Plant Omics Journal POJ

POJ 9(5):319-326 (2016) ISSN:1836-3644
DOI: 10.21475/poj.09.05.16.pne126

In silico analysis of the LRR receptor-like serine threonine kinases subfamily in *Morus notabilis*

Antonios Zambounis[1,*], Fotis E. Psomopoulos[2], Ioannis Ganopoulos[3], Evangelia Avramidou[3], Filippos A. Aravanopoulos[3], Athanasios Tsaftaris[1,4] and Panagiotis Madesis[4]

[1]Laboratory of Genetics and Plant Breeding, Faculty of Agriculture, Forestry & Natural Environment, Aristotle University of Thessaloniki, P.O. Box 261, Thessaloniki GR-54124, Greece
[2]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
[3]Laboratory of Forest Genetics and Tree Breeding, Faculty of Agriculture, Forestry & Natural Environment, Aristotle University of Thessaloniki, P.O. Box 238, Thessaloniki GR54006, Greece
[4]Institute of Applied Biosciences, CERTH, Thermi, Thessaloniki, 570 01, Greece

***Corresponding author: Antonios Zambounis; antbio@yahoo.gr**

Abstract

Mulberries are important trees crops for orchards and agroforestry systems, which are plagued by many phytopathogenic fungal species. Leucine rich repeats (LRRs) receptor-like serine threonine kinases (LRR-RSTK) subfamily plays an important role in plant defense-related reactions against fungal attacks. In the present study, we mined this subfamily on *Morus notabilis*, a mulberry species whose relevant annotated genome assembly has recently become publicly available. Our aim was to decipher *in silico* the expansion and phylogeny of these genes, their homology relationships against their orthologous in woody angiosperm plant species, and the existence of positive selective pressures acting upon their LRRs. This subfamily was found to be quite abundant and diverged, comprising by 142 annotated gene members and containing a range of conserved functional domains in their C-termini, whilst their LRRs number ranged from one to 17 repeats. A phylogenetic investigation revealed 12 distinct clades based on their diverse structural profiles, mainly as a result of the fused functional domains at their C-termini. The interspecific expansion of these *M. notabilis* LRR-RSTKs has been investigated by a homology analysis across 12 other woody angiosperm species, showing that the highest proportion of homologous best BLAST hits observed primarily in *Prunus persica*, *Malus domestica* and *Theobroma cacao*. Using a series of maximum likelihood analyses, extensive episodes of positive selective pressures acting across the LRRs were observed. This overall evidence supports a potential crucial role of this diverged LRR-RSTK subfamily as a surveillance mechanism of *M. notabilis* against fungal attacks by providing rapidly evolving ligand-binding specificities.

Keywords: disease resistance; genomics-assisted breeding; leucine-rich repeat receptor-like serine threonine kinase (LRR-RSTK); phylogenetic analysis; positive selection.
Abbreviations: LRR_leucine rich repeat; MCL_markov cluster algorithm; NBS_nucleotide-binding site domain; PAML_phylogenetic analysis by maximum likelihood; PAMP_pathogen associated molecular pattern; RAxML_randomized accelerated maximum likelihood.

Introduction

Mulberry (*Morus* L.) is a genus of fast growing deciduous and outbreeding tree species which is belonging to the Moraceae family (Clement and Weiblen, 2009). At present, mulberries, such as *M. alba* and *M. nigra* are systematically cultivated in Europe and other parts of the world as valuable orchard and agroforestry species (Vijayan, 2010), since they are easily adopted in various local environments and are considered rather naturalized species (Singhal et al., 2010). Their fruits are consumed either as fresh or as industrial derivatives such as jams, marmalades, juices and wines (Vijayan, 2010). In spite of their high commercial importance, mulberry molecular breeding has been limited mainly due to absence of genomic resources. Recently, He et al., (2013) reported the first annotated genome assembly (http://www.ncbi.nlm.nih.gov/genome/17692) of a mulberry species (*M. notabilis*), with an estimated genome size of 357 Mb and 29,338 protein-coding genes to being predicted across its 14 chromosomes.

Fungal diseases are a crucial limiting factor in mulberry cultivation worldwide. Among the most destructive are the powdery mildew (*Phyllactinia corylea* (Pers.) P. Karst), leaf spot (*Cercospora moricola* Cooke), leaf rust (*Cerotelium fici* (Castagne)) and red rust (*Aecidium mori*) (Srivastava and Mehra, 2004). The development of fungal resistance cultivars through genomics-assisted breeding is an important determinant towards the establishment of efficient disease breeding strategies (Lalli et al., 2005), especially in mulberry where it is time consuming to select for fungal resistance.

Molecular recognition of phytopathogenetic fungi is mediated mainly by PAMPs at infection sites, a first crucial step of defense reactions in plants. It is commonly mediated by a plethora of rapid-evolving receptors, many of which are containing ligand-binding domains, such as Leucine-Rich Repeats (LRRs) (Diévart et al., 2011; Matsushima and Miyashita, 2012). These LRRs domains evolve novel ligand-binding specificities through a wealth of various mechanisms, such as point mutations at variable residues, variations in their repeat numbers, tandem gene duplications and conversions events, unequal crossing-over and other short rearrangements (Friedman and Baker, 2007; Meyers et al., 2003; Yang et al., 2008; Zambounis et al., 2012). Therefore, the evolution patterns of plant LRRs-containing genes appears to be rather a complex process (Sekhwal et al., 2015; Zhou et al., 2004), where signatures of positive selective pressures acting upon them are often quite evident (Khan et al., 2015). Positive selection is defined as the rapid fixation of beneficial non-synonymous mutations and is an important evolutionary force for a number of defense-related genes towards their functional diversification, which may be selectively favored in plants under intensive biotic pressures (Delph and Kelly, 2014). Even, the location of positively selected amino acid residues is crucial for obtaining novel gene functions (Mondragon-Palomino et al., 2002); Thus, previous studies have reported that solvent-exposed regions of the LRRs repeats are under strong positive selection. Such evidence has been interpreted as an additional sign of the involvement of these regions in fungal pathogens recognition (Khan et al., 2015; Parniske et al., 1997; Zambounis at al., 2012).

In plants, the receptor-like kinase (RLK) gene family is being involved in signal perception under fungal attacks and is being commonly subjected to positive selective pressures, particularly upon their LRRs domains (Fischer et al., 2009; Lefti-Shiu et al., 2009; Zan et al., 2013). Therefore, RLKs appear to play a crucial role in signaling pathways during pathogen recognition and the subsequent activation of plant defense reactions (Afzal et al., 2008). This family is subdivided into three main subfamilies according to their different substrate specificities of kinase domains, such as the receptor tyrosine kinases (RTK) that are involved in the phosphorylation of tyrosine residues, the receptor serine/threonine kinases (RSTK) that phosphorylate serine and threonine residues and the histidine kinases that are implicated in the phosphorylation of histidine residues (Becraft, 2002).

Our study focuses on members of the LRR-containing RSTK subfamily which interact with other proteins influencing a wide array of processes ranging from disease resistance reactions and developmental regulation of recognition (Diévart et al., 2011). The major group of these proteins contains members belonging to the NBS (nucleotide-binding site)-LRR class, which is quite abundant in plant genomes (Dangl et al., 2013) covering resistance across a wide range of phytopathogenic fungi (Wan et al., 2010).

These evidence underlie the basis of the present survey in order to decipher and gain insights into: (i) the expansion and phylogeny of LRR-RSTK-coding genes in the genome of *M. notabilis*, (ii) the homology relationships of the above genes across their orthologous in woody angiosperm plant species with annotated genomes, and (iii) the existence of positive selective pressures with their signatures acting upon LRRs of these genes. We hypothesize, by assessing the evolutionary profiles upon these LRRs repeats that successive episodes of positive selection might contribute to the acquisition of novel pathogens recognition repertoires at these *LRR-RSTK* genes in *M. notabilis*. Such results might provide a foundation for the overall ongoing disease resistance breeding efforts in mulberries at the future (Aravanopoulos et al., 2015).

Results

Expansion and phylogenetic analysis of LRR-RSTK gene subfamily in M. notabilis

An ample number of LRR-RSTKs were mined from the genome of *M. notabilis* in line with their common abundance in other plant genomes (Zan et al., 2013). The overall structure of these proteins was typical for the LRR-RSTK protein subfamily. It consisted of conserved LRRs regions at the N-terminus comprising of up to 17 (such as in XM_010102302) tandem LRR repeats arranged in a single continuous block, which were downstream linked to various conserved functional regions with homologies mainly to the Pkinase domain. The LRRs domain mediates the establishment of ligand-specific interactions, whilst the RSTK domains are involved in signal transductions (Zan et al., 2013). Most of these proteins gave a match at their N-terminus with the Pfam profile LRRNT_2 (PF08263), a conserved region often found in LRRs-containing proteins. Besides, high variation and divergence was observed regarding the composition and numbers of tandem LRR repeats among these proteins; four different LRR repeats [LRR_1 (PF00560.30); LRR_4 (PF12799.4); LRR_6 (PF13516.3); LRR_8 (PF13855.3)] were revealed usually in various combinations. A notable number of proteins (30 out of the 142) exhibited only the LRRs domains in their structure profiles, without any apparent sequence homology to any known conserved functional domain towards their C-terminus.

The Pfam matches at the C-termini for the 142 LRR-RSTKs are showed in Table 1. The most abundant functional domains were these of Pkinase (60 times) and of the Pkinase_Tyr (51 times); in addition, Malectin and Malectin_like domains were also predicted 24 times in total. Remarkably, these 142 LRR RSTKs were extremely diverse in terms of functional domains fusions towards their C-termini (Table 1). Pkinase_Tyr domain was found to be mostly fused with Malectin and Malectin_like domains, eight and seven times, respectively.

In order to gain insights into the origin of the expansion of these LRR-RSTKs, a RAxML-based phylogenetical analysis was conducted at their amino acid level (Stamatakis, 2014). A significant portion of the phylogenetic clades (five out of the 12) contained only a few genes (ranging from two up to five), while the backbone topology of the tree was adequately resolved. The tree consisted of 283 branches, reflecting a series of consecutive gene duplications at its terminal branches (Fig 1). The overall pairwise identity among the 142 amino acid sequences was 13.3%, indicative of a rather high degree of divergence.

The proteins comprising our dataset were grouped in 12 clades (I - XII) as a consequence of their significant variation both in their LRRs number at their N-termini and the substantial divergence of the conserved functional domains in their C-termini (Fig 1). The majority (75.35%) of LRR-RSTKs were sub-grouped in five clades (I, V, VI, XI and XII). The 60 LRR-RSTKs which were contained primarily the Pkinase domain grouped mainly in clades I and V, whilist these ones harboring the Pkinase_Tyr domain were clustered in clades I, V, VII and VIII. Finally, the LRR-RSTKs which were contained the Malectin and the Malectin-like functional domains grouped in the XI (or in the XII) and in the VI clades, respectively.

Table 1. Pfam matches of the 142 LRR-RSTKs at their C-terminus, Pfam IDs, and their occurrences along with the combinations of integrated functional domains fusions.

Functional domains at C-terminus	Pfam IDs	No of indivindual occurences	Combinations of fused functional domains	
			Pkinase	Pkinase_Tyr
Pkinase	PF00069.22	60		
Pkinase_Tyr	PF07714.14	51	2	
Malectin	PF11721.5	15	5	8
Malectin_like	PF12819.4	9	2	7
Peptidase_S10	PF00450.19	2	1	
2OG-FeII_Oxy	PF03171.17	2	1	
Atx10homo_assoc	PF09759.6	1	1	
DUF4216	PF13952.3	1		
Ribosomal_L14	PF00238.16	1		1

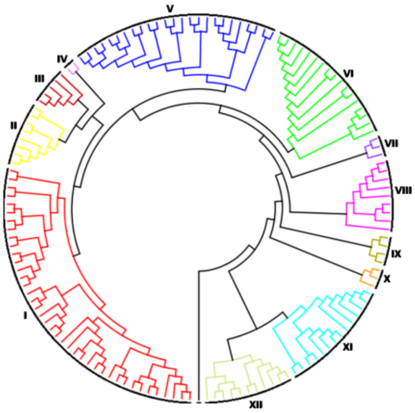

Fig 1. RAxML-based phylogenetic relationships of the 142 *M. notabilis* LRR-RSTKs amino acid sequences, which were aligned using the Muscle program. Clades in different colors are corresponding to LRR-RSTKs sequences being grouped to these 12 distinct clades (I-XII).

Table 2. Parameters of the evolutionary analysis upon the 458 LRRs which were calculated by employing the MEGA 5 and CODEML software.

LRRs dataset	LRRs sequences	alpha (gamma shape parameter)	log likelihood value (ℓ0)		$2\Delta\ell s = 2(\ell_{1-}\ \ell_0)$, (df *= 1)	Statistical significance	Positive selection in branches **
			one-ratio model	free-ratio model			
LRR-1	41	0.5	-1028.54	-477.98	1101.12	P < 0.001	YES (0.08)
LRR-2	204	0.5	-5360.05	-5787.19	854.28	P < 0.001	YES (0.012)
LRR-3	213	0.5	-5581.29	-6053.72	944.86	P < 0.001	YES (0.014)

* Degrees of freedom refer to difference in number of parameters between the two models.

** The numbers in parentheses indicate the ratio of branches under positive selection.

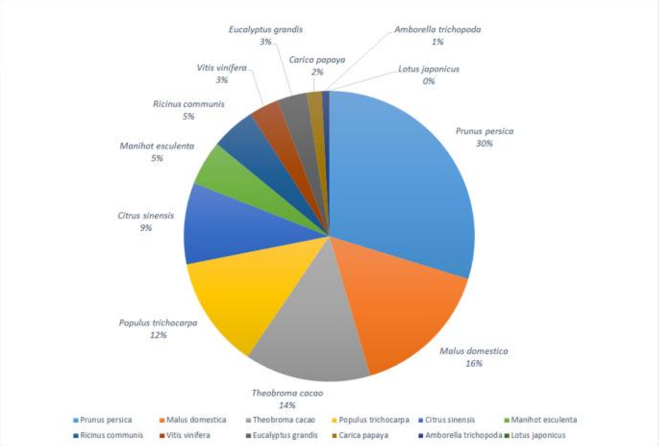

Fig 2. Distribution of the top BLAST hit of each of the 142 *M. notabilis LRR-RSTK* genes across the 12 woody angiosperm plant species in the Plaza Dicots 3.0 database. For each of the LRR-RSTK sequences, only homologue with the best BLAST hit, in means of e-value and identity, was retained and the corresponding target species was reported. The final number of homologues per species was consequently used to construct the pie chart. Over half of the LRR-RSTK homologous sequences (62%) identify matches to just four species, i.e. *Prunus persica*, *Malus domestica*, *Theobroma cacao* and *Populus trichocarpa* in descending order.

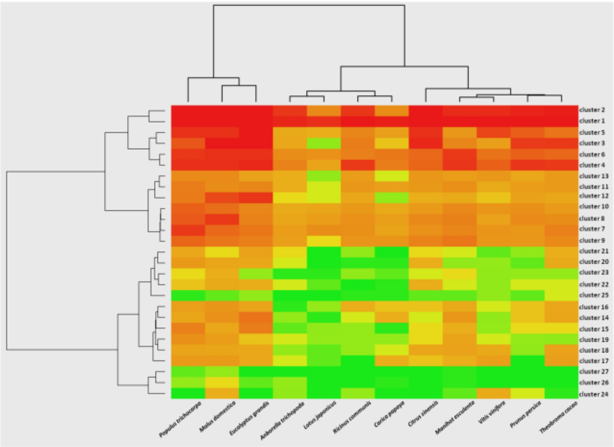

Fig 3. Heatmap of the 27 constructed clusters across the selected 12 woody angiosperm plant species, based on the number of *M. notabilis* homologues sequences identified in each participating genome. The rows of the heatmap correspond to the clusters, whereas the columns to the 12 plant species. Each cell ($c_{i,j}$) represents the number of homologous sequences from species j that are members of cluster i, with actual values being color coded from green to dark red in ascending order. Both rows and columns have been independently and hierarchically clustered, and the corresponding trees are shown around the heatmap.

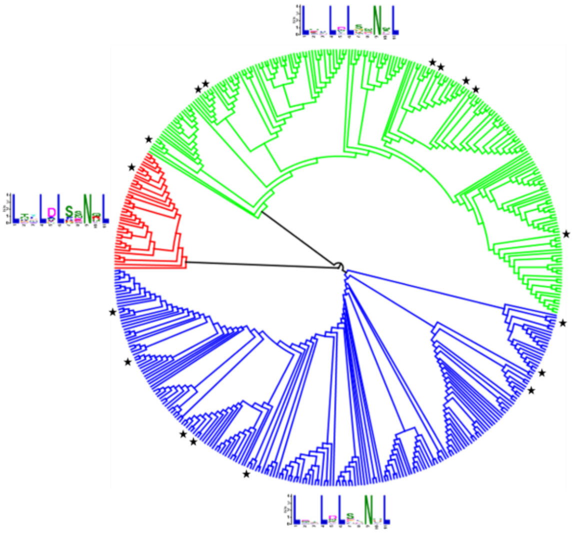

Fig 4. RAxML phylogeny of the 458 LRRs amino acid sequences being identified among the 142 M. notabilis LRR-RSTK genes and which were aligned using the Muscle program. Phylogenetic clades in different colors are corresponding to the three distinct LRRs groups with their conserved motifs being depicted beside. With asterisks are indicated the locations of the 17 tandem LRRs from XM_010102302 LRR-RSTK which is containing the highest number of LRRs among all the 142 M. notabilis LRR-RSTKs genes.

Homologies-based comparative genomics of the LRR-RSTK subfamily

In order to investigate the evolutionary-based distribution of the *LRR-RSTKs* genes of *M. notabilis* across other woody angiosperm plant genomes, we performed a comparative genomic analysis using the 12 selected species listed in Fig 2 as reference genomes. The reference sequence data were retrieved from the Plaza 3.0 Dicots database (Proost et al., 2015). Aiming at the identification of the closest homolog for each of the 142 LRR-RSTKs, only the best BLAST hit in the reference database was retrieved using all-against-all blast(p) searches of the blast algorithm (e-value cutoff < 10^{-10}), keeping all other parameters at their default values. A pie chart distribution of the best BLAST hits across the 12 species is shown in Fig 2, providing a clear overview of the LRR-RSTKs participation in each genome. As it shown in Fig 2, the highest proportion of homologous genes was observed in *Prunus persica* (30%), in *Malus domestica* (16%) and in *Theobroma cacao* (14%). Astonishingly, top homologous BLAST hits were not observed in *Lotus japonicas*, whereas highly homologous genes were observed in *Carica papaya* (2%) and in *Amborella trichopoda* (1%), an indication of the significant expansion and diversity of these genes among the woody angiosperm species tested in this study.

Furthermore, in order to bring forth the evolutionary-based distribution of these 142 LRR-RSTK sequences across the woody angiosperm genomes, the sequences were clustered through the MCL algorithm by a similarity cutoff of 45% and using all identified homologues, i.e. not only the best BLAST hits per sequence. This time, the 27 constructed clusters were consequently re-organized into coherent groups, using the maximum distance as the driving metric, and Ward.D2 as the organizing hierarchical clustering method. Fig 3 clearly shows an internal organization of both the clusters, as well as the corresponding reference genomes that can be interpreted as the evolutionary-based distribution of the LRR-RSTK subfamily across the woody angiosperm plant species studied.

In particular, two main groups of clusters were observed; clusters 1 and 27 were the most conserved in hits abundance among the 12 plant species (Fig 3). These species were classified into three main sub-groups. The largest number of hits (69.04%) was matching to *Malus domestica, Eucalyptus grandis, Populus trichocarpa, Citrus sinensis,* and *Theobroma cacao*.

Conserved motifs, phylogeny and maximum likelihood evolutionary analysis of LRRs

We identified by keyword searches 458 LRRs distributed among the respective LRR domains of the 142 LRR-RSTKs in order to evaluate whether positive selection is acting upon the core sequences of these LRRs. We identified 453 unique LRRs, whilst there were 25 LRRs duplicates being distributed among both identical and different *LRR-RSTK*s genes. These LRRs were distributed in tandem repeats with a maximum number of 17 of such repeats per gene in case of XM_010102302.

The phylogenetic RAxML-based reconstruction revealed that these 458 LRRs in their amino acid level were grouped along three distinct clades with an overall pairwise identity of 53%, implying a rather high diversity among them. An important finding was the totally gene-independent phylogenetical distribution of these repeats; for example, the 17 LRRs repeats of the XM_010102302 gene were distributed across both the three clades of the phylogeny (Fig 4).

The Multiple EM for Motif Elicitation (MEME) motif detection software (Bailey et al., 2006) was used for the identification of LLRs conserved motifs at each of their clade. Three LRRs-related motifs for each clade were identified (Fig 4): LT[YF]LDLSSN[QR]L (LLR-1), LQXLDLSNN[NS]L (LLR-2) and LXXLDLSXNQL (LLR-3). Except of the five highly conserved amino acids (Leu at positions 1,4,6,11 and Asn at position 9), the most conserved amino acid residues were: (i) for the LRR-1 motif, Asp at position 5 and Ser at position 7, and (ii) for the LRR-2 and LRR-3 motifs, Ser at position 7. Furthermore, Ser or Asn at position 8, as well Gln or Arg at position 10 were being often substituted each other in the LRR-1 motif, whereas Asp or Asn at position 5 for the LRR-2 and LRR-3 motifs, respectively (Fig 4).

Finally, we investigated whether accelerated evolution and particularly signatures of positive selection might have also contributed to the LRRs divergence of the LRR-RTSKs in *M. notabilis*. Analyses were performed separately for the three LRRs datasets (LRR-1, LRR-2, LRR-3) (Table 2), each containing the LRRs sequences from each of the three phylogenetic clades (Fig 4). The average codon-based evolutionary divergence was calculated over all sequence pairs for each dataset, using the MEGA 5 software (Tamura et al., 2011). In all instances, the average numbers of synonymous nucleotide substitutions per site (dS) mutations over all sequence pairs were below the value of 2.00, the cut-off above which the corresponding sequences would have to be excluded from CODEML analyses, in order to bypass any saturation effects based on nucleotide substitutions effects (Yang and Nielsen, 2000). Subsequently, all three LRRs datasets were subjected to successive tests of positive selection using the CODEML program implemented in the PAML version 4.8 package (Yang 2007). Statistically significant (P<0.001) evidence of positive selection with omega (ω) ratios [non-synonymous (dN) per synonymous (dS) nucleotide substitution mutations] higher than 1.00 being observed in numerous branches at the respective evolutionary trees in all the three datasets (Table 2). These selective signatures were spreading across the entire trees topologies and were evident both in the terminal and in the ancient branches. Thus, CODEML analysis allows us to confirm that recent episodes of positive selection were overlapping similar more ancient events towards the evolution of these LRRs sequences in the LRR-RSTK subfamily at the *M. notabilis* genome.

Discussion

Trees crops in orchard and agroforestry systems are substantially decreased by harmful invading fungal pathogens both in terms of fruit yield and quality. In order to combat these pathogens, plant species generally rely on their innate immunity in order to deploy a wealth of intricate pathways to recognize the relevant signals of fungal infection (Perazzolli et al., 2014). Immune receptors are merely the onset of these complex signal transduction pathways to enable finally transcription of target defense-related genes (Sekhwal et al., 2015). Plant genes members of the LRR-containing RSTK subfamily are involved both in disease resistance reactions and in the developmental regulation of recognition (Diévart et al., 2011). However, deciphering *LRR-RSTK* genes functions is complicated due to the functional redundancy between these receptors in plants. *Arabidopsis* has 216 *LRR-RSTK* encoding genes, but only very few have been associated with defined biological functions (Afzal et al., 2008). In our study, we focused on *M. notabilis*, a mulberry species whose relevant genome assembly has recently become publicly available (He et al., 2013). We mined the 142 LRR-RSTKs-encoded sequences (Supp Table 1), which were grouped into 12 clades based on their structural and sequence similarities. The variations in motifs composition were analyzed in order to infer their phylogenetic relationships. The dispersal phylogeny pattern of these LRR-RSTKs, which was indicative of potential divergence in their functions, prompted consideration of their role, assuming that LRR-RSTKs belonging to distinct clades perform different functions in a broad spectrum of defense interactions. According to Zan et al., (2013) the high abundance of the *Populus* LRR-RLK gene family is implying a great demand for these genes to participate in more complex transcriptional pathways in woody plant species, as it was also postulated by Lehti-Shiu et al., (2009). Our results showed that there a fairly large number of *LRR-RSTKs*-encoding genes contained a functional domain named Malectin that presents a relatively novel functionality. This putative extracellular, C-terminus domain is likely involved in carbohydrate binding of maltose-like glucose oligomers (Schallus et al., 2008). A similar Malectin-like *LRR-RLK* gene was found to contribute to downy mildew disease in *Arabidopsis* (Hok et al., 2011). Furthermore, we observed evidence of various functional domains fusions, predominantly at the C-termini of *M. notabilis* LRR-RSTKs, mostly integrated by the Pkinase_Tyr and the Malectin or Malectin-like domains. Recent studies demonstrated that LRR-containing proteins with non-typical domain architectures, mainly these that are including “integrated decoys”, play a crucial role in plant immunity with an integration re-occurring and an evolutionarily conserved pattern being evident across numerous species lineages (Sarris et al., 2016). A previous survey (Cesari et al., 2014) revealed similar and unique integrated domain fusions to LRR-containing proteins, commonly occurred in mosses and across numerous lineages of flowering plants, enabling by this way an efficient fungal recognition. According to Sarris et al., (2016) the most abundant group of domains fusions at LRR-containing proteins is involving the kinase domain found in receptor-like kinases that transducer PAMP-triggered immunity in plants; these findings are in line with our results, as such integrated domains were quite frequent in our *M. notabilis* LRR-RSTK dataset. The primary genomic assembly of *M. notabilis* (He et al., 2013) does not include any assembled chromosomes or linkage groups (Li et al., 2014), hence we could not test whether these genes are found at tandem clusters. Such a gene topology seems to be the main driving mechanism for the expansion of the RLK family in *Oryza sativa*, where such tandem duplications may result in gene fusions providing novel functionality (Shiu et al., 2004). Furthermore, our study reveals that the highest proportion of homologous genes counterparts was found in *Prunus persica, Malus domestica, Eucalyptus grandis, Populus trichocarpa, Citrus sinensis,* and *Theobroma cacao,* which is indicative of the high expansion and diversity of these genes among the woody angiosperms. The LRRs domains are formed by the juxtaposition of up to 40 individual repeats, while some of the variable amino acids between the structural leucine residues engage in specific interactions with other ligands (Matsushima and Miyashita, 2012). In turn, the individual LRRs sustain their typical b-sheet structures and are most often key components in acquisition of novel fungal recognition specificities (Zambounis et al., 2012). Since the repetitive structure of LRRs accounts for the rapid generation of new variants by duplications or deletions of entire repeats, the overall LRRs profiles of the *M. notabilis* LRR-RSTKs have been regarded as important parameter in order to reflect their evolutionary and functionality history. Our findings postulate also that LRRs of LRR-RSTKs exhibit signs of positive selection. Such a selective mode is promoting a rapid alteration of gene sequences in different alleles or species by a non-directional mode without altering the main functions, which in cases of fungal attacks, often increases the reservoir of the ligands that can be recognized in these interactions (Delph and Kelly, 2014).Positive selection on NBS-containing gene families in plants and especially among their LRRs domains, were reported in various studies (Chen at al., 2010; Perazzolli et al., 2014; Khan et al., 2015), supporting the view that selection upon them for durable disease resistance might be a crucial component of nearly all plant breeding programs (Mace et al., 2014). Comparative analyses among Rosaceae trees species regarding their resistance (*R)* genes have revealed that solvent-exposed residues of the LRRs domains are hyper-variable, with intensive positive selective pressures acting on them (Perazzolli et al., 2014). Similar results were constantly observed in a range of plant species (Li et al., 2010; Zambounis et al., 2012; Yang et al., 2013; Khan et al., 2015). We hypothesize that these positive selective pressures acting upon LRRs of the LRR-RSTK subfamily in *M. notabilis* might being associated mainly with a continual selective demand for their functional diversification in terms of increasing fungal disease resistance.

Materials and methods

Genes mining, structures and phylogenetic reconstruction of LRR-RSTKs genes in M. notabilis genome

Based on keyword searches, both the transcript and the protein sequences of the 142 *LRR-RSTK* annotated genes from the genomic assembly (ASM41409v2) of *M. notabilis* (http://www.ncbi.nlm.nih.gov/genome/17692) were mined. All amino acid sequences contained valid structures and were searched against the Pfam database (http://pfam.sanger.ac.uk/search#tabview=tab1) to assess their functional domains predictions (Table 1).

The phylogenetic relationships among these *LRR-RSTK* genes were revealed by performing a Muscle alignment (Edgar, 2004) and a tree reconstruction, using the RAxML program (Stamatakis, 2014) with default parameters (Fig 1). All the above analyses were performed using the Geneious R7 platform (Kearse at al., 2012). The 12 clades (I-XII) that were revealed in the phylogenetical analysis were further validated using the following criteria for each clade: (i) overall similarity throughout the majority of the coding sequences, (ii) no or few (up to four) gaps across the aligned sequences, and (iii) at least 50% amino acid average identity.

Comparative analysis of LRR-RSTKs among woody angiosperm plant species

Given the functional and phylogenetical relationships of the LRR-RSTKs and in order to investigate their evolutionary-based distribution across other woody angiosperm plant genomes, we performed a comparative genomic analysis against 12 selected species listed in Fig 2 as reference genomes. All the reference sequence data were retrieved from the Plaza 3.0 Dicots database (Proost et al., 2015).

Initially, aiming at the identification of the closest homologous for each of the 142 LRR-RSTKs, only the best BLAST hit in the reference database was retrieved using the blast(p) algorithm with the default parameters. The sequences were consequently clustered using the MCL (Markov Cluster Algorithm) approach with an inflation value of 2.00 (Enright et al., 2002). The clustering process utilized the identity column of the all-against-all blast(p) output as a distance metric, using the default values for any of the tool options. A pie chart distribution of the best BLAST hits across the 12 woody angiosperm species is shown in Fig 2, providing a clear overview of the LRR-RSTKs participation in each genome.

Furthermore, in order to bring forward the evolutionary-based distribution of these 142 LRR-RSTKs across the 12 woody angiosperm plant species genomes, the sequences were clustered through the MCL algorithm as above using all the identified homologues, in addition to the best BLAST hits, for each *M. notabilis* LRR-RSTK sequence. The 27 constructed clusters were consequently re-organized into coherent groups, by performing a hierarchical clustering analysis both at the cluster level, as well at the target genome species level (Fig 3). The clustering processes employed the maximum distance between vectors as the driving metric, and Ward.D2 as the organizing method in order to produced meaningful groupings.

Identification of LRRs, assignment of conserved motifs, phylogenetic and evolutionary analyses

The core sequences (11 amino acids in length) of the LRRs were retrieved from all the 142 LRR-RSTK proteins by keyword searches using the LRRs typical LxxLxLxxNxL core sequence stretch (Matsushima and Miyashita, 2012). The phylogenetic relationships among the 458 identified LRRs sequences (Fig 4) were investigated as described above for the 142 LRR-RSTKs. In order to exhibit the structural divergence of these repeats across their three clades in their phylogenetic tree, the conserved amino acid motifs were detected using the Multiple EM for Motif Elicitation (MEME) motif software v.4.9.0 (http://meme.sdsc.edu/meme/intro.html) (Bailey et al., 2006). The setting of maximum number of motifs was at 1.00 and of the minimum motif width at 11.00 with default values for all the other parameters.

The 458 LRRs which were phylogenetically grouped in three distinct clades (Fig 4) and counting up to 41 LRRs per clade (Table 2) were assigned separately for positive selection signatures, using the CODEML program from the PAML version 4.8 package (Yang, 2007). We thus searched for evidence of positive selection on three independent LRRs datasets at a clade-dependent mode. Initially, any substitution saturation effects were estimated by calculating the dS rates between the aligned nucleotide sequences using the altered Nei-Gojobori (1986) method presented by Yang and Nielsen, (2000). Then, CODEML program was specifically inferred, separately in all three LRRs datasets, for investigating variable positive selective pressures among branches in the phylogeny by examining significant differences at ω ratios. Two evolutionary scenarios were assigned using the CODEML program: one presuming equal ω-ratio for all branches in phylogeny, and one (free-ratios model) permitting an independent ω-ratio along the different branches. The detailed methodology for these analyses is described in Zambounis et al., (2012). The log-likelihood values for each model were compared by a likelihood ratio test (LRT) assigning a *P*-value. Posterior Bayesian probabilities were estimated for codon substitutions in each branch. All datasets including alignments and outputs of the CODEML program are available upon request.

Conclusion

The number of genes members of the LRR-RSTKs subfamily was fairly abundant in *M. notabilis* genome. These genes were found to contain a range of conserved functional features and domains in their C-termini, whilst the number of LRRs ranged from one up to 17 repeats. According to their phylogenetic relationships, 12 distinct clades were observed, based mainly on their fused patterns of functional domains. Homology analysis revealed that the highest proportion of homologous best BLAST hits among the woody angiosperms studied, were observed primarily in *Prunus persica*, in *Malus domestica* and in *Theobroma cacao*. Extensive episodes of positive selective pressures acting across the 458 LRRs of these *M. notabilis* LRR-RSTKs were also postulated in the relevant LRRs lineages. This evidence is enriching the rationale regarding the crucial role and the engagement of these genes in specific interactions with other ligands, providing novel, rapidly evolving ligand-binding specificities and assigning them as a surveillance mechanism against fungal attacks. As the genomic resources of other mulberry species as *M. alba* increase, the interspecific validity of the above results, may form a notable component in mulberries molecular breeding towards fungal disease resistance.

Acknowledgments

This research has been co-financed by the European Union (European Social Fund-ESF) and Hellenic national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracletus-II.

References

Afzal AJ, Wood AJ, Lightfoot DA (2008) Plant receptor-like serine threonine kinases: roles in signaling and plant defense. Mol Plant Microbe Interact. 21: 507-517.

Aravanopoulos FA, Ganopoulos I, Tsaftaris A (2015) Population and conservation genomics in forest and fruit trees. Adv Bot Res. 74: 125-156.

Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucl Acids Res. 34: 369-373.

Becraft PW (2002) Receptor kinase signaling in plant development. Ann Rev Cell Dev Biol. 18: 163-192.

Cesari S, Bernoux M, Moncuquet P, Kroj T, Dodds PN (2014) A novel conserved mechanism for plant NLR protein pairs: the ‘integrated decoy’ hypothesis. Front Plant Sci. 25: 606.

Chen Q, Han Z, Jiang H, Tian D, Yang S (2010) Strong positive selection drives rapid diversification of *R*-genes in *Arabidopsis* relatives. J Mol Evol. 70: 137-148.

Clement WL, Weiblen GD (2009) Morphological evolution in the mulberry family (*Moraceae*). Syst Bot. 34: 530-552.

Dangl JL, Horvath DM, Staskawicz BJ (2013) Pivoting the plant immune system from dissection to deployment. Science. 341: 746-751.

Delph LF, Kelly JK (2014) On the importance of balancing selection in plants. New Phytol. 201: 45-56.

Diévart A, Gilbert N, Droc G, Attard A, Gourgues M, Guiderdoni E, Périn C (2011) Leucine-rich repeat receptor kinases are sporadically distributed in eukaryotic genomes. BMC Evol Biol. 11: 367.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics. 5: 1-19.

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30: 1575-1584.

Fischer I, Diévart A, Droc G, Dufayard J-F, Chantret N (2009) Evolutionary dynamics of the leucine-rich repeat receptor-like kinase (LRR-RLK) subfamily in angiosperms. Plant Physiol. 150: 12-26.

Friedman AR, Baker BJ (2007) The evolution of resistance genes in multi-protein plant resistance systems. Curr Opin Genet Dev. 17: 493-499.

He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, Lee TH, Wang X, Cai Q, Li D, Lu M, Liao S, Luo G, He R, Tan X, Xu Y, Li T, Zhao A, Jia L, Fu Q, Zeng Q, Gao C, Ma B, Liang J, Wang X, Shang J, Song P, Wu H, Fan L, Wang Q, Shuai Q, Zhu J, Wei C, Zhu-Salzman K, Jin D, Wang J, Liu T, Yu M, Tang C, Wang Z, Dai F, Chen J, Liu Y, Zhao S, Lin T, Zhang S, Wang J, Wang J, Yang H, Yang G, Wang J, Paterson AH, Xia Q, Ji D, Xiang Z (2013). Draft genome sequence of the mulberry tree *Morus notabilis*. Nat Commun. 4: 2445.

Hok S, Danchin EGJ, Allasia V, Panabieres F, Attard A, Keller H (2011) An *Arabidopsis* (malectin-like) leucine-rich repeat receptor-like kinase contributes to downy mildew disease. Plant Cell Environ. 34: 1944-1957.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P, Drummond A (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28: 1647-1649.

Khan AM, Khan AA, Azhar MT, Amrao L, Cheema HM (2015) Comparative analysis of resistance gene analogues encoding NBS-LRR domains in cotton. J Sci Food Agr. 96: 530-538.

Lalli DA, Decroocq V, Blenda AV, Schurdi-Levraud V, Garay L, Le Gall O, Damsteegt V, Reighard GL, Abbott AG (2005) Identification and mapping of resistance gene analogs (RGAs) in *Prunus*: a resistance map for Prunus. Theor Appl Genet. 111: 1504-1513.

Lehti-Shiu MD, Zou C, Hanada K, Shiu SH (2009) Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. Plant Physiol. 150: 12-26.

Li T, Qi X, Zeng Q, Xiang Z, He N (2014) MorusDB: a resource for mulberry genomics and genome biology. Database (Oxford), 2014, bau054.

Li J, Ding J, Zhang W, Zhang Y, Tang P, Chen J-Q, Tian D, Yang S (2010) Unique evolutionary pattern of numbers of gramineous NBS-LRR genes. Mol Genet Genomics. 283: 427-438.

Mace ES, Tai SS, Innes DJ, Godwin ID, Hu WS, Campbell BC, Gilding EK, Cruickshank A, Prentis PJ, Wang J, Jordan DR (2014) The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes. BMC Plant Biol. 14: 253.

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

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

Mondragón-Palomino M, Meyers BC, Michelmore RW, Gaut BS (2002) Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Res. 12: 1305-1315.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. Mol Biol Evol. 3: 418-426.

Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BBH, Jones JDG (1997) Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. Cell. 91: 821-832.

Perazzolli M, Malacarne G, Baldo A, Righetti L, Bailey A, Fontana P, Velasco R, Malnoy M (2014) Characterization of resistance gene analogues (RGAs) in apple (*Malus x domestica* Borkh.) and their evolutionary history of the Rosaceae family. PLoS ONE. 9: e83844.

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K (2015) PLAZA 3.0: an access point for plant comparative genomics. Nucl Acids Res. 43: 974-981.

Sarris PF, Cevik V, Dagdas G, Jones JD, Krasileva KV (2016) Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. BMC Biology. 14: 8.

Schallus T, Jaeckh C, Fehér K, Palma AS, Liu Y, Simpson JC, Mackeen M, Stier G, Gibson TJ, Feizi T, Pieler T, Muhle-Goll C (2008) Malectin: a novel carbohydrate-binding protein of the endoplasmic reticulum and a candidate player in the early steps of protein N-glycosylation. Mol Biol Cell. 19: 3404-3414.

Sekhwal MK, Li P, Lam I, Wang X, Cloutier S, You FM (2015) Disease resistance gene analogs (RGAs) in plants. Int J Mol Sci. 16: 19248-19290.

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH (2004) Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. Plant Cell. 16: 1220-1234.

Singhal BK, Khan MA, Dhar A, Baqua FM, Bindroo BB (2010) Approaches to industrial exploitation of mulberry (mulberry sp.) fruits. J Fruit Ornam Plant Res. 18: 83-99.

Srivastava MP, Mehra R (2004) Diseases of minor tropical and sub-tropical fruits and their management. Diseases of Fruits and Vegetables: Diagnosis and Management, Volume II, S.A.M.H Naqvi, Kluwer, Spinger. pp: 559-632.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30: 1312-1313.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28: 2731-2739.

Vijayan K (2010) The emerging role of genomic tools in mulberry (*Morus*) genetic improvement. Tree Genet Genomes. 6: 613-625.

Wan H, Zhao Z, Malik AA, Qian C, Chen J (2010) Identification and characterization of potential NBS-encoding resistance genes and induction kinetics of a putative candidate gene associated with downy mildew resistance in *Cucumis*. BMC Plant Biol. 10: 186.

Yang S, Li J, Zhang X, Zhang Q, Huang J, Chen J-Q, Hartl DL, Tian D (2013) Rapidly evolving *R* genes in diverse grass species confer resistance to rice blast disease. Proc Natl Acad Sci. 110: 18572-18577.

Yang S, Zhang X, Yue J-X, Tian D, Chen J-Q (2008) Recent duplications dominate NBS-encoding gene expansion in two woody species. Mol Genet Genomics. 280: 187-198.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24: 1586-1591.

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17: 32-43.

Zambounis A, Elias M, Sterck L, Maumus F, Gachon CMM (2012) Highly dynamic exon shuffling in candidate pathogen receptors…What if brown algae were capable of adaptive immunity? Mol Biol Evol. 29: 1263-1276.

Zan Y, Ji Y, Zhang Y, Yang S, Song Y, Wang J (2013) Genome-wide identification, characterization and expression analysis of *Populous* leucine-rich repeat receptor-like protein kinase genes. BMC Genomics 14: 318.

Zhou T, Wang Y, Chen JQ, Araki H, Jing Z, Jiang K, Shen J, Tian D (2004) Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Genet Genomics. 271: 402-415.