RESEARCH ARTICLE

Positive selection in the leucine-rich repeat domain of *Gro1* genes in *Solanum* species

VALENTINO RUGGIERI, ANGELINA NUNZIATA and AMALIA BARONE*

Department of Agricultural Sciences, University of Naples Federico II, Via Università 100, 80055 Portici (NA), Italy

Abstract

In pathogen resistant plants, solvent-exposed residues in the leucine-rich repeat (LRR) proteins are thought to mediate resistance by recognizing plant pathogen elicitors. In potato, the gene *Gro1-4* confers resistance to *Globodera rostochiensis*. The investigation of variablity in different copies of this gene represents a good model for the verification of positive selection mechanisms. Two datasets of *Gro1* LRR sequences were constructed, one derived from the *Gro1-4* gene, belonging to different cultivated and wild *Solanum* species, and the other belonging to paralogues of a resistant genotype. Analysis of nonsynonymous to synonymous substitution rates (K_a/K_s) highlighted 14 and six amino acids with $K_a/K_s > 1$ in orthologue and paralogue datasets, respectively. Selection analysis revealed that the leucine-rich regions accumulate variability in a very specific way, and we found that some combinations of amino acids in these sites might be involved in pathogen recognition. The results confirm previous studies on positive selection in the LRR domain of R protein in *Arabidopsis* and other model plants and extend these to wild *Solanum* species. Moreover, positively selected sites in the *Gro1* LRR domain show that coevolution mainly occurred in two regions on the internal surface of the three-dimensional horseshoe structure of the domain, albeit with different evolutionary forces between paralogues and orthologues.

[Ruggieri V., Nunziata A. and Barone A. 2014 Positive selection in the leucine-rich repeat domain of *Gro1* genes in *Solanum* species. *J. Genet.* **93**, 755–765]

Introduction

Plant pathogen resistance genes have been divided into five main classes based on the presence or absence of typical domains (Hammond-Kosack and Parker 2003). Four such classes contain a LRR domain. For instance, the CC-NBS-LRR (CNL) class comprises resistance genes encoding proteins with a coiled–coil domain (CC), a nucleotide-binding site (NBS) and a leucine-rich repeat (LRR); the TIR-NBS-LRR (TNL) class have a toll-interleukin receptor-like domain (TIR), a NBS and a LRR. Moreover, two classes (RLK and RLP), have an external LRR domain (eLRR) that is connected via a transmembrane domain to a variable cytoplasmic C-terminal region. In R genes, LRR domains are directly involved in pathogen–host recognition (Fluhr 2001; Martin *et al*. 2003; De Young and Innes 2006; Jaillais *et al*. 2011). The crystal structures of the LRR domains from various nonplant organisms have been compared and are found to closely resemble each other (Kobe and Deisenhofer 1995; Zhang *et al*. 2000; Chai *et al*. 2011; Hothorn *et al*. 2011). In these domains, individual repeats of parallel β-strands and α-helix units are arranged consecutively to form a curved shape resembling a horseshoe (Kobe and Deisenhofer 1993). In this scaffold, aliphatic residues provide the proper packing of the hydrophobic core, whereas solvent-exposed residues interact with ligands. LRR from different protein subfamilies retain a similar solenoid fold and nonglobular horseshoe shape but differ in the three-dimensional (3D) structures of individual repeats (Kajava *et al*. 2008). It has been suggested (Thomas *et al*. 1997; Ellis and Jones 1998, Ellis *et al*. 2000; Tor *et al*. 2009) that in resistant plants, solvent-exposed residues in the β-strand portion of the LRR protein mediate pathogen resistance by recognizing plant pathogen elicitors. Results in plant and nonplant organisms (Michelmore and Meyers 1998; Ellis and Jones 1998; Ellis *et al*. 2000; Jann *et al*. 2008) suggest that these domains are under diversifying selection, thus corroborating the hypothesis of a role of LRR protein in pathogen recognition and resistance.

*For correspondence. E-mail: ambarone@unina.it.

Valentino Ruggieri and Angelina Nunziata contributed equally to this work. VR carried out the molecular genetic studies, sequence elaborations and drafted the manuscript. AN carried out the sequence analysis, alignment and drafted the manuscript. AB conceived the study, participated in its design and coordination, and helped draft the manuscript. All authors have read and approved the final manuscript.

Keywords. plant resistance genes; SNP; nonsynonymous substitution; synonymous substitution; orthologues; paralogues.

Analysis of nonsynonymous nucleotide substitutions over the number of synonymous substitutions (K_a/K_s) is the method most frequently used to detect positive or purifying selection (Kreitman 2000; MacCallum and Hill 2006; Roth and Liberles 2006). Active sites or structurally important sites may represent conserved sites. Mutations in such sites may reduce fitness and are therefore more likely to be removed from the population (purified sites) (Graur and Li 2000). Highly variable sites, instead, may represent those under positive Darwinian selection. Mutations in these sites may be interpreted as a molecular adaptation conferring an evolutionary advantage to the organism and have a higher probability of becoming fixed in the population. Sites responsible for host–pathogen recognition could be under positive selection if site mutation confers resistance to a new pathogen strain or a new pathogen. The *Globodera rostochiensis* resistance gene (*Gro1-4*) was mapped on chromosome 7 of potato using the restriction fragment length polymorphism (RFLP) marker CP51c on an *S. tuberosum* × *S. spegazzinii* hybrid (Barone *et al*. 1990). The gene *Gro1-4* was later identified in the diploid potato genotype P6/210 (GenBank AY196151) and demonstrated to confer resistance to *G. rostochiensis* pathotype Ro1 (Paal *et al*. 2004). The gene *Gro1-4* is part of a cluster of at least 11 genes or pseudogenes (*Gro1-1*, *Gro1-2*, *Gro1-3*, *Gro1-4*, *Gro1-5*, *Gro1-6*, *Gro1-8*, *Gro1-10*, *Gro1-11*, *Gro1-12* and *Gro1-14*) sharing high sequence similarity.

The genes of the *Gro1* cluster share the same intron/exon organization. The first exon contains a 500-bp long TIR domain, the second a 1000-bp long NBS domain, the third a 1300-bp long LRR domain of 11 modules. The fourth exon displays no similarity with the known domain motifs. Several studies have demonstrated that TIR domains are responsible for cytoplasm signal transduction, NBS domains for phosphorylation cascades and LRR domains for protein–protein interactions (Cai *et al*. 1997; Jia *et al*. 2000; Dangl and Jones 2001). Within this gene cluster, *Gro1-4* is the only gene conferring resistance to *G. rostochiensis*. However, reverse transcription-polymerase chain reaction (RT-PCR) has shown that the remaining genes are expressed in several tissues such as the roots, flowers, stems, tubers, stolons and leaves (Paal *et al*. 2004), suggesting that they might exert some biological function other than nematode resistance such as that reported for the gene *Mi* from tomato, which confers resistance both to the nematode *Meloidogyne incognita* (Milligan *et al*. 1998) and the aphid *Macrosiphum euphorbiae* (Rossi *et al*. 1998). Among the domains shared by both the genes *Mi* and *Gro1-4*, the LRR plays a key role in mediating direct or indirect interactions with the pathogen elicitors (Dangl and Jones 2001; De Young and Innes 2006; Matsushima and Miyashita 2012) and has been widely proven to be under positive selection in the *NBS-LRR R* genes (Ellis and Jones 1998; Meyers *et al*. 1998; Bergelson *et al*. 2001; Moore and Purugganan 2005).

Evolution of plant disease resistance (*R*) genes that encode an LRR region has been extensively studied (McHale *et al*. 2006; Afzal *et al*. 2008; Wulff *et al*. 2009a; Sanseverino *et al*. 2010). The generation of *R* genes and emergence of new resistance alleles have been ascribed to a number of causes, including primarily gene duplication, genetic recombination, diversifying selection, sequence divergence in the intergenic region, composition of transposable elements, gene conversion and unequal crossover (Mondragon-Palomino *et al*. 2002; Zhou *et al*. 2007). The *Gro-1* cluster of resistance genes is present in numerous related and syntenic wild potato species. Among the latter, the LRR domain is characterized by a high level of similarity, combined with an average rate of synonymous substitutions (K_s) of 0.035 (Nunziata *et al*. 2007). Such data document a relatively long evolutionary history during which the main structure of the domain has been conserved such that single substitutions have occurred in an unchanged domain context. The material presented here was, therefore, particularly suited to our study. Besides, these preliminary data support the hypothesis that the main general function of elicitor recognition of the domain has been conserved during its evolution in *Solanum* species.

The aim of this work is to evaluate the coevolving sites in the 'gene-for-gene' pathogen–host interaction in a set of paralogues and orthologues genes of *Gro1* locus. This was achieved by computing the ratio of nonsynonymous to synonymous substitutions and by mapping the position of all positively selected amino acid residues onto the LRR domain structure. The selected sites might help to predict the regions of LRR domain that are most likely to be involved in plant defense against *G. rostochiensis*.

Materials and methods

Plant material

Plant material used in this study included one accession from each of the 16 wild *Solanum* species listed in table 1. Accessions were provided by the IR-1 Potato Introduction Project (Sturgeon Bay, USA) as true seeds. The study also included one variety of cultivated *S. tuberosum* cv. 'Spunta' and the diploid hybrid P40 (*S. spegazzinii* × *S. tuberosum*). The latter was one of the genotypes originally used to map the locus *Gro1* and to clone and sequence the *Gro1-4* gene (Barone *et al*. 1990; Paal *et al*. 2004).

LRR resequencing

DNA was extracted from 100 mg of leaves using the Qiagen DNeasy® Plant Mini Kit (Valencia, USA) and then eluted from the column using 100 μL of distilled water. The LRR fragments from the 16 listed genotypes were amplified and sequenced using the following primers designed specifically on the *Gro1-4* sequence (GenBank AY196151): GRO1–4LRR1F (5′ – 3′ gcc taa cct tga aag act gg) and GRO1–4LRR2R (5′ – 3′ tga cga cag ttt ctg aat gta g). These primers target the 927-bp long fragment extending from 11,874 to 12,800 bp region of the *Gro1-4* sequence. PCR were performed according to standard protocols using 50–100 ng of DNA per reaction.

Table 1. Name, code, plant introduction number, ploidy level and geographical origin of genotypes used as plant material.

Species	Code	PI	Ploidy level	Origin
S. acaule	ACL 1	210029	2*n*=4*x*	Bolivia
S. boliviense	BLV 1	310974	2*n*=2*x*	Bolivia
S. bulbocastanum	BLB 3	243510	2*n*=2*x*	Mexico
S. canasense	CAN 1	365321	2*n*=2*x*	Peru
S. chacoense	CHC 1	133124	2*n*=2*x*	Uruguay
S. commersonii	CMM 1	243503	2*n*=2*x*	Argentina
S. demissum	DMS 1	205625	2*n*=6*x*	Mexico
S. eutuberosum	ETB 3	558054	2*n*=2*x*	Chile
S. fendleri	FEN 2	458417	2*n*=4*x*	USA
S. hougasii	HOU 1	161726	2*n*=6*x*	Mexico
S. multidissectum	MLT 1	8MLT-MI	2*n*=2*x*	Peru
S. phureja	IVP 35		2*n*=2*x*	Colombia
S. stoloniferum	STO 1	275248	2*n*=4*x*	Mexico
S. tarijense	TAR 1	265577	2*n*=2*x*	Bolivia
S. tuberosum ssp. andigena	TBR1	205624	2*n*=4*x*	Bolivia
S. vernei	VRN 1	230468	2*n*=2*x*	Argentina
S. tuberosum ssp. tuberosum	Spunta		2*n*=4*x*	
S. spegazzini x S. tuberosum	P40		2*n*=2*x*	

PI, plant introduction number.

Sequence alignment and analysis

The sequences obtained as previously described were then compared with the reference sequence *Gro1-4* (GenBank AY196151) from 11,874 to 12,800 bp region by the Seq Scape software (ABI Prism, Foster City, USA). Selection affecting specific sites were detected using the Selecton 2.4 Server (http://selecton.tau.ac.il/), which measures the K_a/K_s rate on each amino acid residue (Nielsen and Yang 1998; Yang and Bielawski 2000; Stern *et al*. 2007). The server program performs a codon-based alignment of the sequences, calculates the K_a/K_s ratio at each site, translates these ratios into colour-coded selection scores and projects these onto the primary, secondary and 3D structure sequences of the protein. In this study, selection scores were computed under both M8 and MEC models. In the M8 model, a proportion p_0 of the sites were estimated to have K_a/K_s values in the interval (0, 1), whereas the remaining sites were constricted into a category called ω_S (values in the interval (1, ω_S)). Both p_0 and ω_S were estimated using the maximum likelihood method. Statistical significance was then evaluated by comparing the M8 likelihood scores against the M8a null model (Swanson *et al*. 2003), which assumes no positive selection (H_O). By contrast, the mechanistic empirical combination model (MEC) takes into account the differences between amino acid replacement probabilities, expanding a 20 × 20 amino acid replacement rate matrix (such as the commonly used JTT matrix) into a 61 × 61 sense-codon rate matrix. Significance for the test was obtained by comparing the Akaike information content (AIC) scores between the MEC and the M8a models. Moreover, single amino-acid sites were considered to be significantly under positive selection in both tests if they displayed a posterior probability >95%. The I-Tasser server (http://zhanglab.ccmb.med.umich.edu/I-TASSER) was used to predict the 3D structure of the domain and to obtain the best PDB template for mapping positive values (see 'Rank of templates representing the top 10 threading templates used by I-TASSER' in electronic supplementary material at http://www.ias.ac.in/jgenet/), and FirstGlance in Jmol (http://bioinformatics.org/firstglance/fgij//index.htm) was used to visualize the 3D structure. Nucleotide and amino acid analysis were performed with the CLC Main Workbench (Aarhus, Denmark). A phylogenetic tree was constructed using TreeView (http://tree.bio.ed.ac.uk/software/figtree) and multialignment representation by using WebLogo application (http://weblogo.berkeley.edu/logo.cgi).

Results and discussion

Gro1 LRR sequence collection

As previously indicated by Paal *et al*. (2004), the Gro1-4 protein (1136 aa) included a TIR homology domain (129 aa), NB domain (255 aa), LRR domain (317 aa) and a IV exon of unknown function. They found 11 repeat leucine modules in the LRR domain, as predicted by using Profilescan. In addition, our analysis highlighted an imperfect 12th module. This 12th module revealed an xIxxIxAxxCxxxIxxL motif (table 2), comparable to the consensus motif of the other LRR modules xLxxLxLxx(C/N/T)x(x)LxxxP. Most of the Gro1-4 LRR repeating units were 24 residues in length, but they ranged from 21 to 26 residues. All the LRR domains formed a single continuous structure and adopted an arc or horseshoe shape (Enkhbayar *et al*. 2004). On the inner concave face, a stack of parallel β-strands (11.7% of residues) were present, and on the outer convex face, a variety of secondary

Table 2. Multialignment of 12 repeat leucine modules in the LRR domain of the Gro1-4 reference sequence. For each module, we report the number of amino acids (AA), start and end positions of each module, consensus sequence and logo representation of multialignment.

LRR modules	No AA	Start-sequence-end																									
1	24	2-	N	L	E	R	L	V	L	E	E	C	T	S	L	V	E	–	I	N	F	S	–	I	E	N	L	G	-25
2	23	26-	K	L	V	L	L	N	L	K	N	C	R	N	L	K	T	–	L	P	K	–	–	R	I	R	L	E	-48
3	24	49-	K	L	E	I	L	V	L	T	G	C	S	K	L	R	T	–	F	P	E	I	–	E	E	K	M	N	-72
4	23	73-	C	L	A	E	L	Y	L	–	G	A	T	S	L	S	E	–	L	P	A	S	–	V	E	N	L	S	-95
5	24	96-	G	V	G	V	I	N	L	S	Y	C	K	H	L	E	S	–	L	P	S	S	–	I	F	R	L	K	-119
6	23	120-	C	L	K	T	L	D	V	S	G	C	S	K	L	K	N	–	L	P	D	–	–	D	L	G	L	L	-142
7	23	144-	G	L	E	E	L	H	C	–	T	H	T	A	I	Q	T	–	I	P	S	S	–	M	S	L	L	K	-166
8	26	167-	N	L	K	H	L	S	L	S	G	C	N	A	L	S	S	Q	V	S	S	S	–	S	H	G	Q	K	-192
9	25	205-	S	L	I	M	L	D	L	S	D	C	N	I	S	D	G	G	I	L	N	N	–	L	G	F	L	S	-229
10	24	230-	S	L	E	I	L	I	L	N	G	N	N	–	F	S	N	–	I	P	A	A	S	I	S	R	F	T	-253
11	21	254-	R	L	K	R	L	K	L	H	G	C	G	R	L	E	S	–	L	P	E	L		P	P	–	–	–	-274
12	21	275-	S	I	K	G	I	F	A	N	E	C	T	S	L	M	S		I	D	Q	L	T	K	Y	P	M	L	-295
Consensus			**X**	**L**	**X**	**X**	**L**	**X**	**L**	**X**	**X**	**C**	**X**	**X**	**L**	**X**	**X**	**x**	**L**	P	**X**	**X**	**x**	**L**	**X**	**X**	**L**	**X**	
Logo																													

structures such as α-helix (17%) or β-turns connected by loops were present, as shown in figure 1. To explore *Gro1* LRR variability, we chose to collect and examine a large number of homologue sequences of the *Gro1* cluster belonging to different species (orthologues 'ORT') or to the same species (paralogues 'PAR'). Sequences from the ORT were genetically isolated from each other by speciation. Sequences from PAR were derived instead from the P6/210 hybrid (Paal *et al*. 2004). Sequences in the ORT were obtained using *GRO1-4* LRR primers on analysed genotypes. As expected, each of the 16 analysed genotypes produced one 927 bp amplified fragment and no other band was detected. Each fragment was sequenced and a consensus sequence corresponding to the 11,874–12,800 bp fragment of the *Gro1-4* gene was established. Out of the 16 consensus sequences thus determined, nine corresponded each to a single haplotype, whereas the remaining seven displayed double peaks corresponding to double base callings. These seven sequences belonged to the diploid species *S. bulbocastanum*, *S. chacoense*, *S. commersonii*, *S. tarijense*, *S. vernei*, the tetraploid species *S. tuberosum* ssp. *andigena*, and the hexaploid species *S. demissum*. In these seven sequences, 73 double bases were scored, varying from a minimum of two in *S. commersonii* to a maximum of 20 in *S. demissum*. No more than one double peak per codon was ever observed, neither were triple nor quadruple peaks observed. To perform

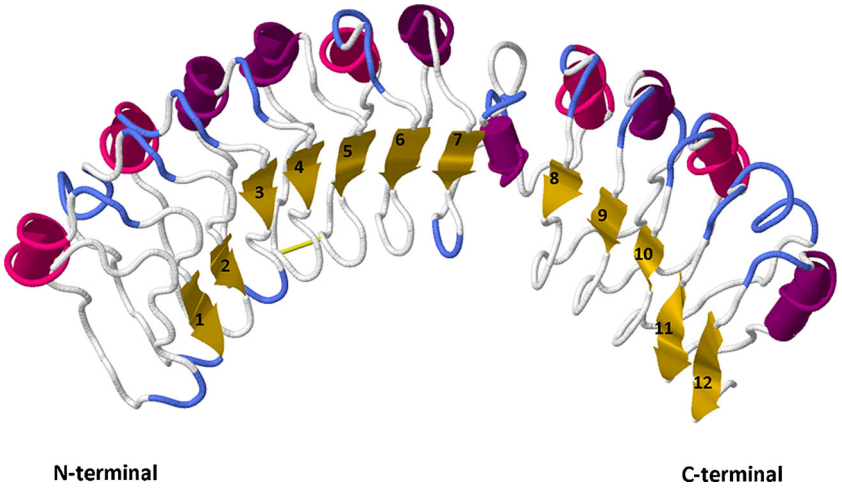

Figure 1. *Gro1-4* LRR 3D prediction. The 12 modules are highlighted and numbered. β-sheets are shown in yellow and α-helices as tubes.

the codon elaborations specified below, the seven sequences with double peaks were exploited in two haplotypes, each carrying one of the two bases pointed out by the double peaks observed. Since no information on the real *cis*/*trans* position of detected SNPs was available, SNPs on each of these haplotype sequences were sorted arbitrarily. This arbitrary haplotype construction did not affect downstream K_a/K_s elaborations as these were performed on SNPs independently, codon by codon. In this regard, tests were made by randomly inverting SNP attribution to the 14 arbitrary haplotypes in ORT built to resolve sequences carrying double peaks. These tests confirmed that arbitrary haplotype construction did not affect downstream elaborations (data not shown).

Out of 12 paralogues of the *Gro1* cluster detected by Paal *et al*. (2004), seven (*Gro1-1*, *Gro1-2*, *Gro1-3*, *Gro1-5*, *Gro1-6*, *Gro1-8* and *Gro1-10*) were collected in PAR by selecting them for absence of stop codons in the LRR domain. These sequences were aligned and used to build a phylogenetic tree. As shown in figure 2, sequences were grouped into two main clusters: the first included *Solanum* genotypes, the second consisted of paralogue genes of the *Gro1* cluster. However, the LRR domain is characterized by a high level of similarity (Nunziata *et al*. 2007) between the different species of *Solanum* and within the cluster *Gro1*. Indeed, the average similarity between the sequences considered is 94% (table 1 in electronic supplementary material). This indicates that during speciation, the structure of the domain was preserved and that subsequent duplication events led to the formation of a highly conserved gene cluster. Clustered genes probably retained the structure but not the same function. For example, some of these paralogues in the cluster may not be functioning or may exert some biological functions other than nematode resistance as shown for the gene *Mi* (Milligan *et al*. 1998; Rossi *et al*. 1998).

K_a/K_s detection

To examine the selective pressures acting on the repertoire of the LRR domain of the *Gro1* genes, Selecton 2.4 software (Doron-Faigenboim and Pupko 2006; Delport *et al*. 2009) was used. This software enabled us to investigate the molecular mechanisms involved in the ligand-protein recognition in the reactions of resistance/susceptibility to *G. rostochiensis* in the *Solanum* species. Selecton implements several codon models: in this case, models M8 and MEC were used to detect positive selection of ORT (orthologues from *Solanum* wild species) and PAR (paralogues from the P40 genotype).

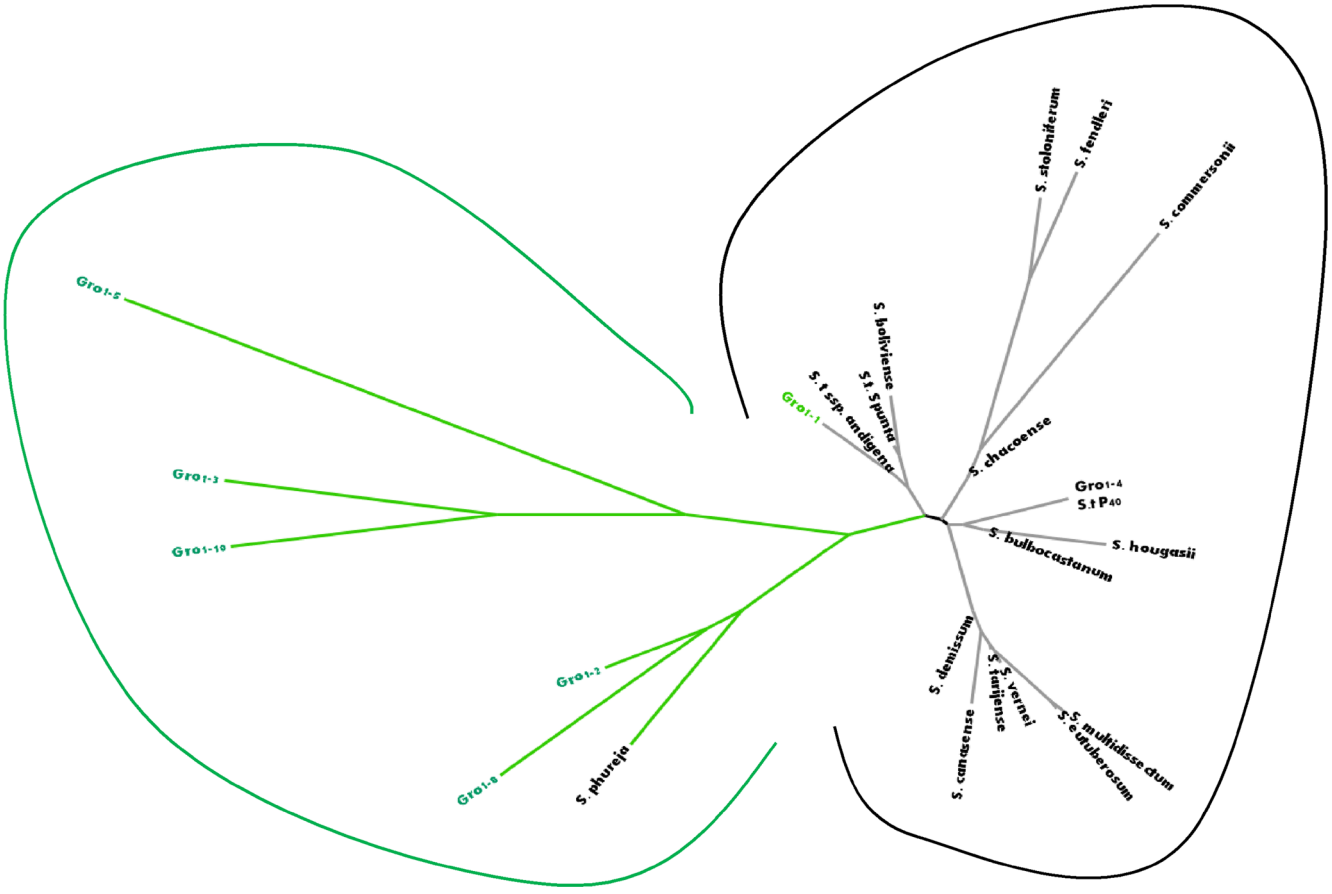

Figure 2. Phylogenetic tree based on *Gro1* LRR sequence similarity in *Solanum* species. Paralogues of *Gro1-4* are in green and orthologues in black.

All sequences were aligned by using the sequence of the *Gro1-4* gene as reference, translated into amino acid sequences and checked for absence of stop codons. We used the M8 model to detect sites under positive selection and the MEC model to gain insights into the different selection forces and their intensities.

In our tests on ORT and PAR, the likelihood resulting from application of the M8 model was always greater than that of the null M8a model, and the likelihood ratio test showed a significance level of 0.001, demonstrating that the LRR domain is under positive selection. As there is no null model nested within the MEC model, to perform statistical analysis AIC scores were compared between models MEC and M8a. AIC scores computed under the MEC and M8a models on our data confirmed the occurrence of positive selection by both sets.

M8 test: Figure 3 shows overall variation of K_a/K_s rates along amino acid sequences in ORT and PAR computed with the M8 codon model (for all tabular values see table 2 in electronic supplementary material) and in particular its distribution on each LRR module of the domain. Most of the variability was found in the C-terminal portion of the domain. In addition, other studies that pointed out the attention on variability of LRR domain in resistance genes highlighted similar results. Studies on L resistance proteins from flax (*Linum usitatissimum*) found positively selected sites mainly in the N-terminal and C-terminal portions of the LRR domain (Ravensdale *et al*. 2011). In addition, studies on patterns of LRR nucleotide variation in 12 different resistance genes showed that positively selected sites were mainly located in the C-terminal half of the LRR repeats in most (78%) *NBS-LRR* genes considered (Jiang *et al*. 2007). The evidences of this high rate of variation in the C-terminal portion of the LRR suggest that this part may be directly involved in the interactions with pathogen elicitors and may confer pathogen recognition specificity. In particular, in our case, in the ORT we identified 35 sites with $K_a/K_s > 1$ but only 14 sites showed statistically significant values (with $K_a/K_s > 1$, where the 95% confidence interval was greater than 1). Moreover, one positively selected site was detected in modules 3, 4, 6, 7 and 9, two sites in modules 8, 10 and 12 and three sites in module 11. By contrast, 74 sites in PAR showed $K_a/K_s > 1$, but only six were statistically significant (one in module 7, two in modules 10 and 11 and one in module 12). This discrepancy can be explained both by the low sequence similarity of gene *Gro1-5* compared to the other paralogues and the small number of sequences (seven) within the dataset. Such factors raised K_a/K_s values for many sites. Indeed, the inferred K_a/K_s ratios are known to depend on several factors, including for instance the number of homologue sequences used and the multialignment quality (Stern *et al*. 2007). However, in the Bayesian method, evaluating the confidence interval of the posterior distribution inferred for the position, only statistically significant sites are considered reliable, thereby diminishing the possibility of errors in positive detection. Comprehensively, comparing positively selected sites of ORT and PAR, we noted that four sites were common to both sets (147E, 233I, 257R and 261H), 10 were specific to ORT (56T, 83S, 131K, 170H, 174S, 215N, 229S, 262G, 278G and 280F) and two to PAR (232 E and 300S).

Moreover, as shown in figure 3, peaks of the K_a/K_s values in the ORT were clearly distinct. This indicates a remarkable conservation of structural sites (aliphatic sites), which had very low K_a/K_s values since they were under purifying selection, and a significant variability in recognition sites (exposed to solvent) with high K_a/K_s values, since they were under positive selection. By contrast, K_a/K_s peaks in the PAR were lower and less ordered, as indicating that they do not follow a common evolutionary path. Finally, purifying and positive selection values in PAR underlay possible changes

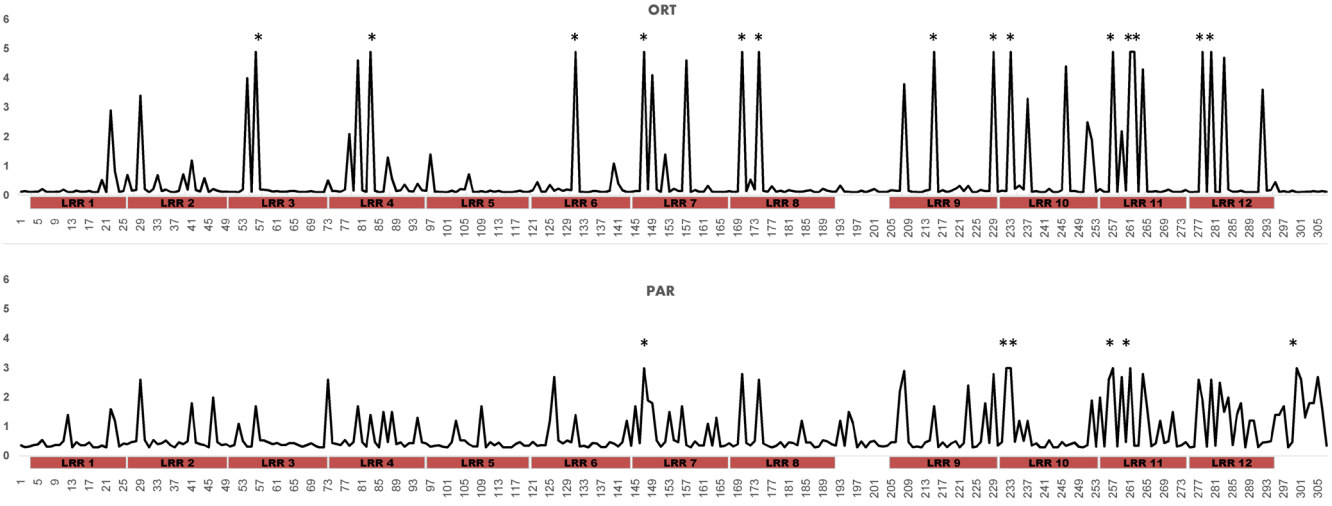

Figure 3. Selection results from M8 tests conducted on ORT and PAR sets. Positively selected sites that are statistically significant are marked using *. The 12 red boxes represent the modules of LRR domain.

in their evolutionary function. This could mean that some of the genes in the *Gro1* cluster gradually lost their ancestor recognition specificity.

MEC tests: To evaluate selection intensity in each site we considered K_a/K_s values inferred by the mechanistic empirical combination (MEC) model. This model takes into account the differences between amino acid replacement probabilities. Thus, under the MEC model, a position with radical replacements will obtain a higher K_a value than a position with more moderate replacements. As expected, a smaller number of positively selected sites was scored under the MEC model than under M8 but the overall trend was reliable. No new sites under this test were detected. Nine sites showed statistically significant values in ORT and two sites in PAR. From a comparison of K_a/K_s values computed by the MEC model in ORT and PAR with M8 values (table 3), the MEC test showed that radical replacements mainly occurred in ORT at sites 56T, 170H, 257R, 262G, 278G and 280F, and in PAR at sites 257R and 261H. Most of these sites belong to a β-strand structure, confirming the high affinity of this region to accumulate variability. Moreover, distribution of these hot spots on protein sequence showed that modules 8, 11 and 12 are very important for elicitor host interaction.

Independent of the test used (either M8 or MEC), our results on positive selection analysis revealed that the leucine-rich repeat domain had accumulated variability in a very typical way and in a particular subdomain (β-sheet). We found that a specific combination of amino acids in these sites are necessary for pathogen recognition. These results confirm previous studies on positive selection in the R protein with the LRR domain. For instance, a genomewide survey of *NBS-LRR R*-gene polymorphisms in *Arabidopsis thaliana* showed that LRR regions tends to be highly variable and that positively selected positions were disproportionately located in the LRR domain, particularly accumulating in the β-strand submotifs (Mondragon-Palomino *et al*. 2002). Other examples are reported also for cultivated flax, where two NBS-LRR resistance protein (P and P2) specificities are a result of just six amino acid polymorphisms found in the LRR β-sheet motif (Dodds *et al*. 2001), and LRR swap experiments in other resistance loci (L protein) showed novel pathogen effector recognition specificity (Dodds *et al*. 2006; Ellis *et al*. 2007). Moreover, three putative solvent exposed residues in the LRR domain of the *Cf-4* resistance gene confer specificity to recognize the fungal avirulence determinant *Avr4* in tomato (Van der Hoorn *et al*. 2001; Wulff *et al*. 2009b). Our results confirmed those of previous studies on the distribution of positively selected sites in the LRR domain and indicated that positive selection is predominantly targeted on β-sheet motif according to the host–pathogen coevolution and selection model (Flor 1971; Dangl and Jones 2001; Jiang *et al*. 2007).

Table 3. Positively selected sites under M8 and MEC tests in the ORT and PAR dataset.

	ORT		PAR	
Site	M8	MEC	M8	MEC
56 T	4.9 *	6.1 *	–	–
83 S	4.9 *	5.0 *	–	–
131 K	4.9 *	–	–	–
147 E	4.9 *	–	3.0 *	–
170 H	4.9 *	6.1 *	–	–
174 S	4.9 *	–	–	–
215 N	4.9 *	5.8 *	–	–
229 S	4.9 *	5.9 *	–	–
232 E		–	3.3 *	–
233 I	4.9 *	–	3.3 *	–
257 R	4.9 *	6.1 *	3.0 *	5.8 *
261 H	4.9 *	–	3.0 *	5.7 *
262 G	4.9 *	6.1 *	–	–
278 G	4.9 *	6.1 *	–	–
280 F	4.9 *	6.1 *	–	–
300 S	–	–	3.0 *	–

Mapping of positively selected sites

To visualize the variability and selection forces operating on individual amino acid sites in the *Gro1* LRR domain in both sets, we combined WebLogo multialignment representation (Crooks *et al*. 2004) with 3D visualization (figure 4). A 3D structure was obtained by using I-Tasser prediction software (http://zhanglab.ccmb.med.umich.edu/I-TASSER/). Values are shown on a 3D model in a discrete colour scale, with positively selected sites in dark yellow, purifying sites in deep purple and neutral selection sites in white.

Figure 4a shows that positively selected sites in ORT accumulate in the concave part of the structure, mainly in non-aliphatic residues of the β-strand motif sequence. In particular, the substitutions in the first part of the domain (56T, 83S, 131K and 147E) occurred with amino acids with similar hydrophobicity (same colour in WebLogo representation), while subsequent substitutions occurred with amino acids with different hydrophobicity values. This indicates that the initial portion of the LRR domain has undergone positive evolution, focussing on maintaining the same hydrophobic profile while the C-terminal portion has undergone positive selection aimed at increasing variability especially in sites of the β-strands exposed to the solvent phase. In particular, the last three modules of the LRR domain showed hypervariable β-strands. Since the LRR domain is strongly repeated in the 3D structure, these hypervariable regions were arranged consecutively, so that sites 170H, 233E, 257R and 280F were very close to each other in a curved horseshoe. The amino acids of β-strands of these modules might play a major role in pathogenic elicitor recognition. By contrast, aliphatic sites (Leu, Iso, Val) of β-strands in these modules are well conserved (deep purple balls in figure 4a), suggesting that selection acted in this region of the protein in a very specific way. In this regard, we also compared the hydrophobic

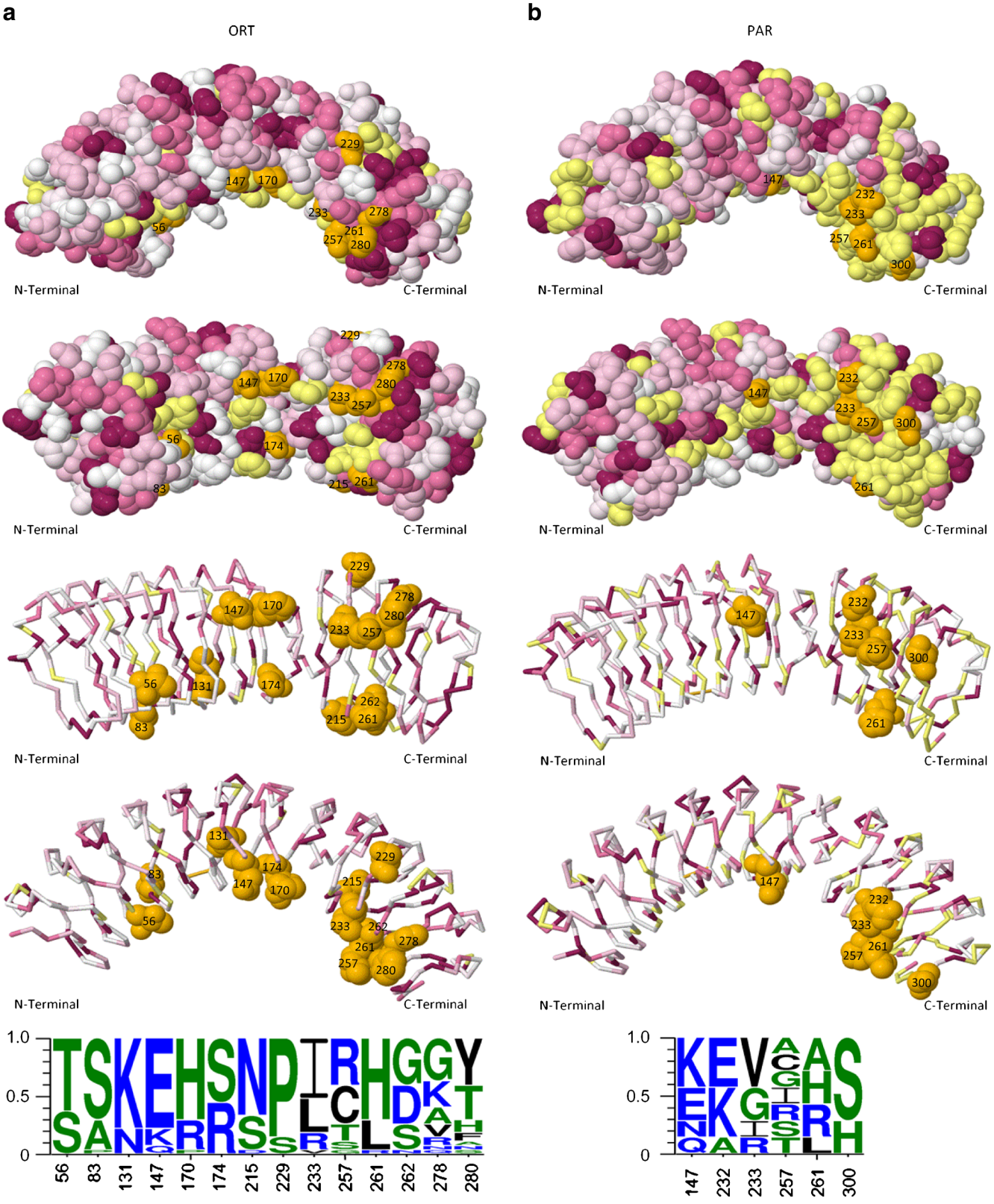

Figure 4. Positively selected sites on a 3D structure of LRR domain of *Gro1-4* for (a) ORT dataset and (b) PAR dataset. Positively selected sites are in dark yellow, purifying sites in deep purple and neutral selection sites in white. 3D structures are represented both as spacefill and backbone with only positively selected sites highlighted. Lateral and frontal side of each representation is reported. Numbers are referred to the position of amino acids. WebLogo results of positively selected sites from ORT and PAR are shown at the bottom of figure. Hydrophobic amino acids: RKDENQ (blue); neutral aa: SGHTAP (green) and hydrophilic aa: YVMCLFIW (black).

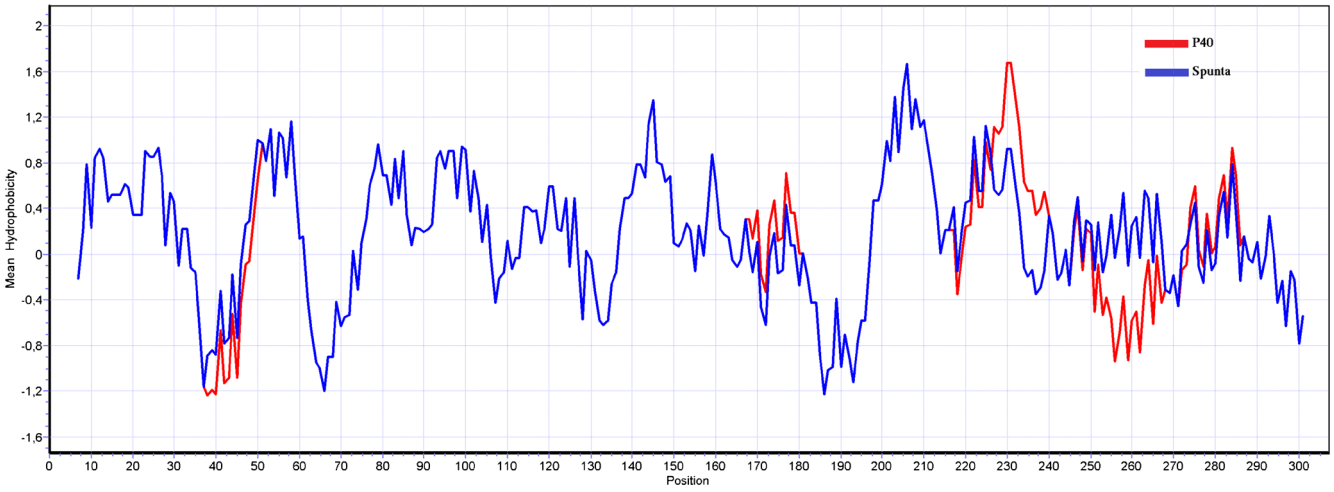

Figure 5. Hydrophobicity profile of *Gro1-4* LRR sequence of resistant genotype 'P40' compared to susceptible genotype 'Spunta.' Hydrophobicity is based on the Kyte–Doolittle scale. Three regions on sequence show divergence between the two genotypes; in particular, regions 257–262 show the opposite trend in hydrophobic profile.

profiles of the two reference species P40 and Spunta for resistance and susceptibility to *G. rostochiensis*, respectively.

As shown by the hydrophobic profiles of P40 (red line) and Spunta (blue line) (figure 5), the region around 170H, 229–233E, and around 257R–262G displayed great differences (unlike the other sites under positive selection in the first part of the sequence where no significant divergences were found), suggesting that the amino acids in this region may be critical in ligand interaction and suggesting that a specific combination of these amino acids in P40 genotype was required to confer resistance to pathotype Ro1 of *G. rostochiensis*.

With regard to PAR (figure 4b), similar to ORT we noted that the first half of the domain preserved the hydrophobicity trend in replacements while the second half showed a variation in hydrophobicity profiles, in particular for sites 233E, 257R and 261H. The smaller number of sites under positive and purifying selection suggests the possibility of having paralogue genes that are nonfunctioning or have different functions in the *Gro1* cluster.

However, detection of four positively selected sites shared by both sets in this study might indicate that interaction mechanisms were very similar in orthologue and paralogue genes. Hence, the common sites 147E, 233I, 257R and 261H could represent the core sites required for mechanisms of elicitor recognition, while the specific sites could confer resistance to different pathogen or specificity to different pathotypes as well as could confer new functions in other tissues or environmental conditions.

Conclusions

Our results confirmed previous studies on the LRR domain in *R* genes, which reported the action of positive selection on particular subdomains of the LRR. Indeed, we documented the action of positive selection on the gene *Gro1* in a number of different *Solanum* species, which were investigated for the first time in one gene controlling resistance to the nematode *G. rostochiensis* in potato. Our analysis of the variability in the LRR domain of *Gro1* genes revealed that this domain has specific conserved regions, mainly related to aliphatic residues, which are important for their structural properties, and an intradomain variability mainly related to residues exposed to solvent phase, which are probably involved in interaction mechanisms. We found that positively selected sites mainly fall in exposed residues of the internal surface of the 3D horseshoe structure of the domain and are localized in two different functioning regions. The first, located in the initial portion of the LRR domain, has undergone positive evolution, focussing on maintaining the hydrophobic profile, while the C-terminal portion has undergone positive selection aimed at increasing variability, especially in sites of the β-strands exposed to the solvent phase. In conclusion, we can argue that different evolutionary forces acted between the orthologues and paralogues. Indeeed, a higher number of amino acids (14) was found to be under positive selection among ORT than to the (6) observed among PAR. A common group of positive-selected sites (147E, 233I, 257R and 261H) was detected between the two sets and this might be required for mechanisms of elicitor recognition between host and pathogen, even if the specificity of this recognition is probably due to other selected sites specific for the two sets. The model proposed here might encourage future studies focussed on the detected sites to ascertain how mutations at such sites could change the specificity of pathogen elicitor recognition.

Data Archiving Statement: Nucleotide sequence data with accession numbers JX281707–JX281734 for this study are available in GenBank.

Acknowledgements

The authors thank Mark Walters for editing the manuscript. This research was supported by the Project AGRONANOTECH funded by MiPAF.

References

Afzal A. J., Wood A. J. and Lightfoot D. A. 2008 Plant receptor-like serine threonine kinases: Roles in signaling and plant defense. *Mol. Plant Microbe Interact.* **21**, 507–517.

Barone A., Ritter E., Schachtschabel U., Debner T., Salamini F. and Gebhardt C. 1990 Localization by restriction fragment length polymorphism mapping in potato of a major dominant gene conferring resistance to the potato cyst nematode *Globodera rostochiensis*. *Mol. Gen. Genet.* **224**, 177–182.

Bergelson J., Kreitman M., Stahl E. A. and Tian D. 2001 Evolutionary dynamics of plant R-Genes. *Science* **292**, 2281–2285.

Cai D. G., Kleine M., Kifle S., Harloff H. J., Sandal N. N., Marcker K. A *et al.* 1997 Positional cloning of a gene for nematode resistance in sugar beet. *Science* **275**, 832–834.

Chai J. J., She J., Han Z. F., Kim T. W., Wang J. J., Cheng W. *et al.* 2011 Structural insight into brassinosteroid perception by BRI1. *Nature* **474**, 472–496.

Crooks G. E., Hon G., Chandonia J. M. and Brenner S. E. 2004 WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190.

Dangl J. L. and Jones G. D. J. 2001 Plant pathogens and integrated defence responses to infection. *Nature* **411**, 826–833.

De Young B. J. and Innes W. R. 2006 Plant NBS-LRR proteins in pathogen sensing and host defence. *Nat. Immunol.* **7** 12, 1243–1249.

Delport W., Scheffler K. and Seoighe C. 2009 Models of coding sequence evolution. *Brief. Bioinform.* **10**, 97–109.

Dodds P. N., Lawrence G. J. and Ellis J. G. 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.

Dodds P. N., Lawrence G. J., Catanzariti A.-M., Teh T., Wang C.-I. A., Ayliffe M. A. *et al.* 2006 Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc. Natl. Acad. Sci. USA* **103**, 8888–8893.

Doron-Faigenboim A. and Pupko T. 2006 A combined empirical and mechanistic codon model. *Mol. Biol. Evol.* **24**, 388–397.

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 J., Dodds P. and Pryor T. 2000 The generation of plant disease resistance gene specificities. *Trends Plant Sci.* **5**, 373–379.

Ellis J. G., Lawrence G. J. and Dodds P. N. 2007 Further analysis of gene-for-gene disease resistance specificity in Flax. *Mol. Plant Pathol.* **8**, 103–109.

Enkhbayar P., Kamiya M., Osaki M., Matsumoto T. and Matsushima N. 2004 Structural principles of leucine-rich repeat (LRR) proteins. *Proteins* **54**, 394–403.

Flor H. H. 1971 The current status of the gene for gene concept. *Annu. Rev. Phytopathol.* **9**, 275–296.

Fluhr R. 2001 Sentinels of disease. Plant resistance genes. *Plant Physiol.* **127**, 1367–1374.

Graur D. and Li W. H. 2000. *Fundamentals of molecular evolution*, 2nd edition. Sinauer Press, Sunderland, USA.

Hammond-Kosack K. E. and Parker J. E. 2003 Deciphering plant-pathogen communication: fresh perspectives for molecular resistance breeding. *Curr. Opin. Biotechnol.* **14**, 177–193.

Hothorn M., Belkhadir Y., Dreux M., Dabi T., Noel J. P., Wilson I. A. *et al.* 2011 Structural basis of steroid hormone perception by the receptor kinase BRI1. *Nature* **474**, 467–471.

Jaillais Y., Belkhadir Y., Balsemao-Pires E., Dangl J. L. and Chory J. 2011 Extracellular leucine-rich repeats as a platform for receptor/coreceptor complex formation. *Proc. Natl. Acad. Sci. USA* **108**, 8503–8507.

Jann O., Werling D., Chang J. S., Haig D. and Glass E. 2008 Molecular evolution of bovine Toll-like receptor 2 suggests substitutions of functional relevance. *BMC Evol. Biol.* **8**, 288.

Jia Y., McAdams S., Bryan G., Hershey H. P. and Valent B. 2000 Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. *EMBO. J.* **19**, 4004–4014.

Jiang H., Wang C., Ping L., Tian D. and Yang S. 2007 Pattern of LRR nucleotide variation in plant resistance genes. *Plant Sci.* **173**, 253–261.

Kajava A. V., Anisimova M. and Peeters N. 2008 Origin and evolution of GALA-LRR, a new member of the CC-LRR subfamily: from plants to bacteria? *PLoS One* **3**, e1694.

Kobe B. and Deisenhofer J. 1993 Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature* **366**, 751–756.

Kobe B. and Deisenhofer J. 1995 Proteins with leucine-rich repeats. *Curr. Opin. Struct. Biol.* **5**, 409–416.

Kreitman M. 2000 Methods to detect selection in populations with applications to the human. *Ann. Rev. Genomics Hum. Genet.* **01**, 539–559.

Martin G. B., Bogdanove A. J. and Sessa G. 2003 Understanding the functions of plant disease resistance proteins. *Ann. Rev. Plant Biol.* **54**, 23–61.

MacCallum C. and Hill E. 2006 Being Positive about Selection. *PLoS Biol.* **4**, e8.

McHale L., Tan X., Koehl P. and Michelmore R. W. 2006 Plant NBS-LRR proteins: Adaptable guards. *Genome Biol.* **7**, 212.

Matsushima N. and Miyashita H. 2012 Leucine-rich repeat (LRR) domains containing intervening motifs in plants. *Biomolecules* **2**, 288–311.

Meyers B. C., Shen K. A., Rohani P., Gaut B. S. and Michelmore R. W. 1998 Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. *Plant Cell* **10**, 1833–1846.

Michelmore R. W. and Meyers B. C. 1998 Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**, 1113–1130.

Milligan S. B., Bodeau J., Yaghoobi J., Kaloshian I., Zabel P. and Williamson V. M. 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.

Mondragon-Palomino M., Meyers B. C., Michelmore R. W. and Gaut B. C. 2002 Pattern of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res.* **12**, 1305–1315.

Moore R. C. and Purugganan M. D. 2005 The evolutionary dynamics of plant duplicate genes. *Curr. Opin. Plant Biol.* **8**, 122–128.

Nielsen R. and Yang Z. H. 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

Nunziata A., Ruggieri V., Frusciante L. and Barone A. 2007 Allele mining at the locus *Gro 1* in *Solanum* wild species. *Acta Hortic.* **745**, 449–456.

Paal J., Henselewski H., Muth J., Meksem K., Menéndez C. M., Salamini F. *et al.* 2004 Molecular cloning of the potato *Gro 1-4* gene conferring resistance to pathotype *Ro 1* of the root cyst nematode *Globodera rostochiensis*, based on a candidate gene approach. *Plant J.* **38**, 285–297.

Ravensdale M., Nemri A., Thrall P. H., Ellis J. G. and Dodds P. N. 2011 Co-evolutionary interactions between host resistance and pathogen effector genes in flax rust disease. *Mol. Plant Pathol.* **12**, 93–102.

Rossi M., Goggin F. L., Milligan S. B., Kaloshian I., Ullman D. E. and Williamson V. M. 1998 The nematode resistance gene *Mi* of tomato confers resistance against the potato aphid. *Proc. Natl. Acad. Sci. USA* **95**, 9750–9754.

Roth C. and Liberles D. A. 2006 A systematic search for positive selection in higher plants (*Embryophytes*). *BMC Plant Biol.* **6**, 12.

Sanseverino W., Roma G., De Simone M., Faino L., Melito S., Stupka E *et al.* 2010 PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res.* **38**, 814–821.

Stern A., Doron-Faigenboim A., Erez E., Martz E., Bacharach E. and Pupko T. 2007 Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res.* **35**, 506–511.

Swanson W. J., Nielsen R. and Yang Q. 2003 Pervasive adaptive evolution in mammalian fertilization proteins. *Mol. Biol. Evol.* **20**, 18–20.

Thomas C. M., Jones D. A., Parniske M., Harrison K., Balint-Kurti P. G., Hatzixanthis K. and Jones J. D. G. 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.

Tor M., Lotze M. T. and Holton N. 2009 Receptor-mediated signalling in plants: molecular patterns and programmes. *J. Exp. Bot.* **60**, 3645–3654.

Van der Hoorn R. A. L., Kruijt M., Roth R., Brandwagt B. F., Joosten M. H. A. J and de Wit P. J. G. M 2001 Intragenic recombination generated two distinct Cf genes that mediate AVR9 recognition in the natural population of Lycopersicon pimpinellifolium. *Proc. Natl. Acad. Sci. USA* **98**, 10493–10498.

Yang Z. and Bielawski J. P. 2000 Statistical methods for detecting molecular adaptation. *Tree* **15**, 496–503.

Wulff B. B., Chakrabarti A. and Jones D. A. 2009a Recognitional specificity and evolution in the tomato-Cladosporium fulvum pathosystem. *Mol. Plant Microbe Interact.* **22**, 1191–1202.

Wulff B. B. H., Heese A., Tomlinson-Buhot L., Jones D. A., Pena M. dl. *et al.* 2009b The major specificity-determining amino acids of the tomato Cf-9 disease resistance protein are at hypervariable solvent-exposed positions in the central leucine-rich repeats. *Mol. Plant. Microbe Interact* **22**, 1203–1213.

Zhang H., Seabra M. C. and Deisenhofer J. 2000 Crystal structure of Rab geranylgeranyltransferase at 2.0 Å resolution. *Structure* **8**, 241–251.

Zhou B., Dolan M., Sakai H. and Wang G. L. 2007 The genomic dynamics and evolutionary mechanism of the Pi2/9 locus in rice. *Mol. Plant Microbe Interact* **20**, 63–71.

Received 16 November 2013, in revised form 7 May 2014; accepted 7 July 2014
Unedited version published online: 11 July 2014
Final version published online: 29 December 2014