolyploid and hybrid origin of the sugarcane genome, can one expect a larger diversity of alleles regarding the expressed resistance genes as compared with diploids such as rice, maize and *Arabidopsis*? The present study attempts to bring to light some of these open questions using a data mining-based analysis of plant disease *R*-genes in the SUCEST database, as compared with available information from other plants deposited in public databases.

MATERIAL AND METHODS

The sugarcane ESTs used in the present study are available in Genbank (NCBI, National Center for Biotechnology Information, www.ncbi.nih.gov/). The clusterized ESTs are available at www.biotec.icb.ufmg.br/sucest. Information regarding the 26 libraries that constitute the SUCEST libraries, including experimental conditions and pipeline routines have been described before (Grivet and Arruda, 2001; Vettore et al., 2001). For practical purposes we combined some libraries that comprised different stages of the same tissue/organ (AM1 and AM2 are designated here simply "AM", for example), resulting in a total of 13 libraries (AD: tissues infected by *Gluconacetobacter diazotroficans*; AM: apical meristem; CL: callus; FL: flower; HR: tissues infected with *Herbaspirillum rubrisubalbicans*; LB: lateral bud; LR: leaf roll; LV: leaves; RT: root; RZ: stem-root transition; SB: stalk bark; SD: seeds; ST: stem) considered for a better evaluation of the results of the present study.

For the identification of sugarcane *R*-genes, a search was carried out using sequences of known *R*-genes selected from the literature against the SUCEST database (see Attachment I in Appendix). Members of the sixth class (reductases) were not included in the present evaluation. The genes selected included 27 *R*-genes previously compiled by Barbosa da Silva et al. (2005) including all five gene classes previously described. To this study, we added three sequences, namely the genes *Pi-ta* and *Pib* from rice and *RPM1* from *A. thaliana* (accession numbers AAK00132, BAA76282 and AC016827_19, respectively) all belonging to the second class (LRR-NBS) described before.

For the identification of *R*-genes, tBLASTn alignments were carried out against SUCEST database using the 30 seed sequences described above. After this search, sugarcane sequences that were found to match *R*-genes with a cut-off of e^{-20} were used for a homology screening of *R*-genes in Genbank (NCBI) using BLASTx (Altschul et al., 1990). The cluster frame of the tBLASTn alignment was used to predict the open-reading frames (ORFs) for each selected cluster.

A second general analysis using a cut-off of e^{-10} was also carried out followed by an elimination of some redundancies (genes that matched more than one gene class due to common domains). For this purpose, matching clusters to each query sequence were annotated on a local database (called 'non-redundant'). Cluster name was adopted as primary key in order to identify and prevent inclusion of the same cluster in different gene classes due to the presence of common domains.

Exclusively in the case of the third class of *R*-genes (LRR + NBS + TIR), an additional tBLASTn search was carried out using only the TIR domain to confirm its presence/absence in sugarcane. Sugarcane clusters were translated using the TRANSLATE tool of Expasy (http://us.expasy.org/) and screened for conserved motifs with the aid of the RPS-BLAST CD-search tool (Altschul et al., 1990). Multiple alignments with CLUSTALx program allowed the structural analysis of the sequences including conserved and diverging sites as well as the elimination of non-aligned terminal segments. For each *R*-gene class, one resistance gene (*Pto*, *Xa1*, *Cf*, and *Xa21*, respectively) was selected to perform a phenetic UPGMA (unweighted pair groups method using arithmetic averages) analysis using a bootstrap function with 1,000 replicates. For this purpose CLUSTALx alignments were submitted to the program MEGA (Molecular Evolutionary Genetic Analysis), version 3, for Windows, kindly provided by the authors (Kumar et al., 2004).

The resulting dendrogram was created with the program TreeView for Windows (Page, 1996) kindly provided by Dr. Robert Page (University Glasgow, Scotland).

A preliminary analysis of *R*-gene distribution patterns in sugarcane libraries was verified by direct correlation of the frequency reads for each cluster in various SUCEST cDNA libraries (Figure 3, see Results).

To generate an overall picture of *R*-genes expression patterns in sugarcane, a hierarchical clustering approach (Eisen et al., 1998) was applied using normalized data and a graphic representation constructed with the aid of the CLUSTER program. Dendrograms including both axes (using the weighted pair-group for each gene class and library) were generated by the TreeView program (Eisen et al., 1998). In the diagrams (Figure 4, see Results), yellow means no expression and red all degrees of expression. This approach was previously employed by other plant EST projects such as in rice (Ewing et al., 1999) and also in sugarcane (Lambais, 2001).

RESULTS AND DISCUSSION

Using 30 well-known *R*-genes as template, we could identify 196 clusters in SUCEST database bearing the complete expected domains. Considering the identity of these genes with the queries, a total of 151 clusters could be identified as non-redundant while 45 other sequences aligned with more than one *R*-gene used as query (Table 1).

The use of several previously described and sequenced *R*-genes as seed sequences proved to be a useful and low time consuming strategy in the search for *R*-gene candidates in plants. This approach allowed the identification of a large set of candidate sequences by using various representative genes per class, while former studies (e.g., Koczyk and Chelkowski, 2003) employed few genes. Previous studies have shown that using only domains or few genes as template per class resulted in double grouping of some genes in different classes and caused some level of redundancy (Meyers et al., 1999). In other cases, a higher stringency had to be used (e^{-50} or less) resulting in the exclusion of important gene candidates (Rossi et al., 2003). For example, the kinase domain (present in classes I and V) or the NBS domain (present in classes II and III) often leads to the generation of mixed grouping. Furthermore, the imperfect nature of the LRR domain alone may bring about some problems regarding automatic annotation and classification, showing that this domain is not adequate for this purpose (Barbosa da Silva et al., 2005). Thus, the strategy of generating a local database (here called non-redundant) by adopting the cluster number as a primary key register was very effective in the solution of this problem, helping in the recognition, classification, elimination of duplicates, and inferences about candidates of orthologs and paralogs. We recommend this procedure for the future development of tools specific for *R*-gene automatic annotation, quantification and classification.

By lowering the cut-off value (from e^{-20} to e^{-10} during tBLASTn), an additional 84 *R*-gene clusters could be identified in the sugarcane transcriptome, but many showed only partial sequences or incomplete domains. Altogether, this means that sugarcane encodes a significant number of transcriptionally active *R*-genes (at least 280) with considerable allelic diversity. This number is much higher than the 88 sugarcane sequences identified for the development of resistance gene analog markers by Rossi et al. (2003). The authors used key word search and 17 resistance gene analog-related seed sequences with a stringent BLASTn cut-off (e^{-50}). Many important sequences bearing complete domains have been excluded using this approach, which

Table 1. Main sugarcane clusters similar to known *R*-genes. tBLASTn results and sequence evaluation of sugarcane *R*-genes including the five best matches of each *R*-gene class.

Cluster features and evaluation						BLASTx information			
Gene class and expected domain	Gene name	Sugarcane cluster No.	Size (n)	ORF (aa)	e-value	Score and frame		NCBI gi|-No.	Plant species
Class I kinase	*Pto*	SCEZLB1010F11.g	1350	450	e^{-112}	402	+2	46981335	*Oryza sativa*
	Pto	SCCCLR1072H06.g	1788	596	e^{-100}	360	-1	34909896	*Oryza sativa*
	Pto	SCCCLR1024A02.g	984	328	6.0 e^{-99}	357	-3	34909896	*Oryza sativa*
	Pto	SCSGHR1066F10.g	759	253	6.0 e^{-62}	234	+1	37536416	*Oryza sativa*
	Pto	SCCCRZ1001B02.g	1926	642	7.0 e^{-61}	230	-3	15222211	*Arabidopsis thaliana*
Class II LRR-NBS-CC and Class III LRR-NBS-TIR	*Rps5, RPP8, RPM1, Hrt, R1, BS2, Gpa2, RX2*	SCSFSB1103A12.g	1830	610	2.0 e^{-74}	277	-1	4519936	*Oryza sativa*
	Xa1, Rpp8, RPM1, Hero, I2, Sw5, R1, RP1, Bs2, Rx2	SCEPAM2014D05.g	1725	575	3.0 e^{-58}	224	+1	27542759	*Sorghum bicolor*
	RPP8, RPM1, Hero, Pib, Sw5, RPP13, Hrt, Bs2, GPA2, RX2	SCRUFL1024D04.g	1050	350	4.0 e^{-41}	166	+1	40253400	*Oryza sativa*
	RPP8, RPM1, Pib, Sw5, Pi-ta, R1, GPA2	SCCCCL3080A08.g	762	254	8.0 e^{-35}	145	+1	22208466	*Sorghum bicolor*
	RPP8, RPM1, Hero, I2, Pib, Sw5, Hrt, Bs2, RP1	SCJLRZ1019B10.g	1107	369	4.0 e^{-33}	140	-2	22652528	*Oryza sativa*
Class IV LRR	*Cf5, Cf9*	SCJLRT1020F05.g	1842	614	5.0 e^{-68}	256	+1	50934027	*Oryza sativa*
	Cf4, Cf5, Cf9	SCCCLR1001A03.g	2976	992	1.0 e^{-58}	224	+3	18390097	*Sorghum bicolor*
	Cf5, Cf9	SCJLRT1020F05.g	1842	614	5.0 e^{-49}	192	+1	50934027	*Oryza sativa*
	Cf5	SCCCLR1001A03.g	2976	992	7.0 e^{-48}	189	+3	18390097	*Sorghum bicolor*
	Cf4, Cf5, Cf9	SCJLRT1020F05.g	1842	614	3.0 e^{-45}	180	+1	50934027	*Oryza sativa*
Class V LRR kinase	*Xa21*	SCCCLR1001A03.g	2970	990	1.0 e^{-93}	341	+1	18390097	*Sorghum bicolor*
	Xa21	SCSGLR1045B05.g	1173	391	3.0 e^{-70}	263	+3	34915440	*Oryza sativa*
	Xa21	SCVPRT2081F09.g	798	266	1.0 e^{-66}	351	-1	7434424	*Oryza sativa*
	Xa21	SCCCRZ2C01E06.g	2184	728	1.0 e^{-65}	248	+2	34898088	*Oryza sativa*
	Xa21	SCQGAM2108A06.g	708	236	7.0 e^{-61}	232	+2	34915440	*Oryza sativa*

I) Features and evaluation results with sugarcane cluster number, cluster size in nucleotides (n), ORF (open-reading frame) size in amino acids (aa), e-value, score and frame. II) Data about BLASTx best alignment: NCBI gi|-number and plant species. Some clusters matched significant alignments with genes belonging to class III but presented no TIR domain. LRR = leucine-rich repeats; NBS = nucleotide-binding site; TIR = Toll interleukin-1 receptor domain.

is justified by some needs for use in comparative *in silico* mapping, the main focus of this study. Despite the undoubted identification of *R*-genes using this approach, no overall picture of *R*-gene abundance and diversity within an EST database is possible by using this procedure.

It is interesting to note that only two of the 26 sugarcane libraries were obtained under influence of microorganisms (tissues infected by *Gluconacetobacter diazotroficans* and tissues infected with *Herbaspirillum rubrisubalbicans*) and that none of them are pathogenic but on the contrary, are symbiotic organisms. With exposure to pathogen, the number of genes will probably increase and one may suppose that additional sequences may be identified.

Clusters representing exclusive *R*-gene classes were: I) kinase: 92; II) LRR-NBS-CC: 62; IV) TM-LRR: 27, and V) kinase-TM-LRR: 15. Clusters that aligned with *R*-gene classes II and III (TIR-NBS-LRR) showed only CC, NBS and LRR domains (no TIR domain) and were therefore included in group II, since the presence of TIR is the distinctive factor between the two classes.

The prediction of cluster-coding regions revealed that ORFs were oriented in both forward and reverse reading frames, with an average of 394 amino acids in length. ORF sizes varied from 992 (cluster SCCCLR1001A03.g of the LRR class) to 102 amino acids. Regarding the average ORF length in each *R*-gene class, we observed 380 amino acids for class I (kinase), 262 amino acids for class II (LRR-NBS-CC), 492 amino acids for class IV (TM-LRR), and 442 amino acids for class V (kinase-TM-LRR) (Figure 1).

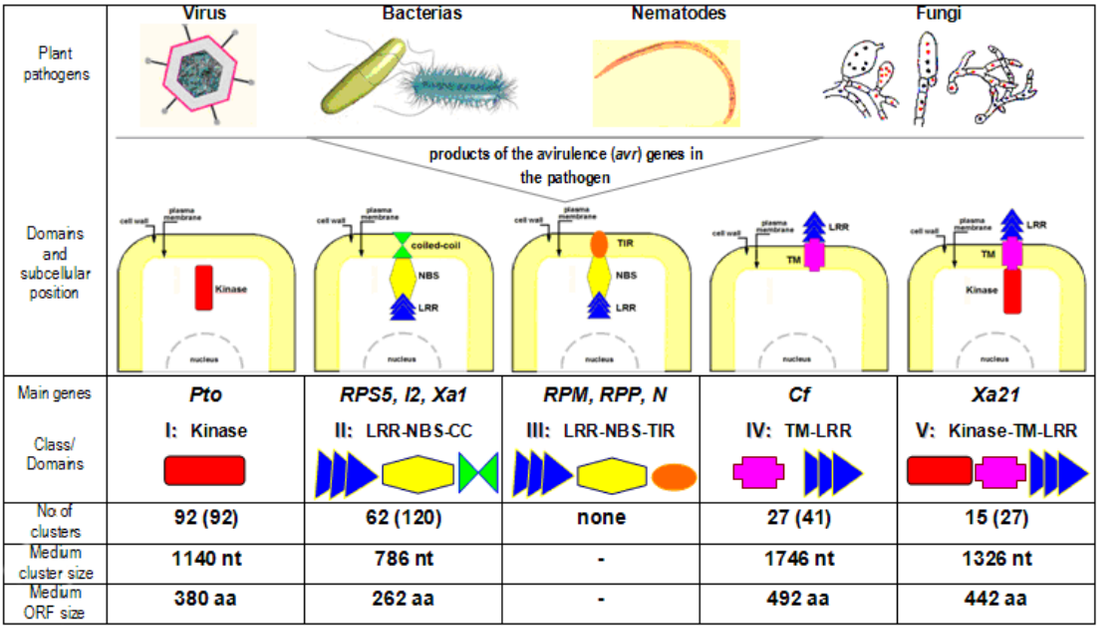

Figure 1. Representation of main *R*-gene classes considering the presence of conserved domains, from SUCEST database. For each class, domains are represented with their putative subcellular position, as well as number of clusters, size range of sequence in nucleotides (nt) and of open-reading frame (ORF) in amino acids (aa). Given cluster numbers were selected using the e^{-20} cut-off for tBLASTn alignments, while numbers between parentheses indicate clusters obtained using a less stringent (e^{-10}) cut-off with the same tool. LRR = 1eucine-rich repeats; NBS = nucleotide-binding site; TIR = Toll interleukin-1 receptor domain; CC = coiled-coil sequence; TM = transmembrane protein; Class I = kinase; Class II = CC-NBS-LRR; Class III = TIR-NBS-LRR; Class IV = TM-LRR; Class V = LRR-TM-kinase.

Most of the 196 clusters that aligned with known *R*-genes were from the monocotyledonous class (169 clusters), represented by eight different species of the Poaceae family, with emphasis on rice. From dicots only three families appeared as best matches (27 clusters), including four different species. A comprehensive inventory of all species that aligned with sugarcane with their taxonomic affiliation is presented in Table 2.

By lowering the cut-off value from e^{-20} to e^{-10} during tBLASTn alignments, additional *R*-gene clusters could be identified in sugarcane (Figure 1), with exception of the kinase group (class I), where the same number of clusters (92) was identified with both approaches. Considering the remaining classes, a higher number of clusters could be identified, even though domains were incomplete or missing in some of them. This is the case of class II (LRR-NBS-CC) where 120 clusters were identified (instead of 62 previously identified at e^{-20}), similar to class IV (TM-LRR) with 41 instead of 27 and

Table 2. Comprehensive inventory of the organisms identified in Genbank as best alignment to each of the 196 identified sugarcane clusters related to known resistance genes.

Gene class	Higher taxonomic affiliation	Family	Species
Class I kinase	Monocots (70)	Poaceae (70)	*Oryza sativa* (64)
			Zea mays (2)
			Sorghum bicolor (4)
	Dicots (22)	Brassicaceae (19)	*Arabidopsis thaliana* (19)
		Leguminosae (1)	*Glycine max* (1)
		Solanaceae (2)	*Nicotiana tabaccum* (2)
Class II and class III CC/TIR LRR-NBS	Monocots (60)	Poaceae (60)	*Oryza sativa* (41)
			Zea mays (5)
			Sorghum bicolor (6)
			Hordeum vulgare (2)
			Saccharum hybrid (1)
			Triticum aestivum (2)
			Triticum monococcum (3)
	Dicots (2)	Brassicaceae (2)	*Arabidopsis thaliana* (2)
Class IV LRR	Monocots (25)	Poaceae (25)	*Oryza sativa* (22)
			Zea mays (1)
			Sorghum bicolor (2)
	Dicots (2)	Brassicaceae (2)	*Arabidopsis thaliana* (2)
Class V LRR-kinase	Monocots (14)	Poaceae (14)	*Oryza sativa* (10)
			Oryza longistaminata (2)
			Sorghum bicolor (2)
	Dicots (1)	Apiaceae (1)	*Daucus carota* (1)

		Number
Grouped by taxonomic affiliation	Dicots	27
	Monocots	169

The organisms are grouped by gene classes (I to V) and taxonomic affiliation (class, family and species). Numbers in parentheses indicate number of members in each taxonomic group or species.

class V (kinase-TM-LRR) with 27 instead of 15 (Figure 1). With this lower cut-off the total number of putative transcribed *R*-genes in sugarcane increased from 196 to 280.

General considerations about conserved domains of sugarcane *R*-genes

Some *R*-genes pertaining to different classes were able to align significantly to the same cluster on SUCEST database, an occurrence also observed during mining of the *Eucalyptus* transcriptome (Barbosa da Silva et al., 2005), probably because known *R*-genes combine a limited number of related functional domains (Ellis et al., 1999).

The conserved domains (CDs) identified during this investigation showed that most of the sugarcane predicted sequences possessed the same motifs shared by previously known disease *R*-genes. The CD with the higher level of sampling was kinase, which was present in class I and class V with a total of 107 occurrences.

The CD search revealed conserved regions (Figure 1) in all of the 196 clusters analyzed. From the 107 clusters that showed the kinase domain, 92 of them matched the *Pto* gene (class I) while 15 matched the *Xa21* gene (class V).

The NBS domain was present in 62 clusters in sugarcane (120 with cut-off of e^{-10}). After a search with genes encoding class III (LRR-NBS-TIR), only non-TIR sequences matching with class II (due to the common NBS domain) could be identified. Also, no significant matches were found after tBLASTn search using exclusively the TIR domain.

NBS domain is highly conserved among plants and is similar to that in mammalian CED-4 and APAF-1 proteins which are involved in apoptosis (Chinnaiyan et al., 1997), with the additional proposition that NB-ARC plays a role in the activation of downstream effectors (Bryan et al., 2000). Transmembrane motifs were found only in 19 of all analyzed sequences, where 14 were related to *Cf* gene and five to *Xa21*.

The other frequent domain shared was LRR, matching 104 occurrences in 77 different clusters in all classes except kinase (class I) represented by the *Pto* gene. LRR can act as a receptor to recognize the avr proteins, as in *Cf* (27 clusters) and *Xa21* (15 clusters) or can be intracellular, like in class II of R-proteins (62 clusters). The LRR motif contains 23-25 amino acids with a consensus sequence (LxxLxxLxLxxNxLt/sgxIpxxLG), but this pattern is often imperfect (Jones, 2001) and may be difficult to recognize with available *in silico* tools. Thus, it is possible that a larger number would be recognized with a lower cut-off and additional manual search.

Class I: *Pto*-like *R*-genes (solely kinase domain)

The first class includes the tomato gene *Pto*, which confers resistance to *Pseudomonas syringae* pv. *tomato*. In tomato, *Pto* has been described as a small gene. The ORF consists of 963 nucleotides, it has no introns, and encodes a functional serine-threonine kinase (Loh and Martin, 1995). A total of 52 alleles of this gene were found in a search including seven *Lycopersicon* species, bearing 41 variant amino acid positions among these alleles (Rose et al., 2005). Ninety-two clusters of sugarcane showed highly significant alignments to the *Pto* seed sequence used (accession 2112354A). The size of the clusters varied between 1926 and 759 nt with ORFs from 642 to 253 amino acids (Table 1, Figure 1), indicating that gene size within this class may vary significantly.

In the sugarcane transcriptome, no redundancy was observed between this class and the *Xa21* class, which also contains a kinase domain. This is in accordance with the observations of Vallad et al. (2001) who used bootstrap analysis to determine that five *Pto*-like kinase families from bean were distinctly different from other kinases. They also found that *Pto*-kinase subdomains VIa, VIb, VIII, and IX of the *Pto*-like class are unique in plant species. This conservation is confirmed by the fact that sugarcane genes of this group aligned to 92 sequences deposited in the Genbank (both cut-offs, e^{-20} and e^{-10}), 22 of them being from dicots (Table 2).

Classes II and III: sugarcane clusters bearing NBS-LRR and NBS-LRR-TIR domains

There is evidence showing that *R*-genes are quite abundant in higher plants, but the most functionally defined *R*-genes belong to the NBS-LRR class (here including classes II and III for a better understanding of their common and distinctive attributes), considered also the largest class of plant disease *R*-genes.

Unlike in *A. thaliana* and other dicots, the NBS-LRR gene class coding for a TIR domain has been shown to be absent in all the monocots studied although mainly members of the Poaceae family have been analyzed. Most cereal genes are similar in structure to the members of the non-TIR class of dicots, although many do not code for a CC domain in their amino termini (Bai et al., 2002).

A total of 85 TIR-NBS-LRR have been identified in *A. thaliana* genome (The Arabidopsis Genome Initiative, 2000) and 93 in *Eucalyptus* transcriptome (Barbosa da Silva et al., 2005). The availability of the rice whole genome sequence enabled the global characterization of NBS-LRR genes, revealing that this crop carries about 500 NBS-LRR genes (at least three to four times the complement found in *A. thaliana*). Over 100 of these genes were predicted to be pseudogenes in the rice cultivar Nipponbare, but some of these were functional in other rice lines. In rice, over 80 other NBS-encoding genes were identified that belonged to four different subclasses, but only two of which are present in dicotyledonous plant sequences present in databases (Monosi et al., 2004). Zhou et al. (2004) considered that 76% of all gene families with a 5-fold size are larger in rice as compared with *A. thaliana*. In the sugarcane transcriptome, this class was represented by 62 clusters bearing complete domains selected with a cut-off of e^{-20}, while 120 clusters could be identified with a lower cut-off (e^{-10}) (Figure 1). Considering the allopolyploid and hybrid nature of sugarcane and also that no libraries under pathogen induction are available in the SUCEST database, an increase in the number of NBS-LRR sequences is expected under different experimental conditions, possibly more than rice which has a smaller genome and is diploid.

In sugarcane as in rice and other cereals, only non-TIR sequences matching NBS-LRR *R*-genes could be identified. No significant matches were found also after tBLASTn search using exclusively the TIR domain. This result confirms previous assumptions that the TIR domain (also present in animals) may be absent in monocotyledonous plants (Ellis and Jones, 2000) such as Poaceae, while being present in all dicotyledonous taxa studied to date. Therefore, it has been suggested that the TIR domain may have been lost in the course of differential evolution between mono- and dicotyledoneous plants (Pan et al., 2000).

As in rice, most alignments in sugarcane using class II genes also occurred with monocots including only two sequences significantly homologous to dicot sequences deposited in databases (Table 2), both from *A. thaliana*. Exceptionally, this is not the most abundant *R*-gene class

found in the SUCEST database when we consider the most stringent procedure (cut-off e^{-20}), with only 62 sequences included in this group versus 92 *Pto*-like sequences that bear exclusively the kinase domain (Figure 1). A more permissive e-value (cut-off e^{-10}) allowed the identification of 120 clusters of the NBS-LRR class, almost double the number identified previously, while in the case of the *Pto*-like sequences (kinase domain) the same number of clusters (92) was revealed in both approaches. Previous researchs identified that overall sequence homology among *R*-genes of the NBS-LRR class is lower than in the kinase class. On the other hand, the NBS contains some sequence motifs, such as P-loop, kinase-2, kinase-3, and GLPAL, that are highly conserved even among distantly related plants (Hammond-Kosack and Jones, 1996). The wide distribution of NBS-LRR genes in the plant kingdom and their prevalence in both monocots and dicots indicate that they are ancient. This was confirmed by Liu and Ekramoddoullah (2003) who amplified TIR-NBS-LRR in gymnosperms (*Pinus monticola*, white pine) confirming that they share a common origin with *R*-genes from angiosperms. Thus, to identify the whole diversity of class II genes, low stringency alignments are advisable.

In the present study, the low number of dicots with best alignments regarding this class of gene may be explained by the imperfect nature of the LRR domain, not always recognizable with available *in silico* tools.

Class IV: *Cf*-like *R*-genes (TM-LRR domains)

Genes of the *Cf* family mediate resistance to the fungal pathogen *Cladosporium fulvum* in tomato. They encode a putative membrane-anchored protein (also named TM) with the LRR motif in the presumed extracellular domain and a short C-terminal tail in the intracellular domain (Jones et al., 1994; Dixon et al., 1996).

In our study, 25 clusters displaying best matches with this protein belonged to monocots while only two matched to dicot (*A. thaliana*) sequences (at e^{-20}). All selected clusters possessed the putative TM-LRR domains. Central for this model for *Cf* protein function is the concept that the highly variable regions within the LRRs are responsible for the recognition of pathogen-encoded avirulence determinants either directly or indirectly through some co-receptor (Dixon et al., 1996). The LRR domains are known to play a role in protein-protein interactions. In tomato, the size of these sequences varied between 968 and 855 amino acids (Dixon et al., 1996), while ORFs identified in sugarcane varied between 992 and 614 amino acids.

As observed in the NBS-LRR group, also in the *Cf*-related genes, the number of putative related clusters is increased to 41 when the cut-off value is e^{-10} (as compared with 27 clusters with stringent e^{-20} conditions). It seems that also in this case the imperfect nature of the LRR domains makes it difficult to identify sequences of this group, and additional mining is needed since under stringent conditions new variants may be not recognized. On the other hand, permissive conditions often result in redundancies that may lead to the classification of the same cluster in different gene classes.

Class V: LRR-TM-kinase (*Xa21*-like *R*-genes)

The structure of this class indicates an evolutionary link between classes I and IV of *R*-genes presented here (Song et al., 1997). Overall annotation revealed that *Arabidopsis* also

carries homologues to the LRR-kinase-*Xa21* group (Jones, 2001), while eight clusters with significant homology to *Xa21* were also found in distantly related woody dicots as in the case of the *Eucalyptus* transcriptome (Barbosa da Silva et al., 2005).

In the present evaluation, 15 of the clusters analyzed corresponded to this class with high e-values, but this number will probably increase if only the receptor-like kinase sequence is used as template, since the LRR may be quite variable between sugarcane and rice. Regarding the best alignments, one of the best matches revealed LRR-kinase of carrot (a dicot from the family Apiaceae) and 14 of monocots (all belonging to Poaceae with 10 rice sequences; Table 2).

Using a PCR-based approach, Song et al. (1997) cloned seven *Xa21* members in rice and found the presence of 15 transposable element sequences, two of them in coding sequences, confirming the influence of such sequences in the evolution of these genes. Whole sequencing in the rice genome revealed this class of genes in two chromosomes: in the short arm of chromosome 12 (including 12 tandem arrays of *Xa21*-like sequences), while the first described sequence for this gene was found on the long arm of chromosome 11 (The Rice Chromosomes 11 and 12 Sequencing Consortia, 2005). According to these authors both regions are full of defense-related genes, confirming the clusterized organization of these sequences in chromosomes. These results reveal that the number of *Xa21* representatives in rice (13) is lower than that found in the sugarcane transcriptome in non-induced libraries (15), suggesting that they are probably still more abundant in sugarcane, including some of the largest contigs included in our evaluation (up to 2184 bp).

Exceptional *R*-genes have proven to provide durable disease control, due to the fast evolving pathogen genome that breaks resistance. The *Xa21* gene is an important exception to this rule that reveals the full potential of *R*-genes for breeding purposes (Rommens and Kishore, 2000). This may be very valuable for sugarcane breeding, especially considering the possibility of pyramidization of such genes in important crops, increasing the potentiality of an effective specific *R*-*Avr* interaction.

Bootstrap analysis of selected *R*-gene groups

The phenetic UPGMA bootstrap analysis revealed grouping between monocots and dicots in most dendrograms considering the four genes used as template (*Pto*, *Xa1*, *Cf9*, and *Xa21*). In all cases, sugarcane R clusters appeared in two or more clades within each dendrogram. Also, grouping of species belonging to different taxonomic families was observed in all cases (Figure 2).

Sometimes, best alignments matched to *O. sativa*, as in the case of *Pto* and *Cf9* groups (Figure 2A), but in all cases where *Sorghum bicolor* sequences were available they showed the best matches, as seen with the *Xa1* and *Xa21* groups (Figure 2B,D). On the other hand, in all four evaluations, some sugarcane clusters displayed lower levels of similarity, and in three cases (*Pto*, *Xa1* and *Cf9*) a sugarcane cluster was the most divergent, remaining in a basal position in the cladogram that included sequences from mono- and dicotyledoneous plants. This makes it clear that sugarcane bears a high allelic diversity of *R*-genes, also considering that the present study did not include libraries obtained under conditions of pathogen stress.

It is interesting to note that all trees grouped species pertaining to different plant families as expected, since phenetic analysis considers only similarity aspects, not evolutionary.

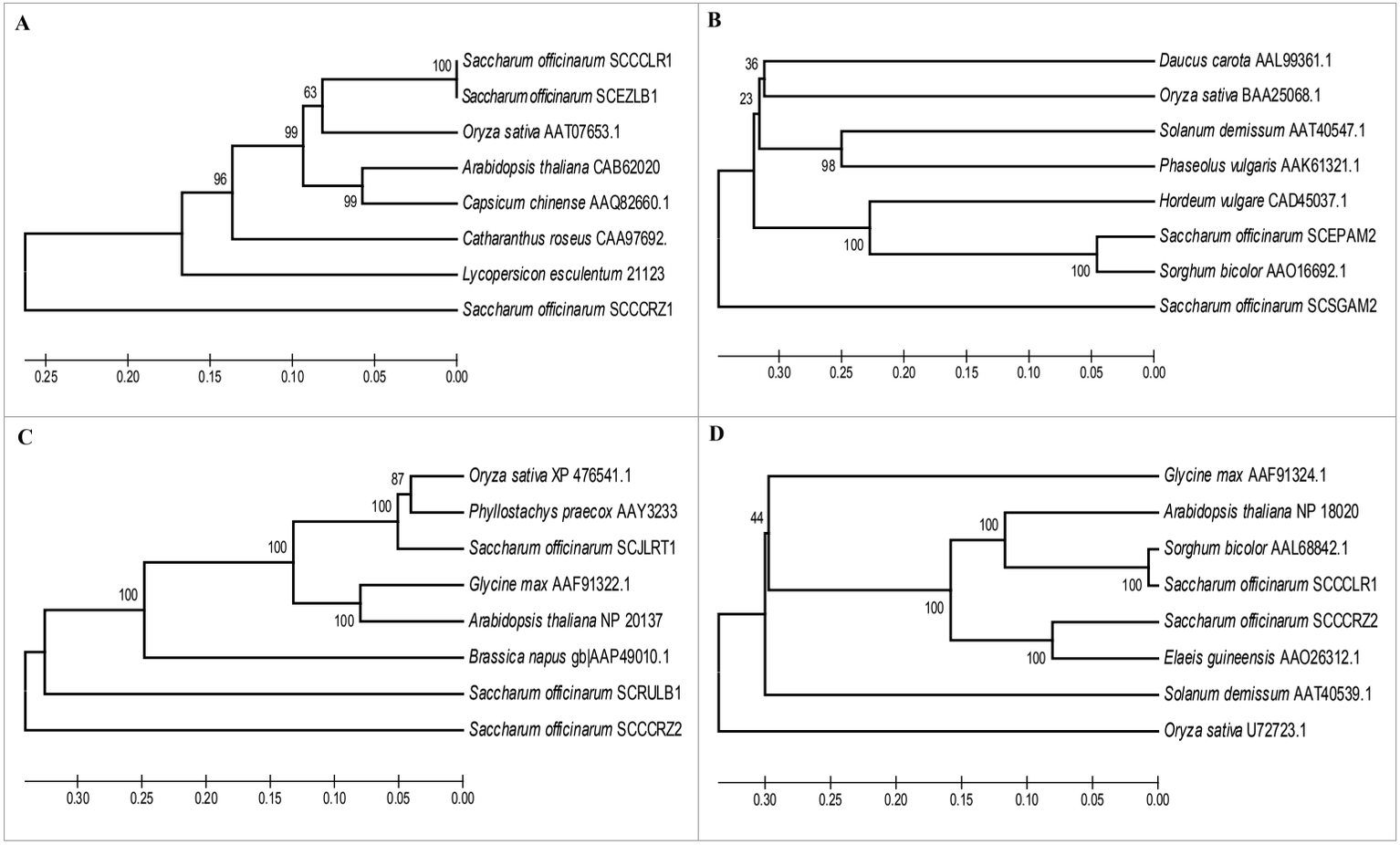

Figure 2. UPGMA analysis of selected dendrograms generated by MEGA 3 program using bootstrap function (1000 replicates). Numbers in the base of clades mean bootstrap values and bar means genetic distance. Codes after taxonomic identities refer to the accession number in NCBI or cluster identity in the case of sugarcane. **A.** *Pto* gene; **B.** *Xa1* gene; **C.** *Cf9* gene, and **D.** *Xa21* gene.

These groupings confirm that *R*-genes, including these four chosen for the phenetic analysis, appeared before divergent evolution of monocots and dicots.

Distribution of expressed sequence tags in the SUCEST libraries

Considering the distribution of the 2108 reads (contained in the 196 clusters selected) in the 13 libraries analyzed, a higher prevalence could be observed in flower (FL, 28%), stem-root transition tissues (RZ = 12.47%) and apical meristem (AM = 12%). One would expect to find higher levels of expression in root (RT), since this is the main entrance for many pathogenic bacteria, fungi and nematodes. Considering both libraries, RT and RZ tissues together included 18% of the *R*-genes expressed. Keeping in mind that the tissues had been cultivated under controlled non-stress conditions, one may suppose that these results represent *R*-genes that are regularly expressed in root and root to shoot tissues.

Surprisingly, a higher prevalence of *R*-genes could be detected in growing tissues, such as AM leaf roll (LR) and lateral bud (LB), which altogether comprised 24% of the expressed *R*-genes. If we consider also flower libraries (that included five different early stages of development) as young/growing tissues, 46% of the expressed *R*-genes annotated here were expressed within this group (Figure 3A). Otherwise, FL libraries together comprised 63,774 reads (26.8% of all SUCEST reads), which may explain the presence of some clusters comprising reads mainly from FL libraries.

The lowest prevalence of reads representing *R*-genes was observed in leaves (LV, 1%), tissues infected with *H. rubrisubalbicans* (HR, 2%), callus (CL, 4%) and seed (SD, 4%) libraries, respectively (Figure 3A,B).

Striking differences could be observed between the prevalence of *R*-genes in the two libraries obtained under influence of microorganisms (tissues infected by *G. diazotroficans* and with *H. rubrisubalbicans*). While the first showed a significant presence of *R*-genes (7%) identified among the 18,144 reads sequenced (7.6% of all SUCEST reads), the second (that comprised 12,000 reads/5% of all SUCEST reads) had one of the lowest (2%) levels of expressed *R*-genes. Both libraries (AD1 and HR1) were constructed with plantlets inoculated with *G. diazotroficans* or *H. rubrisubalbicans*, which are endophytic nitrogen-fixing bacteria that naturally colonize sugarcane tissues (Lee et al., 2000).

The association of endophytic diazotrophic bacteria with plants is quite different than other nitrogen-fixing associations. Diazotrophic bacteria colonize intercellular spaces and vascular tissues of most organs of the host plant, without causing visible plant anatomical changes or disease symptoms (Reinhold-Hurek and Hurek, 1998).

It has been described that the endophytic diazotrophs produce plant growth-regulating hormones, such as auxin (Fuentes-Ramirez et al., 1993), and more recently noted, gibberellin (Bastian et al., 1998). The mechanisms involved in the establishment of this particular type of interaction and what kind of molecules mediate signaling between plant and bacteria remain unclear. In addition, very little is known about the role of the plant in the association. Differences in the contribution of biological nitrogen fixation to the plant nitrogen balance in distinct sugarcane cultivars suggest that the plant is controlling, at least in part, the efficiency of the process (Urquiaga et al., 1992). The plant could control bacterial colonization by sending the proper signals and/or providing the best physiological conditions for bacterial survival. An-

other question to be addressed is how the association benefits the plant. The endophytic diazotrophs promote plant growth when inoculated into sugarcane plantlets, possibly by supplying nitrogen and/or plant hormones (Sevilla et al., 2001).

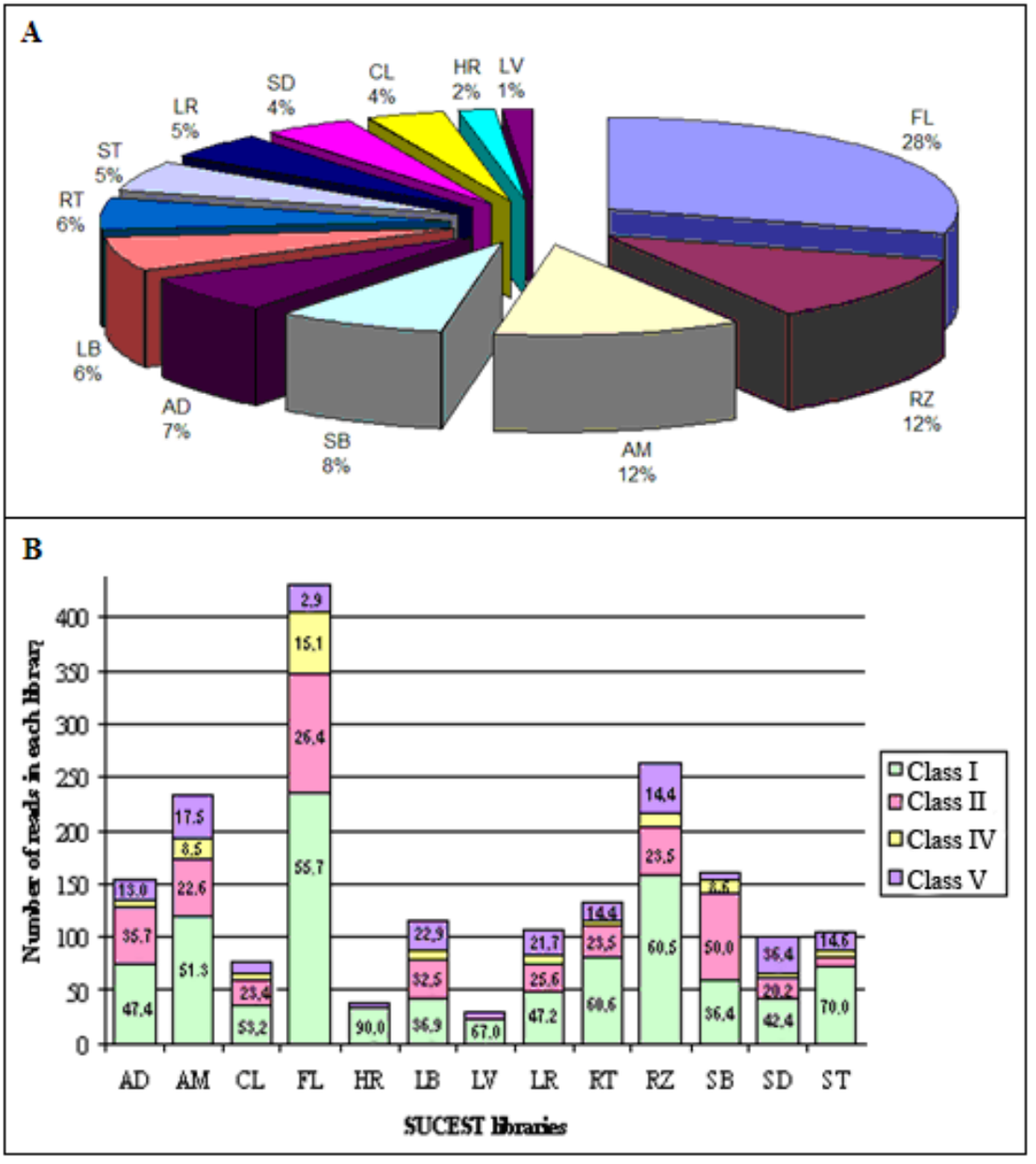

Figure 3. Prevalence of sugarcane *R*-genes in the SUCEST libraries. **A.** General distribution of *R*-gene clusters in the SUCEST libraries. **B.** Prevalence of gene classes in the libraries. Numbers inside columns refer to the percentage of reads in each library for each *R*-gene class. Library code: AD: tissues infected by *Gluconacetobacter diazotroficans*; AM: apical meristem; CL: callus; FL: flower; HR: tissues infected with *Herbaspirillum rubrisubalbicans*; LB: lateral bud; LR: leaf roll; LV: leaves; RT: root; RZ: stem-root transition; SB: stalk bark; SD: seeds; ST: stem.

Nogueira et al. (2001) evaluated the SUCEST data regarding both libraries, but for this purpose they pooled both libraries together preventing the identification of distinctive sequence classes between the two. Most functional categories identified in the study included transporters, transcription factors and protein kinases. Our results suggest that the interaction of sugarcane with each bacterial species was clearly distinctive, indicating that the reaction to the infection with *H. rubrisubalbicans* is clearly permissive with a very low level of *R*-gene expression, while *R*-genes seemed to be activated during the interaction with *G. diazotroficans*. These results suggest that a comparative analysis of both libraries using non-infected seedlings as control would reveal different physiological conditions.

Considering the distribution of reads correlated with the classification of *R*-genes (Figure 3B), it is clear that reads of class I (with kinase domain) were most abundant in all libraries, followed by class II (LRR-NBS-CC) and class V (kinase-TM-LRR), with exception of FL library and stalk bark where the third more representative group was class IV (TM-LRR).

On the contrary, leaves - traditionally the main entrance for viral infections - comprised only 1% of the sequenced *R*-genes. Otherwise, one may consider that this was one of the smallest libraries, comprising only 6342 reads (2.7% of all SUCEST reads). However, all tissues showed *R*-gene expression, suggesting that *R*-genes are overall expressed at low constitutive levels. This is in accordance with the observation of Tang et al. (1999) who detected a basal *Pto*-kinase activity maintained at a low level even when *avrPto* was not present. In the presence of the pathogen, *Pto*-kinase is immediately available and its abundance immediately increased. In contrast, *Xa1* (a gene that confers resistance to rice against *Xantomonas oryzae* pv. *oryzae*) mRNA was detected from rice leaves at 5 days after cutting and inoculation of both compatible and incompatible strains of *Xantomonas oryzae* pv. *oryzae*, but was not detected in intact leaves (Yoshimura et al., 1998). These findings suggest that *R*-gene expression may be induced either by the stimulus of wounding involved in the pathogen infection or in tissues subjected to attacks.

One of the lowest levels of *R*-gene expression (4%) was observed in *in vitro* cultivated calli treated with contrasting temperatures (cold and hot) and alternating dark and light exposure. This low expression suggests that such abiotic stimuli may recruit some responses that indirectly suppress *R*-gene expression.

Expression pattern

In silico evaluation of gene expression can be inferred only with normalized differential display data, an approach that considers reordered data matrices. This method also allows the identification of clusters bearing similar expression patterns in cDNA libraries, suggesting that they may be co-regulated *in vivo*. Lambais (2001) in studying defense-related proteins (PR-class) in sugarcane, a signal cascade also induced by *R*-genes, argued that genes with similar functions or cDNA libraries are expected to have similar patterns of gene expression and also to cluster together in chromosomes. Classical genetic mapping has demonstrated that *R*-genes tend to be clustered in few chromosomes in the genome (Winter et al., 2000; Benko-Iseppon et al., 2003). In *A. thaliana*, they are clustered in two chromosome arms (The Arabidopsis Genome Initiative, 2000), as shown as well in rice (The Rice Chromosomes 11 and 12 Sequencing Consortia, 2005). The same *R*-genes have been observed clustered and almost in

the same order in tomato (Ku et al., 2000) and chickpea (Benko-Iseppon et al., 2003), confirming that gene order and proximity are important for their proper functionality.

In the case of the present analysis, four different approaches (Figure 4) were used to evaluate sites of expression and patterns of co-regulation of *R*-genes in sugarcane, considering the clusters significantly aligned with the genes *Xa1* (Figure 4A), *Cf* (Figure 4B), *Xa21* (Figure 4C), and *Pto* (Figure 4D). The prevalence of expression in libraries constructed from FL tissues is clear for all four groups, with most clusters co-expressing within this library. In the case of the genes *Xa1* and *Cf*, the same clusters were also expressed in apical meristems and stalk bark, while in *Xa21* almost only flower and apical meristem showed significant expression (Figure 4D).

For *Cf9* and *Xa21*, an almost complete lack of expression was observed in most libraries (Figure 4B,C), while in the case of *Xa1*, higher expression (2- to 10-fold) could be detected in most libraries. *Pto* analogs represent the most interesting case study in sugarcane, since many allelic variants of this group could be found. They appeared expressed in all libraries studied, but a clear prevalence of expression in FL, followed by RZ and also RT could be observed with many clusters co-expressing in these libraries. Also in the case of *Pto*-like clusters, high co-expression levels regarding five sugarcane clusters could be observed in the AD library (seedlings infected with *G. diazotroficans*).

The co-expressed clusters presented in Figure 4 represent important candidate sequences for fine mapping of *R*-gene-rich linkage groups in sugarcane, especially using primers designed for conserved flanking sequences which may reside closely in the same linkage group.

It is interesting to note that also after normalized expression analysis of all four selected *R*-genes, their representation in leaves was very low, suggesting that the level of expression in this tissue may be low (less than 1% in the case of *Xa1* and *Pto*) or absent (*Cf9* and *Xa21*).

CONCLUDING REMARKS

Using bioinformatic tools, it was possible to identify and classify *R*-genes in the sugarcane transcriptome, and also to make some inferences regarding their expression pattern under non-induced conditions. All five classes of *R*-genes with their respective conserved domains could be found in sugarcane except the TIR domain which has been found to be absent in all monocots previously studied.

The 196 identified sequences represent valuable resources for the development of markers for molecular breeding and development of resistance gene analogs or gene-specific markers specific for sugarcane and other related cereal crops. The identified clusters constitute also excellent probes for physical mapping of genes in sugarcane, giving support to genetic mapping programs and synteny studies. This may be especially useful for a comparative mapping between sugarcane and *Sorghum*, getting around the difficulties of mapping a large and complex genome as in the case of sugarcane.

The sequences studied probably represent only part of the diversity and number of *R*-genes that are present in cultivated sugarcane. It is expected that an evaluation of tissues under stress conditions induced by pathogen infection would reveal additional information about *R*-genes and their expression, especially considering the huge size and complexity of the sugarcane genome, as compared with most angiosperms.

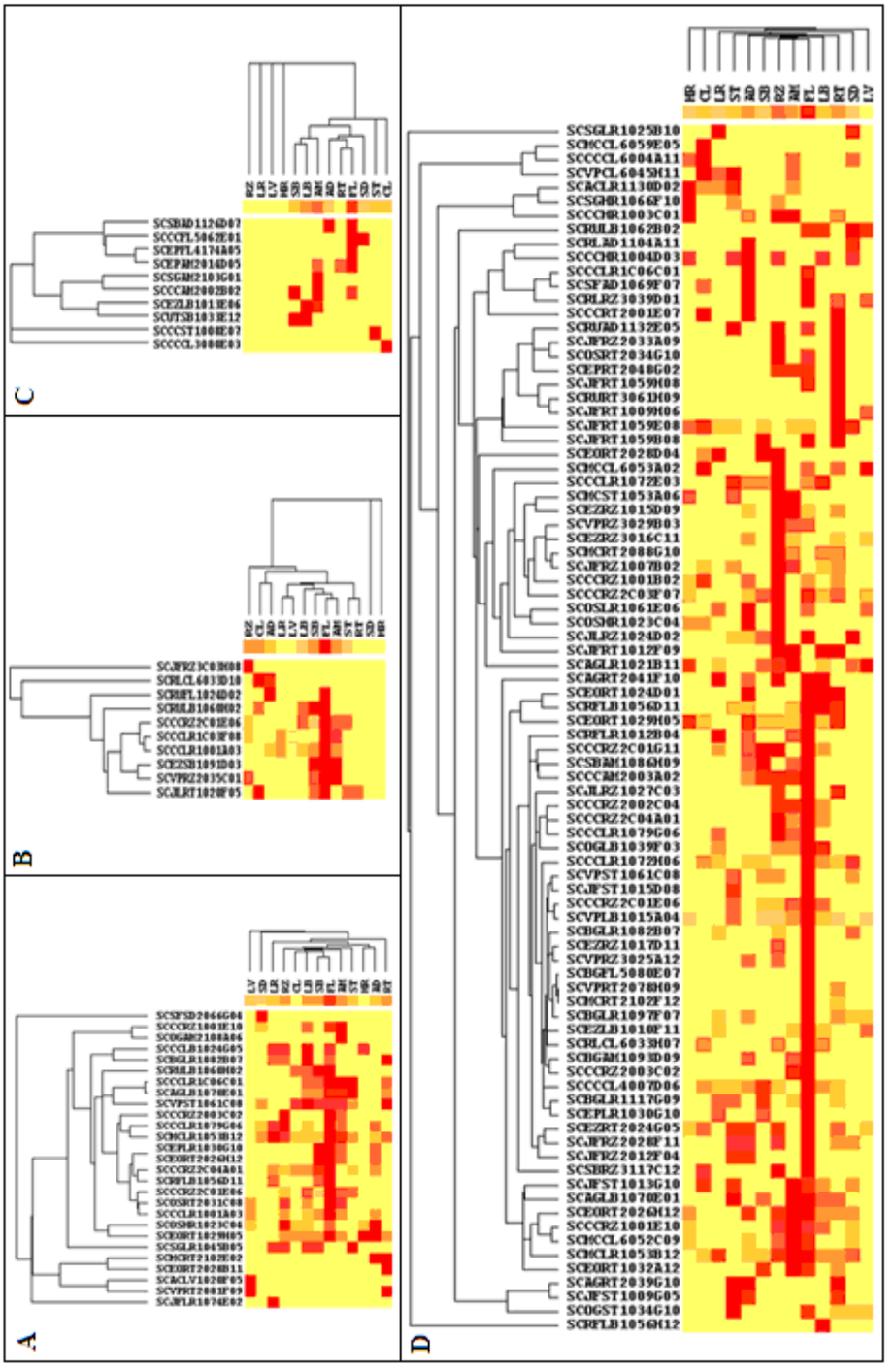

Figure 4. Differential display of sugarcane clusters representing selected *R*-gene classes. **A.** Graphic representing expression in *Xa1*. **B.** *Cf9*. **C.** *Xa21*. **D.** *Pto*. Yellow means no expression and red means all levels of expression. Library codes: AD: tissues infected by *Gluconacetobacter diazotroficans*; AM: apical meristem; CL: callus; FL: flower; HR: tissues infected with *Herbaspirillum rubrisubalbicans*; LB: lateral bud; LR: leaf roll; LV: leaves; RT: root; RZ: stem-root transition; SB: stalk bark; SD: seeds; ST: stem.

Furthermore, it is necessary to manipulate the expression of these genes in economically important plant species in order to improve disease resistance. Although the field is still very much in its infancy, some reports indicate that this strategy is feasible.

ACKNOWLEDGMENTS

We thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FACEPE (Fundação de Amparo à Pesquisa do Estado de Pernambuco) for the fellowships awarded. We are also grateful to FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and SUCEST coordination for the access to the Sugarcane EST databank.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Anderson JC, Pascuzzi PE, Xiao F, Sessa G, et al. (2006). Host-mediated phosphorylation of type III effector AvrPto promotes *Pseudomonas* virulence and avirulence in tomato. *Plant Cell* 18: 502-514.

Bai J, Pennill LA, Ning J, Lee SW, et al. (2002). Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. *Genome Res.* 12: 1871-1884.

Barbosa da Silva A, Wanderley-Nogueira AC, Silva RRM, Belarmino LC, et al. (2005). *In silico* survey of resistance (*R*) genes in *Eucalyptus* transcriptome. *Genet. Mol. Biol.* 28: 562-574.

Bastian F, Cohen A, Piccoli P, Luna V, et al. (1998). Production of indole-3-acetic acid and gibberellins A(1) and A(3) by *Gluconacetobacter diazotrophicus* and *Herbaspirillum seropedicae* in chemically-defined culture media. *Plant Growth Regul.* 24: 7-11.

Benko-Iseppon AM, Winter P, Huettel B, Staginnus C, et al. (2003). Molecular markers closely linked to *fusarium* resistance genes in chickpea show significant alignments to pathogenesis-related genes located on *Arabidopsis* chromosomes 1 and 5. *Theor. Appl. Genet.* 107: 379-386.

Bent AF (1996). Plant disease resistance genes: function meets structure. *Plant Cell* 8: 1757-1771.

Bittner-Eddy PD, Crute IR, Holub EB and Beynon JL (2000). RPP13 is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica*. *Plant J.* 21: 177-188.

Bonas U and Van den Anckerveken G (1999). Gene-for-gene interactions: bacterial avirulence proteins specify plant disease resistance. *Curr. Opin. Microbiol.* 2: 94-98.

Borevitz JO and Ecker JR (2004). Plant genomics: the third wave. *Annu. Rev. Genomics Hum. Genet.* 5: 443-477.

Botella MA, Parker JE, Frost LN, Bittner-Eddy PD, et al. (1998). Three genes of the *Arabidopsis* RPP1 complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. *Plant Cell* 10: 1847-1860.

Brommonschenkel SH, Frary A, Frary A and Tanksley SD (2000). The broad-spectrum tospovirus resistance gene Sw-5 of tomato is a homolog of the root-knot nematode resistance gene *Mi*. *Mol. Plant Microbe Interact.* 13: 1130-1138.

Bryan GT, Wu KS, Farrall L, Jia Y, et al. (2000). A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. *Plant Cell* 12: 2033-2046.

Chinnaiyan AM, Chaudhary D, O'Rourke K, Koonin EV, et al. (1997). Role of CED-4 in the activation of CED-3. *Nature* 388: 728-729.

Dixon MS, Jones DA, Keddie JS, Thomas CM, et al. (1996). The tomato *Cf-2* disease resistance locus comprises two functional genes encoding leucine-rich repeat proteins. *Cell* 84: 451-459.

Dodds PN, Lawrence GJ and Ellis JG (2001). Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the P and P2 rust resistance specificities in flax. *Plant Cell* 13: 163-178.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863-14868.

Ellis J and Jones D (2000). Structure and function of proteins controlling strain-specific pathogen resistance in plants. *Curr. Opin. Plant Biol.* 1: 288-293.

Ellis JG, Lawrence GJ, Luck JE and Dodds PN (1999). Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell* 11: 495-506.

Ernst K, Kumar A, Kriseleit D, Kloos DU, et al. (2002). The broad-spectrum potato cyst nematode resistance gene (*Hero*) from tomato is the only member of a large gene family of NBS-LRR genes with an unusual amino acid repeat in the LRR region. *Plant J.* 31: 127-136.

Expert Protein Analysis System (ExPASy). http://expasy.org. Accessed May 25, 2006.

Ewing RM, Ben KA, Poirot O, Lopez F, et al. (1999). Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* 9: 950-959.

Flor HH (1956). The complementary genetic systems in flax and flax rust. *Adv. Genet.* 8: 29-54.

Flor HH (1971). Current status of the gene-for-gene concept. *Annu. Rev. Plant Pathol.* 9: 275-296.

Fuentes-Ramirez LE, Jimenez-Salgado T, Abarca-Ocampo IR and Caballero-Mellado J (1993). *Gluconacetobacter diazotrophicus*, an indole acetic acid-producing bacterium isolated from sugarcane cultivars in Mexico. *Plant Soil* 154: 145-150.

Gassmann W, Hinsch ME and Staskawicz BJ (1999). The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. *Plant J.* 20: 265-277.

Grant MR, Godiard L, Straube E, Ashfield T, et al. (1995). Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. *Science* 269: 843-846.

Grivet L and Arruda P (2001). Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* 5: 122-127.

Hammond-Kosack KE and Jones JD (1996). Resistance gene-dependent plant defense responses. *Plant Cell* 8: 1773-1791.

Johal GS and Briggs SP (1992). Reductase activity encoded by the *HM1* disease resistance gene in maize. *Science* 258: 985-987.

Jones DG (2001). Putting the knowledge of plant disease resistance genes to work. *Curr. Opin. Plant Biol.* 4: 281-287.

Jones DA, Thomas CM, Hammond-Kosack KE, Balint-Kurti PJ, et al. (1994). Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. *Science* 266: 789-793.

Koczyk G and Chelkowski J (2003). An assessment of the resistance gene analogues of *Oryza sativa* ssp. *japonica*, their presence and structure. *Cell Mol. Biol. Lett.* 8: 963-972.

Kruijt M, Kip DJ, Joosten MH, Brandwagt BF, et al. (2005). The *Cf-4* and *Cf-9* resistance genes against *Cladosporium fulvum* are conserved in wild tomato species. *Mol. Plant Microbe Interact.* 18: 1011-1021.

Ku HM, Vision T, Liu J and Tanksley SD (2000). Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA* 97: 9121-9126.

Kumar S, Tamura K and Nei M (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* 5: 150-163.

Lambais MR (2001). *In silico* differential display of defense-related expressed sequence tags from sugarcane tissues infected with diazotrophic endophytes. *Genet. Mol. Biol.* 24: 1-4.

Lawrence GJ, Finnegan EJ, Ayliffe MA and Ellis JG (1995). The *L6* gene for flax rust resistance is related to the *Arabidopsis* bacterial resistance gene *RPS2* and the tobacco viral resistance gene *N*. *Plant Cell* 7: 1195-1206.

Lee S, Reth A, Meletzus D, Sevilla M, et al. (2000). Characterization of a major cluster of *nif*, *fix*, and associated genes in a sugarcane endophyte, *Acetobacter diazotrophicus*. *J. Bacteriol.* 182: 7088-7091.

Liu JJ and Ekramoddoullah AK (2003). Isolation, genetic variation and expression of TIR-NBS-LRR resistance gene analogs from western white pine (*Pinus monticola* Dougl. ex. D. Don.). *Mol. Genet. Genomics* 270: 432-441.

Loh YT and Martin GB (1995). The *Pto* bacterial resistance gene and the *Fen* insecticide sensitivity gene encode functional protein kinases with serine/threonine specificity. *Plant Physiol.* 108: 1735-1739.

Lu YH, D'Hont A, Paulet F, Grivet L, et al. (1994). Molecular diversity and genome structure in modern sugarcane varieties. *Euphytica* 78: 217-226.

McDowell JM, Dhandaydham M, Long TA, Aarts MG, et al. (1998). Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of *Arabidopsis*. *Plant Cell* 10: 1861-1874.

Mestre P and Baulcombe DC (2006). Elicitor-mediated oligomerization of the tobacco N disease resistance protein. *Plant Cell* 18: 491-501.

Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, et al. (1999). Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.* 20: 317-332.

Meyers BC, Kozik A, Griego A, Kuang H, et al. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 15: 809-834.

Meyers BC, Kaushik S and Nandety RS (2005). Evolving disease resistance genes. *Curr. Opin. Plant Biol.* 8: 129-134.

Milligan SB, Bodeau J, Yaghoobi J, Kaloshian I, et al. (1998). The root knot nematode resistance gene *Mi* from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. *Plant Cell* 10: 1307-1319.

Mindrinos M, Katagiri F, Yu GL and Ausubel FM (1994). The *A. thaliana* disease resistance gene RPS2 encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell* 78: 1089-1099.

Monosi B, Wisser RJ, Pennill L and Hulbert SH (2004). Full-genome analysis of resistance gene homologues in rice. *Theor. Appl. Genet.* 109: 1434-1447.

National Center for Biotechnology Information (NCBI). http://www.ncbi.nlm.nih.gov. Accessed June 7, 2006.

Nelson RR (1972). Stabilizing racial populations of plant pathogens by use of resistance genes. *J. Environ. Qual.* 3: 220-227.

Nogueira EM, Vinagre F, Masuda HP, Vargas C, et al. (2001). Expression of sugarcane genes induced by inoculation with *Gluconacetobacter diazotrophicus* and *Herbaspirillum rubrisubalbicans*. *Genet. Mol. Biol.* 24: 1-4.

Ohtsuki A and Sasaki A (2006). Epidemiology and disease-control under gene-for-gene plant-pathogen interaction. *J. Theor. Biol.* 238: 780-794.

Ori N, Eshed Y, Paran I, Presting G, et al. (1997). The I2C family from the wilt disease resistance locus I2 belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. *Plant Cell* 9: 521-532.

Page RD (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* 12: 357-358.

Pan Q, Wendel J and Fluhr R (2000). Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J. Mol. Evol.* 50: 203-213.

Parker JE, Coleman MJ, Szabo V, Frost LN, et al. (1997). The *Arabidopsis* downy mildew resistance gene *RPP5* shares similarity to the toll and interleukin-1 receptors with *N* and *L6*. *Plant Cell* 9: 879-894.

Piffanelli P, Zhou F, Casais C, Orme J, et al. (2002). The barley *MLO* modulator of defense and cell death is responsive to biotic and abiotic stress stimuli. *Plant Physiol.* 129: 1076-1085.

Pryor T and Ellis J (1993). The genetic complexity of fungal resistance genes in plants. *Adv. Plant Pathol.* 10: 281-305.

Rehmany AP, Gordon A, Rose LE, Allen RL, et al. (2005). Differential recognition of highly divergent downy mildew avirulence gene alleles by *RPP1* resistance genes from two *Arabidopsis* lines. *Plant Cell* 17: 1839-1850.

Reinhold-Hurek B and Hurek T (1998). Life in grasses: diazotrophic endophytes. *Trends Microbiol.* 6: 139-144.

Richter TE and Ronald PC (2000). The evolution of disease resistance genes. *Plant Mol. Biol.* 42: 195-204.

Rommens CM and Kishore GM (2000). Exploiting the full potential of disease-resistance genes for agricultural use. *Curr. Opin. Biotechnol.* 11: 120-125.

Rose LE, Langley CH, Bernal AJ and Michelmore RW (2005). Natural variation in the *Pto* pathogen resistance gene within species of wild tomato (*Lycopersicon*). I. Functional analysis of *Pto* alleles. *Genetics* 171: 345-357.

Rossi M, Araujo PG, Paulet F, Garsmeur O, et al. (2003). Genomic distribution and characterization of EST-derived resistance gene analogs (RGAs) in sugarcane. *Mol. Genet. Genomics* 269: 406-419.

Salmeron JM, Oldroyd GE, Rommens CM, Scofield SR, et al. (1996). Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. *Cell* 86: 123-133.

Salvaudon L, Heraudet V and Shykoff JA (2005). Parasite-host fitness trade-offs change with parasite identity: genotype-specific interactions in a plant-pathogen system. *Evol. Int. J. Org. Evol.* 59: 2518-2524.

Sevilla M, Burris RH, Gunapala N and Kennedy C (2001). Comparison of benefit to sugarcane plant growth and 15N2 incorporation following inoculation of sterile plants with *Acetobacter diazotrophicus* wild-type and Nif- mutants strains. *Mol. Plant Microbe Interact.* 14: 358-366.

Song WY, Wang GL, Chen LL, Kim HS, et al. (1995). A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science* 270: 1804-1806.

Song WY, Pi LY, Wang GL, Gardner J, et al. (1997). Evolution of the rice *Xa21* disease resistance gene family. *Plant Cell* 9: 1279-1287.

Sugarcane EST Genome Project (SUCEST). http://sucest.lad.dcc.unicamp.br. Accessed May 2, 2006.

Tang X, Xie M, Kim YJ, Zhou J, et al. (1999). Overexpression of *Pto* activates defense responses and confers broad resistance. *Plant Cell* 11: 15-29.

The Arabidopsis Genome Iniciative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.

The Rice Chromosomes 11 and 12 Sequencing Consortia (2005). The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. *BMC Biol.* 3: 20.

União dos Produtores de Bioenergia (UDOP). http://www.udop.com.br. Accessed February 12, 2007.

Urquiaga S, Cruz HS and Boddey RM (1992). Contribution of nitrogen fixation to sugarcane: nitrogen-15 and nitrogen balance estimates. *Soil Sci. Soc. Am. J.* 56: 105-114.

Vallad G, Rivkin M, Vallejos C and McClean P (2001). Cloning and homology modeling of a Pto-like protein kinase family of common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* 103: 1046-1058.

van der Biezen EA, Freddie CT, Kahn K, Parker JE, et al. (2002). *Arabidopsis RPP4* is a member of the *RPP5* multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signalling components. *Plant J.* 29: 439-451.

Vettore AL, da Silva FR, Kemper EL and Arruda P (2001). The libraries that made SUCEST. *Genet. Mol. Biol.* 24: 1-7.

Wang GL, Holsten TE, Song WY, Wang HP, et al. (1996). Construction of a rice bacterial artificial chromosome library and identification of clones linked to the *Xa21* disease resistance locus. *Plant J.* 7: 525-533.

Wang ZX, Yano M, Yamanouchi U, Iwamoto M, et al. (1999). The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. *Plant J.* 19: 55-64.

Whitham S, Dinesh-Kumar SP, Choi D, Hehl R, et al. (1994). The product of the tobacco mosaic virus resistance gene *N*: similarity to toll and the interleukin-1 receptor. *Cell* 78: 1101-1115.

Whitham S, McCormick S and Baker B (1996). The *N* gene of tobacco confers resistance to tobacco mosaic virus in transgenic tomato. *Proc. Natl. Acad. Sci. USA* 93: 8776-8781.

Winter P, Benko-Iseppon AM, Hüttel B, Pfaff T, et al. (2000). A linkage map of the chickpea (*Cicer arietinum* L.) genome based on recombinant inbred lines from a *C. arietinum* x *C. reticulatum* cross: localization of resistance genes for *fusarium* wilt races 4 and 5. *Theor. Appl. Genet.* 101: 1155-1163.

Xu WH, Wang YS, Liu GZ, Chen X, et al. (2006). The autophosphorylated Ser686, Thr688, and Ser689 residues in the intracellular juxtamembrane domain of *XA21* are implicated in stability control of rice receptor-like kinase. *Plant J.* 45: 740-751.

Yoshimura S, Yamanouchi U, Katayose Y, Toki S, et al. (1998). Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proc. Natl. Acad. Sci. USA* 95: 1663-1668.

Zhou T, Wang Y, Chen JQ, Araki H, et al. (2004). Genome-wide identification of NBS genes in *japonica* rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol. Genet. Genomics* 271: 402-415.

APPENDIX

Attachment 1. Classification and features of R-gene seed sequences used as query against the SUCEST database.

R-gene class	Accession No.	Source species	Gene	Sequence size (aa)	Domain range (initial-last aa)							
					LRR		Kinase		NBS		TIR	
					Start	End	Start	End	Start	End	Start	End
Class I	2112354A	*Lycopersicon esculentum*	*Pto*	321	-	-	41	236	-	-	-	-
Class II	AF234174_1	*Arabidopsis thaliana*	*HRT*	909	579	868	-	-	150	460	-	-
	NP_172686.1	*Arabidopsis thaliana*	*Rps5*	889	540	636	-	-	140	444	-	-
	AF118127_1	*Lycopersicon esculentum*	*I2*	1266	578	1231	-	-	154	457	-	-
	AAG31014.1	*Lycopersicon esculentum*	*Sw5*	1246	-	-	-	-	519	818	-	-
	BAA25068.1	*Oryza sativa*	*Xa1*	1802	771	1773	-	-	283	593	-	-
	BAA76282	*Oryza sativa*	*Pib*	1251	930	1210	-	-	390	712	-	-
	AAK00132	*Oryza sativa*	*Pi-ta*	928	712	925	-	-	207	503	-	-
	AAP81262.1	*Zea mays*	*Rp1*	1269	596	1228	-	-	145	457	-	-
	AC016827_19	*Arabidopsis thaliana*	*RPM1*	926	559	665	-	-	173	467	-	-
Class III	AAC72977.1	*Arabidopsis thaliana*	*RPP1*	1189	668	1011	-	-	226	505	54	184
	RP13_ARATH	*Arabidopsis thaliana*	*RPP13*	835	-	-	-	-	147	453	14	148
	AF440696_1	*Arabidopsis thaliana*	*RPP4*	1135	642	1053	-	-	185	441	15	145
	AAF08790.1	*Arabidopsis thaliana*	*RPP5*	1361	643	1151	-	-	188	465	14	148
	RPP8_ARATH	*Arabidopsis thaliana*	*RPP8*	908	577	867	-	-	149	459	15	145
	BAB11393.1	*Arabidopsis thaliana*	*Rps4*	1232	663	889	-	-	198	473	21	149
	AAP41025.1	*Lactuca serriola*	*RGC2*	352	49	235	-	-	-	-	21	149
	AF093649_1	*Linum usitatissimum*	*L*	1294	607	1277	-	-	220	521	63	195
	T18548	*Linum usitatissimum*	*M*	1305	744	1288	-	-	235	534	78	210
	AF310960_2	*Linum usitatissimum*	*P*	1211	693	1023	-	-	205	238	23	153
	AF202179_1	*Capsicum chacoense*	*Bs2*	905	-	-	-	-	152	439	63	195
	A54810	*Nicotiana glutinosa*	*N*	1144	597	908	-	-	172	447	14	147
	AF195939_1	*Solanum tuberosum*	*Gpa2*	912	561	863	-	-	119	422	14	147
	CAA61264.1	*Solanum tuberosum*	*Rx1*	248	-	-	-	-	-	-	23	153
	CAB56299.1	*Solanum tuberosum*	*Rx2*	938	561	859	-	-	138	422	78	210
	CAD29728.1	*Solanum tuberosum*	*HERO*	1283	-	-	-	-	504	811	54	184
Class IV	T07015	*Lycopersicon esculentum*	*Cf4*	855	81	758	-	-	-	-	-	-
	AAC78591.1	*Lycopersicon esculentum*	*Cf5*	968	96	855	-	-	-	-	-	-
Class V	AAC80225	*Oryza longistaminata*	*Xa21*	1025	83	333	708	922	-	-	-	-

The genes used are grouped into five *R*-gene classes (I: kinase; II: LRR + NBS; III: LRR + NBS + TIR; IV: only LRR; V: LRR + kinase) with respective accession number at NCBI, source species, gene name, and domain range (in amino acids, aa). LRR = leucine-rich repeats; NBS = nucleotide-binding site; TIR = Toll interleukin-1 receptor domain.