Genetics and Molecular Biology, 28, 3 (suppl), 562-574 (2005)

www.sbg.org.br

Research Article

In silico survey of resistance (*R*) genes in *Eucalyptus* transcriptome

Adriano Barbosa-da-Silva, Ana C. Wanderley-Nogueira, Raphaela R.M. Silva, Luiz C. Berlarmino, Nina M. Soares-Cavalcanti and Ana M. Benko-Iseppon

Universidade Federal de Pernambuco, Centro de Ciências Biológicas, Departamento de Genética, Laboratório de Genética e Biotecnologia Vegetal, Recife, PE, Brazil.

Abstract

A major goal of plant genome research is to recognize genes responsible for important traits. Resistance genes are among the most important gene classes for plant breeding purposes being responsible for the specific immune response including pathogen recognition, and activation of plant defence mechanisms. These genes are quite abundant in higher plants, with 210 clusters found in *Eucalyptus* FOREST database presenting significant homology to known R-genes. All five gene classes of *R*-genes with their respective conserved domains are present and expressed in *Eucalyptus.* Most clusters identified (93) belong to the LRR-NBS-TIR (genes with three domains: Leucine-rich-repeat, Nucleotide-binding-site and Toll interleucine 1-receptor), followed by the serine-threonine-kinase class (49 clusters). Some new combinations of domains and motifs of *R*-genes may be present in *Eucalyptus* and could represent novel gene structures. Most alignments occurred with dicots (94.3%), with emphasis on *Arabidopsis thaliana* (Brassicaceae) sequences. All best alignments with monocots (5.2%) occurred with rice (*Oryza sativa*) sequences and a single cluster aligned with the gymnosperm *Pinus sylvestris* (0.5%). The results are discussed and compared with available data from other crops and may bring useful evidences for the understanding of defense mechanisms in *Eucalyptus* and other crop species.

Key words: serine-threonine kinase, nucleotide binding site, leucine-rich repeats, gene-for-gene interaction.

Received: May 28, 2004; Accepted: March 28, 2005.

Introduction

Pathogen attack can severely affect crop production, with losses that can achieve 80% of the production especially in tropical countries. At the global level, losses have been estimated to accomplish around 12% of the world crop production (James *et al.*, 1990). The most important group of genes that has been used by breeders for disease control is the plant resistance (*R*) genes: single determinant of an effective and specific resistance that can often be characterized by localized necrosis at attempted infection sites (Rommens and Kishore, 2000).

It is proposed that pathosystems are usually highly specific, with a matching *R*-gene on vegetal cell that recognizes elicitor proteins (called Avr-effector) of each infective pathogen. Plant will be resistant and the growth of the pathogen will be arrested only when both genes, *R* and *Avr*, are present (Ellis *et al.*, 2000a). So, for each *R*-gene a correspondent *Avr* gene co-exists: this is the basis of the gene-for-gene concept, suggested by Flor (1956, 1971).

Send correspondence to Ana Maria Benko-Iseppon, Universidade Federal de Pernambuco, Centro de Ciências Biológicas, Departamento de Genética, Laboratório de Genética e Biotecnologia Vegetal, Av. Prof. Moraes Rego s/ n., 50732-970 Recife, PE, Brazil. E-mail: celisep@hotlink.com.br.

Avirulence gene products actually described do not comprise a defined family of related proteins, since no sharing similar motifs or domains could be found. On the opposite, *R*-gene products are separated into distinct but related protein classes, according to their conserved structural domains. Conserved domain function identified for R proteins suggests two fundamental mechanisms during pathogenic infection: (I) the pathogen recognition, conducted mainly by leucine-rich repeats (LRR) regions, which play a direct role in protein-protein specific recognition event; and (II) signaling of pathogen presence in order to activate defense related genes (Richter and Ronald, 2000).

The TIR (Toll interleukine 1-receptor) and CC (coiled coil) regions are involved in signal transduction during many cell processes (Martin *et al.*, 2003), while the NBS (Nucleotide Biding Site) usually signalizes for programmed cell death in animal cells (van der Biezen and Jones, 1998). Additionally, a kinase catalytic region is present in some *R*-genes. This domain plays a direct role in both signaling processes and pathogen effectors. Additionally the NBS region contains not only the three motifs involved in nucleotide binding but additional motifs as well. This extended region of homology is referred to as the NB-ARC domain (Richter and Ronald, 2000). Sometimes this do-

main contains a distinct predicted nucleoside triphosphatase (NTPase) domain known as NACHT, common in animal, fungal and bacterial proteins, implicated with apoptosis induction and transcription activation (Koonin and Avarind, 2000).

Resistance genes are members of a very large multigene family, are highly polymorphic and have diverse recognition specificities. They are commonly clustered in the genome, often in tandem direct repeats, what is consistent with the theory that they originated through gene duplication and that they are continuously evolving through unequal exchange (Song *et al.*, 1997).

Most of the resistance genes that have been cloned and characterized resemble components involved in signal transduction. These can be classified into five categories based on their predicted protein structure (Song *et al.*, 1997, Ellis and Jones, 1998).

The first class is represented by the *Pto* gene of tomato, which encodes a protein with a catalytic serine-threonine kinase (ser-thre-kinase) and a myristoylation motif in his amino terminal region (Martin *et al.*, 1993).

The second class comprises many proteins that present a region rich in repetitions of leucine (LRR, Leucine-rich repeats), a Nucleotide Binding Site (NBS) and a leucine zipper (LZ) or a coiled-coil (CC) sequence. Many genes encode proteins of this class: *I2* (Ori *et al.*, 1997), *Mi* (Milligan *et al.*, 1998) and *Sw5* (Brommonschenkel *et al.*, 2000) from tomato; *RPM1* (Grant *et al.*, 1995), *RPP8* (McDowell *et al.*, 1998), *RPS2* (Mindrinos *et al.*, 1994) and *RPP13* (Bittner-Eddy *et al.*, 2000) from *Arabidopsis thaliana*; *Pib* (Wang *et al.*, 1999), *Pi-ta* (Bryan *et al.*, 2000) and *Xa1* (Yoshimura *et al.*, 1998) from *Oryza sativa* (rice); *Gpa2* (Van der Vossen *et al.*, 2000), *Hero* (Ernst *et al.*, 2002), *R1* (Ballvora *et al.*, 2002), *Rx1a* (Bendahmane *et al.*, 1995) and *Rx2* (Bendahmane *et al.*, 2000) from potato; *Rp1* from maize (Collins *et al.*, 1999); *Mla* from barley (Halterman *et al.*, 2001) and *Dm3* from lettuce (Meyers *et al.*, 1998).

The third class includes similar proteins as described for class II, presenting a toll receptor for interleukine-I (IL-1R) instead of a CC sequence at the amino terminal region (Meyers *et al.*, 1999). This class is referred as TIR-NBS-LRR, including the genes *L* (Lawrence *et al.*, 1995), and *P* (Dodds *et al.*, 2001) of flax; *RPP1* (Botela *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* (Whithan *et al.*, 1996) of tobacco. This class (also present in animals) is supposed to be absent in monocotyledonous plants (Ellis and Jones, 1998), being present in all dicotyledonous taxa actually studied.

The proteins encoded by the three classes of genes previously cited do not present a transmembrane sequence and are therefore classified as intracellular *R*-proteins (Martin *et al.*, 2003).

The fourth class of resistance genes belongs to the tomato *Cf*-family, encoding similar proteins with an extracellular LRR and a short cytoplasmatic tail, but no NBS or any further recognizable domain (Dixon *et al.*, 1996). Member of this family are *Cf-2* (Dixon *et al.*, 1998), *Cf-4* (Joosten *et al.*, 1994; Thomas *et al.*, 1997), *Cf-5* (Dixon *et al.*, 1998) and *Cf-9* (Jones *et al.*, 1994).

The fifth class includes a single gene, *Xa21* from rice that presents an extracellular LRR, a transmembrane region (TM) and a cytoplasmatic ser-thre-kinase. Thus, the structure of *Xa21* indicates an evolutionary link between different classes of plant disease resistance genes (Song *et al.*, 1997).

There is still a sixth class that presents genes with no conserved domains, as described for the previous five classes. This group comprises the gene *Hm1* from maize, a reductase that confers resistance to the fungus *Cochliobolus carbonum* (Johal and Briggs, 1992); *Mlo* from barley, a putative regulator of defense against *Blumenaria graminis* (Piffanelli *et al.*, 2002) possibly associated to the plasma membrane (Buschges *et al.*, 1997); and *RPW8* from *A. thaliana*, that confers non-specific resistance to the fungus *Erysyphe chicoracearum* (Xiao *et al.*, 2001).

Due to its qualities as high level of adaptability, fast growing capacity and wood quality, *Eucalyptus* plantations are carried out in all tropical areas in diverse continents. *Eucalyptus* is the most widely used tree for delivering raw material for the paper industry used in the production of cellulose and to regenerate degraded areas. Over the past 50 years large-scale planting of fast growing exotic *E. grandis*, *E. urophyla*, *E. saligna* and many hybrids (particularly *grandis* x *urophyla*) has occurred in Brazil aiming to reforest some regions and to create an adequate supply of wood, timber and fuel for different purposes (McNabb, 2002). In the late 2001s growing areas reached 138.132 ha, generating more than 7,398 direct employments (BRACELPA, 2004).

The advance of plantations to hot and humid areas resulted in favourable conditions to the development of diseases especially in young individuals that are often severely attacked by fungal (*e.g. Mycosphaerella cryptica*, *Dichomera versiformis*, *Cylindrocladium* spp. and *Phaeophleospora epicpccoides*) and bacterial pathogens (Barber *et al.*, 2003, Mafia and Alfenas, 2003).

Eucalyptus Genome Sequencing Consortium (FOREST) aimed to identify over 15,000 expressed genes from 100,000 sequenced EST from 19 libraries from specific tissues and stages.

The present work aimed to perform a data mining-based identification of plant disease *R*-genes in FOREST database, by using well known *R*-genes sequences as template, comparing the identified sequences with known *R*-genes deposited in public DNA and protein databases.

Materials and Methods

Amino-acid sequences of known genes have been used as query in the search for *R*-gene homologues and analogs in *Eucalyptus* transcriptome database. Accession numbers at NCBI (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov) of sequences used are shown in Table 1, together with sequences features and accession numbers. They are grouped according to the conserved domains previously described. Members of the sixth class (reductases and other *R*-genes with no recognizable conserved domains) have not been included in the present evaluation.

All *Eucalyptus* sequences used during this work were obtained from FOREST project and derived from cDNA libraries specific to different tissues, organs or conditions of growth from the species *E. grandis*, *E. globulus*, *E. saligna* and *E. urophylla*. For detailed information see https://forests.esalq.usp.br/Librariesinfo.html**.**

Reverse alignments were realized on 'FOREST EG_Clusters' database using the program TBLASTN (Altschul *et al.*, 1990), the e-value cutoff adopted was $1e^{-23}$. Matching clusters to query sequences were then annotated on a local database called 'non-redundant' made with aid of the Microsoft Access® program. Cluster name was adopted as primary key in order to prevent data redundancy regarding clusters aligning with more than one query sequence. In the few cases when this occurred the name of both queries has been also annotated for the respective cluster.

The clusters frame of the TBLASTN alignment was used to predict the Open Reading Frames (ORFs) for each searched cluster. For this purpose, the Expasy Translate Tool (bo.expasy.org/tools/dna.html) was used, which predicts the correct ORF for a DNA sequence in the corresponding amino acid FASTA sequence. The obtained ORFs were subsequently submitted to a Reverse Position Specific BLAST (RPS-BLAST) against Conserved Domain Database (Marchler-Bauer *et al.*, 2002) aiming to identify patterns or motifs in predicted cluster products.

Reciprocal alignments were conducted for ORFs by downloading the nr databank and stand alone BLAST package from NCBI ftp site for local use at our server (Laboratório de Genética e Biotecnologia Vegetal, UFPE) performing a high-throughput alignment approach. Matched sequences were annotated for latter comparison.

Predictions of subcellular localization have been inferred by using TargetP program available at CBS (Center for Biotechnology Sequence Analysis) Prediction Servers

Table 1 - Classification and features of *R*-genes used as query against the FOREST database. The used genes are grouped in 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).

Class of *R*-gene	Accession number	Source species	Gene name	Sequence size	Domain range (initial-last aa)							
					LRR		Kinase		NBS		TIR	
				(aa)	Start	End	Start	End	Start	End	Start	End
I	2112354A	*Lycopersicon esculentum*	*Pto*	321	-	-	41	236	-	-	-	-
	AF234174_1	*Arabidopsis thaliana*	*HRT*	909	579	868	-	-	150	460	-	-
	NP_172686.1	*Arabidopsis thaliana*	*Rps5*	889	540	636	-	-	140	444	-	-
II	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	-	-
	AAP81262.1	*Zea mays*	*Rp1*	1269	596	1228	-	-	145	457	-	-
	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
III	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
IV	T07015	*Lycopersicon esculentum*	*Cf4*	855	81	758	-	-	-		-	-
	AAC78591.1	*Lycopersicon esculentum*	*Cf5*	968	96	855	-	-	-	-	-	-

site (http://www.cbs.dtu.dk/services/). Additionally, transmembrane helix segments were inferred with aid of the TMHMM program as well.

Results

After the TBLASTN alignments performed at FOREST EG_Clusters database, a total of 478 clusters aligned with the diverse *R*-genes (Table 1) used as query (data not showed). These clusters were, as described in section 'Material and Methods', inserted on a local database called 'non-redundant'. This procedure generated a set of 210 non-redundant clusters which have been annotated for one or more than one *R*-gene (data summarized in Figure 1 and Tables 2 and 3).

Clusters representing exclusive *R*-gene classes were: (I) serine-threonine kinase (here named KINASE): 49; (II) LRR+NBS: 21; (III) LRR+NBS+TIR: 93; (IV) Only LRR + Transmembrane (LRR+TM): 17 and (V) LRR+TM+ Kinase: 8 (Figure 1).

Regarding the sequence identity of the best alignment, 22 clusters showed equally significant similarity to two different classes of *R*-genes. From these, 18 included LRR plus LRR-Kinase here called MIX I (sequence data presented in Table 3); three included NBS-LRR plus TIR-NBS-LRR (called MIX II) and one LRR plus Kinase (called MIX III).

Sizes of *Eucalyptus* clusters aligned to *R*-genes varied from 3,316 (cluster EGEQRT3301C03 classified to group MIX-III) to 520 nucleotides. The prediction of clusters coding regions revealed that ORFs were coded in both forward and reverse reading frames, with an average of 304 amino acids (aa) in length. ORF sizes varied from 990 (cluster EGEQRT3301C03 of the LRR-KINASE class) to 134aa. Regarding the average ORF length in each *R*-gene class, we observed 417aa for KINASE, 276aa for NBS, 238aa for TIR-NBS-LRR, 247aa for LRR-TM, 352aa for LRR-KINASE, 372aa for MIX I, 343aa for MIX II and 990aa for MIX III class.

The search for conserved domains (CD-Search) revealed conserved regions (Figure 1, Table 1) in 166 of the 210 here analyzed clusters. A total of 40 clusters presented the kinase domain, 37 of them matched to *Pto* gene (class I) after the TBLASTN alignment, with only three grouping into KINASE-LRR (two of them) and MIX III (one of them) classes. These two classes also showed associated LRR segments as well. Regarding the LRR domains, these could be identified in 67 different clusters in all classes (except KINASE class I, represented by *Pto*) with a total of 442 occurrences. This number is higher than the number of clusters due to their occurrence in tandem repetitions. Sometimes these sequences are imperfect and may be difficult to recognize with available *in silico* tools, so it is possible that a larger number may be identified manually.

Twenty clusters showed the NB-ARC domain. In a specific case, this domain occurred associated to a different TIR domain as was cited above. Additionally, a NACHT domain (closed-related to NB-ARC) was identified exclu-

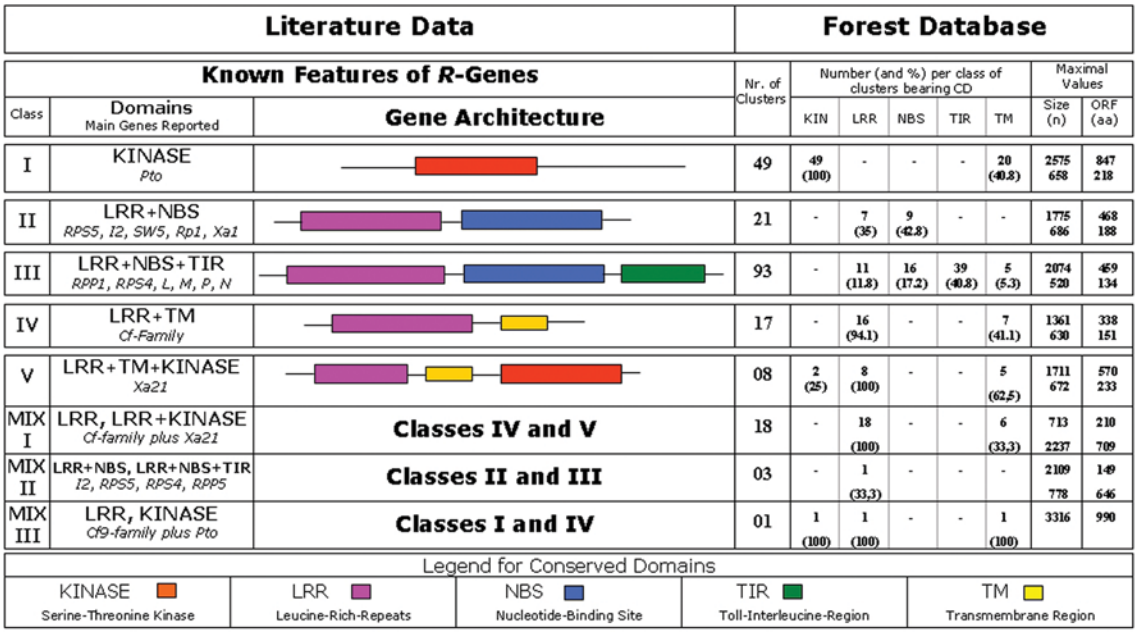

Figure 1 - Representation of main *R*-genes classes considering the presence and position of conserved domains from literature data, as compared with *Eucalyptus* clusters from FOREST database. For each class the data about significant alignments to *R*-genes is given, including following information: number of clusters identified for each class (clusters aligning with more than one class are not included), number and percentage of clusters per class bearing indicated conserved domains, size range (maximal and minimum) of sequence in nucleotides (n) and of ORF in amino-acids (aa). Abbreviation: CD = Conserved domains.

Table 2 - Blast results and sequence evaluation of *Eucalyptus* R genes, including the best matches of each *R* gene and MIX classes: (I) data about the query: gene class and name, NCBI gi|-number, species and family. (II) Features and evaluation results of *Eucalyptus* clusters related to R-genes: cluster number, cluster size in nucleotides (n), ORF (Open Reading Frame) size in amino-acids (aa), e-value; score and frame.

(I) Query Information				(II) Cluster features and evaluation					
Gene class & expected domain	Gene name	NCBI gi|-nr.	Plant species and family	*Eucalyptus* cluster n.	Size (n)	ORF (aa)	E-value	Score and frame	
Class I KINASE	*Pto*	27754635	*Arabidopsis thaliana*	EGEQRT3100D07	2460	722	0.0	1036,6	2
	Pto	15235204	*Arabidopsis thaliana*	EGEQRT3104A12	2575	847	0.0	909,8	1
	Pto	18418211	*Arabidopsis thaliana*	EGUTFB1098H02	2511	616	0.0	904,0	3
	Pto	10177052	*Arabidopsis thaliana*	EGCBRT3133E11	1728	575	0.0	738,0	3
	Pto	25405628	*Arabidopsis thaliana*	EGMCRT3148C12	1705	568	0.0	709,9	1
Class IV LRR	*Cf5*	14626935	*Gossypium hirsutum*	EGEQSL5001G09	1223	321	$2.00e^{-139}$	496,1	2
	Cf5	15240263	*Arabidopsis thaliana*	EGCCRT3339F06	922	307	$3.8e^{-83}$	309,3	2
	Cf5	15239124	*Arabidopsis thaliana*	EGCBST2063A06	697	232	$2.0e^{-61}$	236,5	-2
	Cf4, Cf5	27754637	*Arabidopsis thaliana*	EGACRT3321G06	1361	338	$1.4e^{-60}$	234,6	3
	Cf4, Cf9	14269077	*Lycopersicom esculentum*	EGJMCL1299H10	682	226	$1.9e^{-53}$	209,9	3
Class V LRR KIN	*Xa21*	9651941	*Glycine max*	EGRFRT3357D01	1584	527	0.0	869,4	3
	Xa21	15239540	*Arabidopsis thaliana*	EGEQCL1200B12	1711	570	0.0	658,7	1
	Xa21	19881587	*Oryza sativa*	EGJEST2023F09	716	238	$5.6e^{-51}$	201,8	1
	Xa21	15218385	*Arabidopsis thaliana*	EGSBCL1280C05	725	241	$7.8e^{-48}$	191,4	3
	Xa21	15218385	*Arabidopsis thaliana*	EGSBCL1280C05	725	241	$3,00e^{-04}$	191.4	3
Class II NBS LRR	*Hrt, I2, Sw5, Xa1, Rp1, R1*	18652501	*Oryza sativa*	EGUTRT3110A12	1041	346	$1.2e^{-62}$	241,5	3
	I2, Xa1, Rpm1, Rp1, R1, Pib, Mi1	28300299	*Manihot esculenta*	EGJFSL4202E08	876	291	$5.3e^{-55}$	215,7	3
	Bs2, Gpa2, I2, Rx1, Rx2, Sw5, Xa1, Rp1Mi1	15487949	*Theobroma cacao*	EGEQCL1001F08	934	311	$2.1e^{-52}$	207,2	1
	Gpa2, Rx2, Rpm1, R1, Pib, I2	28300299	*Manihot esculenta*	EGJERT3026C12	804	267	$7.7e^{-50}$	198,4	2
	Gpa2, Hrt, Rpp13, Rpp8, Rx2, Sw5, Rpm1R1, Pi-Ta, Pib	22775643	*Oryza sativa*	EGCECL1282E03	779	231	$1.1e^{-48}$	194,1	1
Class III TIR NBS LRR	*L, M, N, P, Rpp1, Rpp4, Rpp5, Rps4*	7488903	*L. usitatissimum*	EGJMFB1107C10	1395	445	$1.7e^{-83}$	311.2	1
	L, M, N, P, Rpp1, Rpp4, Rpp5, Rps4	9965103	*Glycine max*	EGMCLV2264D03	1155	329	$7.4e^{-75}$	282,0	3
	L, M, N, P, Rpp1, Rpp4, Rpp5, Rps4	12056928	*Glycine max*	EGJEST2234G10	1270	420	$8.8e^{-74}$	278,9	-1
	L, M, N, P, Rpp1, Rpp4, Rpp5, Rps4	27764536	*Glycine max*	EGJMST6019E06	1155	351	$3.0e^{-69}$	263,5	1
	L, M, N, P, Rpp1, Rpp5, Rps4	23477203	*Populus balsamifera*	EGCBRT6029A01	1227	378	$2.6e^{-66}$	253,8	1
MIX I (LRR and LRR-KIN)	*Cf4, Cf5, Cf9, Xa21*	25287710	*Arabidopsis thaliana*	EGBMRT3129F10	2129	709	0.0	705,3	2
	Cf4, Cf5, Cf9, Xa21	21391894	*Lycopersicon peruvianum*	EGCEST2256F04	1687	561	0.0	704,1	3
	Cf4, Cf5, Cf9, Xa21	15240215	*Arabidopsis thaliana*	EGUTFB1136E01	1911	595	$5.00e^{-172}$	605,9	2
	Cf4, Cf5, Cf9, Xa21	15240528	*Arabidopsis thaliana*	EGEZST2207A10	2237	465	$2.00e^{-143}$	510,8	1
	Cf5, Xa21	15230539	*Arabidopsis thaliana*	EGEQRT3201E07	1229	376	$4.00e^{-127}$	455,7	1
MIX II TIR-NBS-LRR and NBS-LRR	*Rpp5, Rps4, I2*	15218365	*Arabidopsis thaliana*	EGEZRT3006B12	2109	646	$1.1e^{-50}$	203,0	3
	Rpp5, Rps5	15221252	*Arabidopsis thaliana*	EGEQST6001H02	778	234	$8.9e^{-33}$	141,4	1
	Rpp5, Rps5	15487963	*Theobroma cacao*	EGJECL1208G03	871	149	$5.8e^{-24}$	110,5	1
MIX III KINASE and LRR	*Cf5, Cf9, Pto*	26450791	*Arabidopsis thaliana*	EGEQRT3301C03	3316	990	0.0	1293,9	1

sively in two TIR-NBS-LRR related clusters (EGCCCL1328B05.g and EGSBRT3118H01).

Most of the 44 clusters with no conserved domains presented shorter ORFs (262 aa in average), with four of them presenting a putative transmembrane region.

A graphic representation of the distribution of conserved domains as compared with class-grouped clusters is presented in Figure 2.

Considering the best matches to the 210 clusters identified, 198 were from plants of Dicotyledonous families, with emphasis on *A. thaliana*. From monocots only rice (*O. sativa*) sequences appeared as best matches (11 clusters). One of the sequences from MIX III group aligned with *Pinus silvestris* (Gymnosperm), the only non-Angiosperm included in the present study. A comprehensive inventory of all species that aligned with *Eucalyptus* with their taxonomic affiliation and habit (herbaceous or woody) is presented in Table 4.

The post-translational inferences carried out for cluster products (TargetP program) revealed a large number of

Table 3 - FOREST clusters classified in the MIX I group, resembling to genes which belong to LRR and LRR-KINASE classes, including: respective templates (query sequences), cluster number and size in nucleotides (n), ORF-size in amino-acids (aa), range of LRR domain after CD-search, identity and results of the best alignment (BLASTp) in NCBI (GI number, species, score and e-value).

Template	Cluster	Size (n)	ORF (aa)	LRR-domain		GI	Description	Score	E-value
				Start	End				
	EGBMRT3129F10.g	2129	709	139	620	25287710	*Arabidopsis thaliana*	705.3	0.0
	EGUTFB1136E01.g	1911	595	91	523	15240215	*Arabidopsis thaliana*	605.9	$5e^{-172}$
Cf4	EGEZST2207A10.g	2237	465	22	262	15240528	*Arabidopsis thaliana*	510.8	$2e^{-143}$
Cf5	EGEQST2201G12.g	2049	412	36	223	25402587	*Arabidopsis thaliana*	360.9	$1.7e^{-98}$
Cf9	EGUTRT3368G02.g	799	265	38	254	15225805	*Arabidopsis thaliana*	323.2	$2e^{-87}$
Xa21	EGCEST2256F04.g	1687	561	22	501	21391894	*Lycopersicon peruvianum*	704.1	0.0
	EGUTSL4018B05.g	1384	447	96	432	3894385	*Lycopersicon esculentum*	257.3	$3e^{-67}$
Cf5, Cf9	EGSBRT3314G03.g	1263	412	118	380	15223460	*Arabidopsis thaliana*	408.3	$9e^{-113}$
Xa21	EGBMRT3131G11.g	1155	384	2	336	15237312	*Arabidopsis thaliana*	349.4	$4.7e^{-95}$
	EGEQRT3201E07.g	1229	376	106	348	15230539	*Arabidopsis thaliana*	455.7	$4e^{-127}$
	EGABST2047C09.g	773	210	13	180	15225805	*Arabidopsis thaliana*	249.6	$1.9e^{-65}$
	EGBMSL4023G05.g	729	242	4	219	15237426	*Arabidopsis thaliana*	226.9	$1.7e^{-58}$
	EGCBST6013F02.g	808	265	32	254	18700171	*Arabidopsis thaliana*	214.5	$1.0e^{-54}$
Cf5	EGCESL5078H03.g	771	257	27	245	15237426	*Arabidopsis thaliana*	211.1	$1.1e^{-53}$
Xa21	EGBGLV3221H06.g	734	235	24	213	3894383	*Lycopersicon esculentum*	200.7	$1.2e^{-50}$
	EGCBRT6048F01.g	713	237	13	227	3641252	*Malus* X *domestica*	327.4	$8.9e^{-89}$
	EGEZST2003B08.g	1586	353	107	346	21952787	*Oryza sativa* (cv. *japonica*)	304.7	$1.2e^{-81}$
	EGEPSL4003G09.g	857	278	59	251	12054894	*Pinus sylvestris*	240.4	$1.9e^{-62}$

predictions (Figure 3). The reliability class (RC), which is a confidence measure for the prediction, showed that only 11 sequences were defined into RC1 (higher than 80%), and 53 for RC2 (higher than 60%) class. Most of the sequences are predicted to be located at unspecific subcellular localization (133 sequences) while 35, 20 and 19 were predicted to contain mitochondrial targeting, signal and chloroplast transit peptides, respectively (Figure 3).

After evaluation with the TargetP program, sequences with motifs specific for transmembrane anchoring could be identified in 44 of all analyzed sequences. From these 19 belonged to LRR or LRR-KINASE-related sequences and, unexpectedly, five showed to be TIR-NBS-LRR and 20 to be KINASE-related sequences.

Discussion

The reverse alignment (TBLASTN) strategy (Altschul *et al.*, 1997) adopted by our group identified a set of 210 clusters similar to the major classes of disease *R*-genes in the current version of the FOREST database,

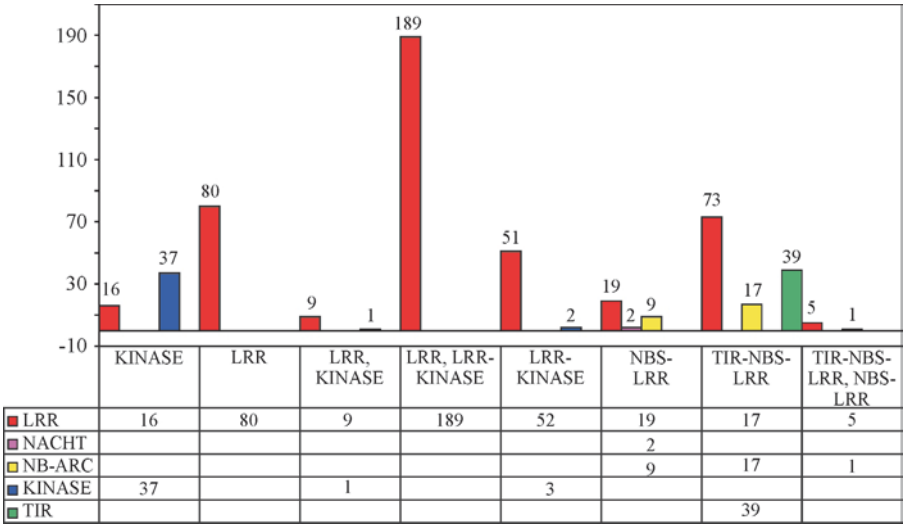

Figure 2 - Graphic representation of the distribution of conserved domains against class-grouped clusters. Values on the base after each domain indicate the number of clusters of each class presenting the indicated domain (also represented in the corresponding columns). Abbreviations: LRR = Leucine-rich-repeats; NB-ARC = Nucleotide-binding-site and additional motifs; NACHT = NB-ARC related domain, including an NTPase implicated in apoptosis and MHC transposition activation.

Table 4 - Inventory of the organisms that appeared as best alignment to each of the 210 here identified *Eucalyptus* clusters related to known resistance genes. The organisms are grouped by gene class (I to V and MIX I to III), taxonomic affiliation (class, subclass, family and species) and habit (herbaceous or woody). Numbers in parenthesis indicate amount of gene members in each taxonomic group or species.

Gene class	Higher taxonomic affiliation	Family	Species	Habit
I KINASE	Dicots (46)	Brassicaceae (43)	*Arabidopsis thaliana* (42)	H
			Brassica napus (1)	H
		Curcubitaceae (1)	*Cucumis melo* (1)	H
		Salicaceae (1)	*Populus nigra* (1)	W
		Solanaceae (2)	*Capsicum anuum* (1)	H
			Nicotiana tacabum (1)	H
	Monocots (2)	Poaceae (2)	*Oryza sativa* (2)	H
II LRR-NBS	Dicots (14)	Asteraceae (1)	*Lactuca sativa* (1)	H
		Brassicaceae (6)	*Arabidopsis thaliana* (6)	H
		Euphorbiaceae (3)	*Manifot esculenta* (3)	W
		Leguminosae (2)	*Glycine max* (1)	H
			Phaseolus vulgaris (1)	H
		Stercullariaceae (2)	*Theobroma cacao* (2)	W
	Monocots (7)	Poaceae (7)	*Oryza sativa* (7)	H
III NBS-LRR-TIR	Dicots (93)	Brassicaceae (10)	*Arabidopsis thaliana* (10)	H
		Curcubitaceae (5)	*Cucumis melo* (5)	H
		Leguminosae (10)	*Glycine max* (10)	H
		Asteraceae (14)	*Helianthus annuus* (14)	H
		Linaceae (34)	*Linum usitatissimum* (34)	H
		Euphorbiaceae (1)	*Manihot esculenta* (1)	W
		Salicaceae (12)	*Populus balsamifera (*10)	W
			Populus tremula (2)	W
		Solanaceae (7)	*Lycopersicon esculentum* (1)	H
			Solanum tuberosum (6)	H
IV LRR	Dicots (17)	Brassicacea (7)	*Arabidopsis thaliana* (7)	H
		Leguminosae (1)	*Glycine max* (1)	H
		Malvaceae (1)	*Gossypium hirsutum* (1)	W
		Solanaceae (8)	*Lycopersicon esculentum* (3)	H
			Lycopersicon hirsutum (2)	H
			Nicotiana tabacum (1)	H
			Petunia X *hybrida* (1)	H
			Solanum tuberosum (1)	H
V LRR-KINASE	Dicots (7)	Brassicaceae (6)	*Arabidopsis thaliana* (6)	H
		Leguminosae (1)	*Glycine max* (1)	H
	Monocot (1)	Poaceae (1)	*Oryza sativa* (1)	H
MIX I	Dicots (16)	Brassicaceae (12)	*Arabidopsis thaliana* (12)	H
		Solanaceae (3)	*Lycopersicon esculentum* (2)	H
			Lycopersicon peruvianum (1)	H
		Rosaceae (1)	*Malus* X *domestica* (1)	W
	Monocot (1)	Poaceae (1)	*Oryza sativa* (1)	H
	Gymnosperm (1)	Pinaceae (1)	*Pinus sylvestris* (1)	W
MIX II	Dicots (3)	Brassicaceae (2)	*Arabidopsis thaliana* (2)	H
		Stercullariaceae (2)	*Theobroma cacao* (1)	W
MIX III	Dicot 1	Brassicaceae (1)	*Arabidopsis thaliana* (1)	H

Synopsis regarding features of aligned species			
		N.	%
Grouped by taxonomic affiliation	Dicots	198	94,3
	Monocots	11	5,2
	Gymnosperm	1	0,5
Grouped by habit	Herbaceous	187	89,0
	Woody	23	10,9

what comprises 0.63% of the actually generated clusters. This approach allowed the identification of a large set of candidate sequences by using various representative genes per class, while some recent works employed few genes (Koczyk and Chelkowski, 2003). Using several previously described and sequenced *R*-genes as template was a useful and low-time consuming strategy in the search for *R*-genes candidates in plants. In this approach it was expected that some similar genes grouped at the same class should cause some level of redundancy (Meyers *et al.*, 1999). The strategy of generating a local database (called non-redundant) by adopting the cluster number as a primary key register was very effective in the solution of this problem. Additionally, this approach was useful in the identification of the respective *R*-gene class for each *Eucalyptus* cluster.

The number of *R*-genes here identified is quite high, especially considering that none of the 19 libraries were obtained under pathogen stress condition. By the other hand, when additional ESTs are generated especially under infection by pathogen, many of the identified clusters may be united in larger clusters of *R*-genes that may include more domains.

Evidences have shown that *R*-genes are quite abundant in higher plants, but the most functionally defined *R*-genes belong to the supergene LRR-NBS family. After completing the whole genome sequencing of the model plant *A. thaliana* a total of 85 TIR-NBS-LRR have been identified (The Arabidopsis Genome Initiative, 2000), less than the number of clusters (93) actually identified in *Eucalyptus*. Especially genes containing NBS-LRR domains were estimated to be in number of ca.166 for *A. thaliana* and ca.600 for rice (*O. sativa*) by Richly *et al.* (2002), but this later number is still not confirmed.

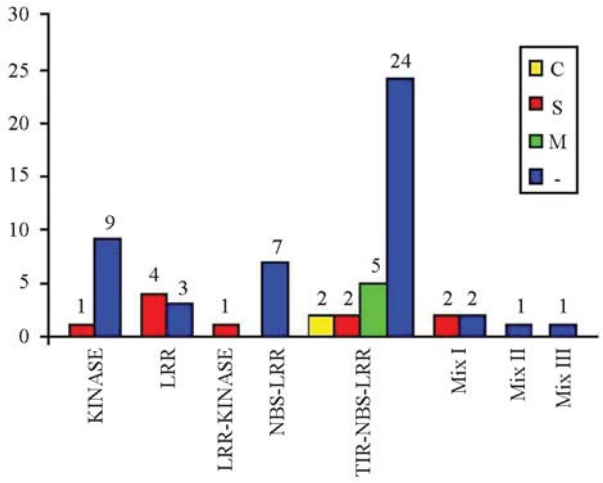

Figure 3 - Subcellular prediction for each class of analyzed *R* genes in *Eucalyptus* transcriptome, considering the predictions whit RC1 and RC2 of Target-P program. Legend: C = Chloroplast transit peptides; S = Secretory pathway; M = Mitochondrial targeting; '-' = No specific localization; LRR = Leucine-rich-repeat; NBS = Nucleotide-binding-site; TIR = Toll-interleucine-1 receptor.

A recent work reevaluated and reannotated all NBS-LRR encoding genes in *A. thaliana* genome database, revealing 149 genes of this class (including 94 TIR and 55 non-TIR sequences) in the genome of *A. thaliana* (Meyers *et al.*, 2003). In our evaluation of FOREST database we found 114 clusters (93 and 21, respectively) of this class. It is interesting to note that in the evaluation of Meyers *et al.* (2003) not only the presence of the TIR or of the CC motif was determinant for the grouping of both distinct classes. Also the NBS-LRR domains co-evolved and were determinant in the divergent evolution of the two groups, with the CC-bearing sequences forming four subgroups and the TIR-bearing sequences forming eight subgroups, regarding the size, composition and order or introns and exons.

Pan *et al.* (2000) compared tomato and *Arabidopsis* sequences of this class by systematically amplifying the tomato genome using a variety of primer pairs based on ubiquitous NBS motifs, generating 70 sequences, from which 10% were putative pseudogenes. The sequences were also used in mapping approaches, revealing a clustering *R*-gene homologues between tomato and potato (*Solanum tuberosum*, also from the Solanaceae family). Clustering of *R*-genes was also detected in *A. thaliana*, with most of the genes located in chromosomes 1 (49) and 5 (55), confirming the initial hypothesis that these genes are clustered in few chromosomes (The Arabidopsis Genome Initiative, 2000). This fact was also observed in other crops, as chickpea (*Cicer arietinum*; Benko-Iseppon *et al.*, 2003). In this last case, with some synteny and colinearity within this species and *Arabidopsis*. The clustering of *R*-genes in specific chromosomes and the existence of conserved domains have allowed the establishment of interesting strategies for identification, mapping and breeding directed to the incorporation of such genes from wild relatives. Considering the number of genes from this group in this last species, it is to expect that they are also clustered in *Eucalyptus*, what can also be valuable for the establishment of *Eucalyptus* breeding strategies in the future, especially considering the previous existence of mapping populations for this crop.

Overall annotation revealed that *Arabidopsis* also carries homologues of other *R*-gene classes, including 174 genes encoding LRR-kinases (*Xa21* group), but many of which are likely to play a role in development rather than defense (Jones, 2001). The present work revealed only eight clusters with significant homology to *Xa21* but this number can increase if only the kinase sequence is used as template, since the LRR may be quite variable between rice and *Eucalyptus*. 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 especially considering the possibility of pyramidization of such genes in important crops, increasing the potentiality of an effective specific *R-Avr* intection.

Another abundant family of *R*-genes in plants is the ser-thr-kinase with about 50 genes in *Arabidopsis* encoding protein kinases that are strongly homologous to tomato's *Pto* gene (Jones, 2001). In *Eucalyptus* we found almost the same number (49) of clusters also with high homology to the *Pto* sequence.

Regarding *R*-gene classes identified in *Eucalyptus*, an interesting phenomenon was observed in the present work: *R*-genes pertaining to different classes were able to align significantly to the same cluster on *Eucalyptus* database. This can be explained by the evidences that known *R*-genes combine a limited number of related functional domains (Ellis *et al.*, 1999, 2000a). Then, similar motifs would be present in different *R*-genes, and it is possible that a gene resembling to a determined class may search another belonging to a different class by local similarity at the site of the conserved motif. But in the practice, previous works do not speculate this possibility, once that the genes identified for specific *R*-genes are directly assigned to its own class as shown by evidences raised from works previously reported (Ronald, 1997; Jones, 2001; Romeis, 2001).

The MIX class one (MIX I) included 18 clusters resembling to genes which belong to both LRR and LRR-KINASE classes. These clusters were searched basically by using Cf (Jones *et al.*, 1998) and Xa21 (Song *et al.*, 1995) amino acid sequences as queries. In this case, the most plausible explanations would be the presence of the LRR domain, common to both classes, being responsible for the alignment and grouping of some clusters in both classes. By the other hand, LRRs are referred as fast evolving sequences and are in some cases quite imperfect, making manual annotation necessary. Often their amino-acid sequences are quite specific to their gene group (Dixon *et al.*, 1998; Ellis *et al.*, 1999). For example, using the LRR of *Xa21* against GenBank database will reveal significant alignments only to *Xa21* genes of rice (and some other Poaceae) and less significantly to *Arabidopsis*, but no sequence including other gene classes align significantly. A similar approach to the present work was used for the analysis of SUCEST (Sugarcane EST project, also running in Brazil) database (Morais, 2003) with no similar results. Song *et al.*, (1997) suggested that the structure of *Xa21* (here referred as class V) itself indicates an evolutionary link between different classes (I and IV) of plant disease resistance genes. May this be the case of this cluster that present a new link between two classes and can represent a new gene for Angiosperms?

Another surprising result was obtained by analyzing the unique cluster with both domains LRR and KINASE. It would be expected to find both domains in genes resembling *Xa21* but this cluster (EGEQRT3301C03.g) showed itself similar to both *Pto* (class I, described by Martin *et al.*, 1993) and *Cf* (Class IV, described by Jones *et al*, 1994)

genes. This double similarity occurred on different motifs. The *Pto* gene is known to encode a ser-thre-kinase protein (Martin *et al.*, 1993) and it was at this motif that the cluster showed similarity to this gene. On the other hand, *Cf* genes encode extracellular LRRs and it was at the LRR motif that the similarity was found. This cluster could be grouped in the LRR-KINASE class. So, why did it not align with *Xa21*, the single known gene with both LRR-KINASE domains? It should be answered by analyzing the KINASE-related clusters. Despite of the conservation of this region (Romeis, 2001), none of the *Pto* (KINASE) or *Xa21* (LRR and a receptor-KINASE) related clusters were mixed (aligned together) during the annotation process. This shows that the kinase segment is less-redundant than LRR at least during our *in silico* gene prediction, once that the kinase CD is present in both *Pto* and *Xa21* genes, they do not caused the mixture of their matching clusters on a mixed class.

The last case of mixture occurred to MIX class II including the motif TIR-NBS-LRR. Two of the three clusters pertaining to this mixed class (EGEQST6001H02.g and EGJECL1208G03.g) were searched at the FOREST database by the genes *RPP5* (TIR-NBS-LRR; Parker *et al.*, 1997) and *RPS5* (NSB-LRR; Noel *et al.*, 1999). The third cluster (EGEZRT3006B12.g) was obtained through search using *RPP5* and *RPS4* (both TIR-NBS-LRR; Gassmann *et al.*, 1999) and *I2* (NBS-LRR; Simmons *et al.*, 1998) queries. We initially supposed that the redundancy was due to the presence of NB-ARC (NBS) conserved motif. However, the first two clusters did not show any motif after *in silico* CD-search and, again, the region that apparently caused the mixture of the classes was the LRR motif, once that it was predicted in cluster EGEZRT3006B12.g.

In view of the results discussed above, could we speculate that *Eucalyptus* bears some new classes of *R*-genes? Before taking further conclusions and in order to solve the questions raised by the present work, we intend to evaluate these groups of clusters in regard to their domain and interdomain structure and organization, evaluating also the clusterization process, before taking further conclusions.

The conserved domains (CDs) identified during our investigation showed that most of the *Eucalyptus* predicted sequences possess the same motifs shared by disease *R*-genes. The CD with the higher level of sampling was LRR, which was present in all classes (except KINASE class I, represented by *Pto*) with a total of 442 occurrences. The other frequent domain shared by *R*-genes, the NB-ARC, was observed in 27 sequences, notably in TIR-NBS-LRR and NBS-LRR predicted clusters. This motif is commonly found in such sequences, and it is proposed that NB-ARC plays a role in activation of downstream effectors (van der Biezen and Jones, 1998) by their sequence similarity to mammalian CED-4 and APAF-1 proteins which are involved in apoptosis (Chinnaiyan *et al.*, 1997). In plants the TIR motif is found only associated to NBS regions of dicotyledones, being possibly absent in monocotyledones (Meyers *et al.*, 1999). In *Eucalyptus* (a eudicot genus of the Myrtaceae family) TIR domains were quite abundant, as expected, being found in 39 clusters (all from TIR-NBS-LRR-class).

Another very common motif present in two classes of disease *R*-genes is the kinase domain. This motif is shared by *Pto* (ser-thre-kinase) and *Xa21* (receptor-kinase) genes, members of the KINASE and LRR-KINASE classes, respectively. We found that all kinase domains found were associated to the classes KINASE, LRR-KINASE and MIX III. As commented here, despite of its conservation, this domain generally does not cause redundancy while searching in databases.

Transmembrane motifs were found only in 44 of all analyzed sequences. Of these clusters five TM were, unexpectedly, found in TIR-NBS-LRR-related sequences (a group of *R*-genes that acts at the intracellular level), while the remaining 19 were as expected LRR or LRR-KINASE-related sequences.

Information regarding the localization of disease resistance proteins in plant cells is still scarce (Martin, 1999). Spatial organization is usually variable among distinct gene classes and tissues affected, and there are no strong evidences in favor of conserved correspondence between R and Avr products spatial occurrence (Bonas and Lahaye, 2002). However, immunocytochemistry approaches allowed the subcellular localization of some Avr and R components (Boyes *et al.*, 1998). Here, we adopted an *in silico* approach which uses neural network-based methods to predict the topology (*i.e.* localization) of protein sequences of the selected clusters. In spite of the large number of predictions obtained, only 11 sequences were defined into RC1 (reliability class 1 $\geq$ 80%), and 53 for RC2 ($\geq$ 60%). Of these significant predictions, we observed that neural network was able to predict the localization of only a small number of proteins (29.62%) compared to the total sample of *Eucalyptus R*-genes. This percentage of representation is much lower than the 80% obtained for plant test sets carried out by Emanuelsson *et al.* (2000) with the same approach. It is important to note that these predictions are based on the N-terminal information available for sequences. Thus, this low number of predictions can be explained by the fact that the FOREST database was obtained from expressed sequence tags, an approach that usually do not include N-termini for many EST generated.

Our *Eucalyptus* transcriptome cDNA sequence analysis revealed that there are 210 clusters with significant alignment to major classes of plant disease *R*-genes. Differentially from the other genomic efforts, as *O. sativa* (Goff *et al.*, 2002) we used a redundant set of well described *R*-genes to screen for RGAs (Resistance Genes Analogs) on FOREST database. This proved to be a very sensitive approach, since best matches in NCBI present sometimes annotation mistakes and we also observed during the present work that

some of the best GenBank matches to *Eucalyptus R*-clusters presented no conclusive description of function. This was also the case also of the first annotation of *Arabidopsis* genome sequences, as pointed out by Meyers *et al.*, (2003). After reannotation of NBS-LRR sequences a total of 56 of the *A. thaliana R*-genes had to be corrected from earlier evaluations on GenBank (Meyers *et al.*, 2003). These results show how important procedures as annotation and detailed evaluation of generated sequences are. These evidences bring to reflections about the strategic design of many genome and transcriptome projects, considering that the data mining is not expensive (normally only fellowships are needed) but still receive few investments from financing agencies, diminishing the final impact of the results.

The comparison of our results regarding the number (and maybe the organization) of identified *Eucalyptus* clusters was mainly with *A. thaliana*, especially due to the lack of open databases for other plant species with EST projects. Many differences considering the here analyzed *R*-related sequences can be explained by using diverse arguments: (i) The larger genome of *Eucalyptus* (e.g. *E. grandis* with 640 Mbp; Myburg *et al.*, 2003) in contrast with the small and "compact" genome of *A. thaliana* (120 Mbp) (ii) The distant taxonomic position: both are dicots, but distantly related families (Brassicaceae and Myrtaceae) and finally (iii) the different levels of complexity: *Eucalyptus* is a wood perennial plant species and *Arabidopsis* is an annual herb. Herbaceous species are often regarded as faster evolving than woody species considering different morphological and genetic aspects (Bennet, 1972, Enrendorfer, 1982, Morawetz 1984, 1986, Bennet and Leitch, 1995, 2000).

Considering these evidences we observed that most of the information regarding *R*-genes available in databases refer to herbaceous (not woody) crop plants (few wild plants), maybe because most identified and sequenced *R*-genes were consequence of mapping approaches that are very time consuming in woody plants and difficult to realize in open pollinated species. The larger number of sequences from *A. thaliana* representing best alignments to *Eucalyptus* does not represent a higher similarity to this plant species, moreover it reflects the large number of sequences of this model plant deposited in GenBank. In our evaluation, only 23 woody species appeared as best matches for the clusters studied, including 22 species from different dicotyledonous families and one Gymnosperm species (*Pinus sylvestris*). This may justify some of the surprising results obtained in the present work and suggest that identification of *R*-genes in a larger number of taxonomic groups may be a very promissory approach to understand the natural evolution of these sequences when not affected by the influence of man. Regarding the actual knowledge of R-gene structure and diversity, some authors suggested that this gene class evolves faster than other genes (Ellis *et al*., 2000b) what should be evaluated in a larger number of taxonomic entities including wild species and also primitive taxa.

Concluding Remarks

Using bioinformatic tools it was possible to identify classify and verify the actually sequenced *R*-genes in *Eucalyptus* transcriptome. No previous sequences of this type could be found in protein or nucleotide databases for this crop. The identified sequences will be valuable resources for the development of markers for molecular breeding and identification of RGAs (resistance gene analogs) in *Eucalyptus* and other related species. The identified clusters constitute also excellent probes for physical mapping of genes in this species, giving support to genetic mapping programs and synteny studies. Considering the size of some clusters, they may also be used for fluorescent *in situ* hybridization (FISH) on *Eucalyptus* chromosomes, helping also in the comparison of different parental species and the respective hybrids.

The present work on *Eucalyptus*, based on FOREST database brought some light to the existing *R*-gene group in this important crop species and also regarding resistance response in higher plants, leading to the following conclusions:

• All five gene classes of *R*-genes with their respective conserved domains are present and expressed in *Eucalyptus*.

• Some new combinations of domains and motifs of *R*-genes may be present in *Eucalyptus* and could represent novel *R*-gene structures, what should be analyzed in detail.

• Despite the lack of libraries from tissues ellicitated by pathogens a high number of *R*-genes was found in different libraries of FOREST project. This may suggest, that the identified clusters are expressed constitutively but also leads to the supposition that a higher number of *R*-genes may be present in *Eucalyptus* under other experimental conditions.

Besides the detailed analysis of different groups of genes and domains we intend to evaluate the expression of the selected clusters in the different libraries of the project. Furthermore, some additional efforts may be necessary to complete some sequences of *R*-genes, especially considering that their size vary between 321 (in case of *Pto*) and 1802 amino-acids (in case of *Xa1* gene) and many identified sequences possibly present incomplete domains.

Further *in silico*, *in vitro* and *in vivo* evaluations of *Eucalyptus* genome may be a very promissory approach. Manipulation of the expression of these genes in economically important woody plant species aiming to improve disease resistance is necessary. Despite of the challenge that this mission may represent, some reports indicate that this strategy is feasible.

Acknowledgements

The present authors thank Ms. David Anderson de Lima Morais and Dr. Valdir Queiroz Balbino for interesting discussions and instructions about some of the pro-

grams and tools used in the present work. To Dr. Reginaldo de Carvalho and Claudete Maria Marques da Silva we thank for valuable technical support. We thank also CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the concession of a fellowship to the last author (Grant no. 478895/2003).

References

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

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

Ballvora A, Ercolano MR, Weiss J, Meksem K, Bormann CA, Oberhagemann P, Salamini F and Gebhardt C (2002) The *R1* gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. Plant J 30:361-371.

Barber PA, Smith IW and Keane PJ (2003) Foliar diseases of *Eucalyptus* spp. grown for ornamental cut foliage. Austral Plant Pathol 32:109-111.

Bendahmane A, Kohn BA, Dedi C and Baulcombe DC (1995) The coat protein of potato virus X is a strain-specific elicitor of Rx1-mediated virus resistance in potato. Plant J 8:933-941.

Bendahmane A, Querci M, Kanyuka K and Baulcombe DC (2000) *Agrobacterium* transient expression system as a tool for the isolation of disease resistance genes: Application to the Rx2 locus in potato. Plant J 21:73-81.

Benko-Iseppon AM, Winter P, Huettel B, Staginnus C, Muehlbauer FJ and Kahl G (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.

Bennett MD (1972) Nuclear DNA content and minimum generation time in herbaceous plants. Proc R Soc Lond Bot 181:109-135.

Bennet MD and Leitch IJ (1995) Nuclear DNA Amounts in Angiosperms. Ann Bot 76:113-176.

Bennet MD and Leitch, IJ (2000) Variation in nuclear DNA amount (C-value) in monocots and its significance. In: Wilson KL and Morrison DA (eds) Monocots: Systematics and Evolution. 1st edition. CSIRO Publishers, Sydney, pp 137-146.

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 21:177-88.

Bonas U and Lahaye T (2002) Plant disease resistance triggered by pathogen-derived molecules: Refined models of specific recognition. Curr Opin Microbiol 5:44-50.

Botella MA, Parker JE, Frost LN, Bittner-Eddy PD, Beynon JL, Daniels MJ, Holub EB and Jones JD (1998) Three genes of the *Arabidopsis* RPP1 complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. Plant Cell 10:1847-1860.

Boyes DC, Nam J and Dangl JL (1998) The *Arabidopsis thaliana RPM1* disease resistance gene product is a peripheral plasma membrane protein that is degraded coincident with the hypersensitive response. Proc Natl Acad Sci USA 95:15849-15854.

BRACELPA (2004), Associação Brasileira de Celulose e Papel. Brazil. Available from World Wide Web: http://www.bracelpa.org.br, release date 20/March/2004, cited 25/April/2004.

Brommonschenkel SH, 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 13:1130-38.

Bryan GT, Wu KS, Farrall L, Jia Y, Hershey HP, McAdams SA, Faulk KN, Donaldson GK, Tarchini R and Valent B (2000) A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. Plant Cell 12:2033-2046.

Buschges R, Hollricher K, Panstruga R, Simons G, Wolter M, Frijters A, van Daelen R, van der Lee T, Diergaarde P, Groenendijk J, Topsch S, Vos P, Salamini F and Schulze-Lefert P (1997) The barley *Mlo* gene: A novel control element of plant pathogen resistance. Cell 88:695-705.

Chinnaiyan AM, Chaudhary D, O'Rourke K, Koonin E and Dixit M (1997) Role of CED-4 in the activation of CED-3. Nature 388:728-729.

Collins N, Drake J, Ayliffe M, Sun Q, Ellis J, Hulbert S and Pryor T (1999) Molecular characterization of the maize Rp1-D rust resistance haplotype and its mutants. Plant Cell 11:1365-1376.

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

Dixon MS, Hatzixanthis K, Jones DA, Harisson K and Jones JGD (1998) The tomato *Cf-5* disease resistance gene and six homologs show pronounced allelic variation in leucine-rich repeat copy number. Plant Cell 10:1915-1925.

Dodds P, Lawrence G and Ellis J (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 13:163-78.

Ehrendorfer F (1982) Speciation patterns in woody angiosperms of tropical origin. In: Barigozzi C (ed) Mechanisms of Speciation. Alan R. Liss. Inc., New York, pp 479-509.

Ellis J and Jones D (1998) 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 N (1999) Identification of regions in alleles of the flax rust resistance gene *L* that determines differences in gene-for-gene specificity. Plant Cell 11:495-506.

Ellis J, Dodds P and Pryor T (2000a) The generation of plant disease resistance genes specificities. Trends Plant Sci 5:373-379.

Ellis J, Dodds P and Pryor T (2000b) Structure, function and evolution of plant disease resistance genes. Curr Opin Plant Biol 3:278-284.

Emanuelsson O, Nielsen H, Brunak B and von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005-1016.

Ernst K, Kumar A, Kriseleit D, Kloos DU, Phillips MS and Ganal MW (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.

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.

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.

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 and Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92-100.

Grant MR, Godiard L, Straube E, Ashfield T, Lewald J, Sattler A, Innes RW and Dangl JL (1995) Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. Science 269:843-846.

Halterman D, Zhou F, Wei F, Wise RP and Schulze-Lefert P (2001) The MLA6 coiled-coil, NBS-LRR protein confers AvrMla6-dependent resistance specificity to *Blumeria graminis* f. sp. *hordei* in barley and wheat. Plant J 3:335-348.

James WC, Teng PS and Nutter FW (1990) Estimated losses of crops from plant pathogens. In: Pimentel D (ed) CRC Handbook of Pest Management, CRC Press, Boca-Raton, pp 15-50.

Johal GS and Briggs SP (1992) Reductase activity encoded by the *HM1* disease resistance gene in maize. Science 158:958-987.

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

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

Jones JB, Stall RE and Bouzar H (1998) Diversity among xanthomonads pathogenic on pepper and tomato. Ann Rev Phytopathol 36:41-58.

Joosten MH, Cozijnsen TJ and De Wit PJ (1994) Host resistance to a fungal tomato pathogen lost by a single base-pair change in an avirulence gene. Nature 367:384-386.

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.

Koonin EV and Aravind L (2000) The NACHT family - A new group of predicted NTPases implicated in apoptosis and MHC transcription activation. Trends Biochem Sci 25:223-224.

Lawrence GJ, Finnegan EJ, Ayliffe MA and Ellis JG (1995) The *L6* gene for flax rust resistance is related to *Arabidopsis* bacterial resistance gene *RPP2* and tobacco viral gene *N*. Plant Cell 7:1195-1206.

Mafia RG and Alfenas AC (2003) Diferenciação sintomatológica de manchas foliares em *Eucalyptus* spp. causadas por patógenos fúngicos e bacterianos. Fitopatol Bras 28:688-688.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY and Bryant SH (2002) CDD: A database of conserved domain alignments with links to domain three-dimensional structure. Nucl Acids Res 30:281-283.

Martin GB (1999) Functional analysis of plant disease resistance genes and their downstream effectors. Curr Opin Plant Biol 2:273-279.

Martin GB, de Vicente MC and Tanksley SD (1993) Hight resolution linkage analysis and physical characterization of the Pto bacterial locus in tomato. Mol Plant-Microbe Interact 6:26-34.

Martin GB, Bogdanove AJ and Sessa G (2003) Understanding the functions of plant disease resistance proteins. Annu Rev Plant Physiol Plant Mol Biol 54:23-61.

McDowell JM, Dhandaydham M, Long TA, Aarts MG, Goff S, Holub EB and Dangl JL (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.

McNabb K (2002) Clonal propagation of *Eucalyptus* in Brazilian nurseries. In: Dumroese RK, Riley LE and Landis TD (eds) National Proceedings: Forest and Conservation Nursery Associations. USDA Forest Service, Rocky Mountain Research Station, Ogden, pp 165-168.

Meyers BC, Chin DB, Shen KA, Sivaramakrishnan S, Lavelle DO, Zhang Z and Michelmore RW (1998) The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. Plant Cell 10:1817-32.

Meyers BC, Diekcman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW and Young ND (1999) Plant disease resistance genes encodes members of an ancient and diverse protein family within the nucleotide-biding superfamily. Plant J 20:317-332.

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

Milligan SB, Bodeau J, Yaghoobi J, Kaloshian I, Zabel P and Williamson VM (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 10:1307-19.

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

Morais DAL (2003) Análise bioinformática de genes de resistência a patógenos no genoma da cana-de-açúcar. Master Dissertation, Universidade Federal de Pernambuco, Recife.

Morawetz W (1984) How stable are genomes of tropical woody plants? Heterozygosity in C-banded Karyotypes of *Porcelia* as compared with *Annona* (Annonaceae) and Drymys (Winteraceae). Pl Syst Evol 145:29-39.

Morawetz W (1986) Remarks on karyological differentiation patterns in tropical woody plants. Pl Syst Evol 152:49-100.

Myburg AA, Griffin AR, Sederoff RR and Whetten RW (2003) Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F1 hybrid based on a double pseudo-backcross mapping approach. Theor Appl Genet 107:1028-1042.

Noel L, Moores TL, van Der Biezen EA, Parniske M, Daniels MJ, Parker JE and Jones JD (1999) Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. Plant Cell 11:2099-2112.

Ori N, Eshed Y, Paran I, Presting G, Aviv D, Tanksley S, Zamir D and Fluhr R (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.

Pan Q, Liu YS, Budai-Hadrian O, Sela M, Carmel-Goren L, Zamir D and Fluhr R (2000) Comparative genetics of nucleotide binding site leucine-rich repeat resistance gene homologues in the genomes of two dycotyledons: Tomato and *Arabidopsis*. Genetics 155:309-322.

Parker JE, Coleman MJ, Dean C and Jones JGD (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, Jarosch B, Schaffrath U, Collins NC, Panstruga R and Schulze-Lefert P (2002) The barley MLO modulator of defense and cell death is responsive to biotic and abiotic stress stimuli. Plant Physiol 129:1076-1085.

Richly E, Kurth J and Leister D (2002) Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution. Mol Biol Evol 19:76-84.

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

Romeis T (2001) Protein kinases in the plant defense response. Curr Opin Plant Biol 4:407-414.

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

Ronald PC (1997) The molecular basis of disease resistance in rice. Plant Mol Biol 35:179-186.

Simmons G, Groenendijk J, Wijbrandi J, Reijans M, Groenen J, Diergaarde van der Lee T, Bleeker M, Onstenk J, De Both M, Haring M, Mes J, Cornelissen B, Zabeau M and Vos P (1998) Dissection of the fusarium *I2* gene cluster in tomato reveals six homologs and one active gene copy. Plant Cell 10:1055-1068.

Song WY, Pi LY, Wang GL, Gardner J, Holsten T and Ronald PC (1997) Evolution of the rice *Xa21* disease resistance genes family. Plant Cell 9:1279-1287.

Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, Gardner J, Wang B, Zhai WX, Zhu LH, Fauquet C and Ronald PC (1995) A receptor kinase-like protein encoded by the rice disease resistance gene *Xa21*. Science 270:1804-1806.

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

Thomas CM, Jones DA, Parniske M, Harrison K, Balint-Kurti PJ, Hatzixanthis K and Jones JD (1997) Characterization of the tomato *Cf-4* gene for resistance to *Cladosporium fulvum* identifies sequences that determine recognitional specificity in *Cf-4* and *Cf-9*. Plant Cell 9:2209-2224.

van der Biezen EA and Jones JGD (1998) The NB-ARC domais: A novel signaling motif shared by plant resistance gene products and regulators of cell death in animals. Curr Biol 8:R226-R227.

van der Biezen EA, Freddie CT, Kahn K, Parker JE and Jones JD (2002) *Arabidopsis RPP4* is a member of the *RPP5* multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signaling components. Plant J 29:439-51.

van der Vossen EA, van der Voort JN, Kanyuka K, Bendahmane A, Sandbrink H, Baulcombe DC, Bakker J, Stiekema WJ and Klein-Lankhorst RM (2000) Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: A virus and a nematode. Plant J 23:567-576.

Wang ZX, Yano M, Yamanouchi U, Iwamoto M, Monna L, Hayasaka H, Katayose Y and Sasaki T (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.

Whithan 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 93:8776-81.

Xiao S, Ellwood S, Calis O, Patrick E, Li T, Coleman M and Turner JG (2001) Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by RPW8. Science 291:118-20.

Yoshimura S, Yamanouchi U, Katayose Y, Toki S, Wang ZX, Kono I, Kurata N, Iwata N and Sasaki T (1998) Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. Proc Natl Acad Sci USA 95:1663-1668.

Associate Editor: Marcio de Castro Silva Filho