Australian Journal of
Crop Science
AJCS

AJCS 10(9):1324-1329 (2016)
ISSN:1835-2707
DOI: 10.21475/ajcs.2016.10.09.p7825

Evidence of extensive positive selection acting on cherry (*Prunus avium* L.) resistance gene analogs (RGAs)

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

[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]Laboratory of Forest Genetics and Tree Breeding, Faculty of Agriculture, Forestry & Natural Environment, Aristotle University of Thessaloniki, P.O. Box 238, Thessaloniki GR54006, Greece
[3]Institute of Applied Biosciences, CERTH, Thermi, Thessaloniki, 570 01, Greece

***Corresponding authors: antbio@yahoo.gr; pmadesis@certh.gr**

Abstract

The cherry tree (*Prunus avium* L.), is an important tree species which is intensively plagued by many phytopathogenic fungal species. Resistance gene analogs (*RGA*s) are the largest class of resistance (*R*) genes and are pivotal components at breeding projects, serving as useful functional markers linked to *R* genes. In order to assess the evolutionary pressures acting upon *P. avium RGAs* candidates, their 173 homologues that have previously been deposited in GenBank were mined. Their proteins were clustered according to their blast(p) similarities in 12 MCL (Markov Cluster Algorithm) tribes, resulting in unique and well supported paralogous gene groups (PGGs). The extent to which these genes exhibit evidence of adaptive, positive selective pressures, which are causing excessive fixation of non-synonymous mutations, was determined using a series of maximum likelihood analyses using the PAML package. The results postulate existence of robust evidence of positive selection, acting in almost all of the clustered PGGs across their phylogenies. Furthermore, analyses revealed that the majority of the positively selected amino acid residues sites are localized widely across these *RGA*s sequences. We speculate that the clustered distribution of these *RGA*s might also be pronounced of high birth and death genes rates with diversifying episodes acting on their NB-ARC domains, putatively affecting their ligand-binding specificities. Such evolutionary insights shed light on how these NBS-encoding *RGA*s in *P. avium* are being evolved, assigning them as the foremost surveillance mechanism against rapidly evolving fungal pathogens, and providing breeders with effective tools for fast-tracking the development of varieties with more durable resistance.

Keywords: disease resistance; genomics-assisted breeding; non-synonymous nucleotide substitution; positive selection; *Prunus avium*; *RGA*s.
Abbreviations: LRR_leucine rich repeat; MCL_markov cluster algorithm; NBS_nucleotide-binding site domain; NGS_next generation sequencing; PAML_phylogenetic analysis by maximum likelihood; PAMP_pathogen associated molecular pattern; RaxML_randomized axelerated maximum likelihood; RGA_resistance gene analog; PGG_paralogous gene group; SNP_single nucleotide polymorphism.

Introduction

The cherry tree (*Prunus avium* L.), also known as wild cherry (natural populations) and sweet cherry (fruit orchards), is an outbreeding species belonging in the Rosaceae family, carrying a diploid genome (2n = 16) (Arumuganathan and Earle, 1991). This species is usually growing in temperate areas (Ganopoulos et al., 2013; Marti et al., 2012) and is one of the most important tree crops worldwide. Among the most destructive *P. avium* fungal diseases are the powdery mildew (*Podosphaera oxyacanthae*), brown rot (*Monilinia* spp.), leaf spot (*Blumeriella jaapii*) and Cytospora canker (*Leucostoma* spp.) (Kappel et al., 2012). Therefore, the development through breeding of fungal resistance cultivars is of pivotal significance towards the establishment of diseases management strategies (Ganopoulos et al., 2011), especially in trees where it is time consuming to breed for new characteristics and especially for phytopathogenic resistance.

Fungi recognition is the first crucial step of defense reactions in plants which is often mediated by a plethora of rapid-evolving receptors, many of which containing ligand-binding and signal transduction domains, like leucine rich repeats (LRRs) and NB-ARC domains, respectively. Employment of innovative molecular mapping technologies, besides traditional breeding practices, allows breeders to meet the challenge of developing resistant crops more accurately and at a faster pace (Lalli et al., 2005). Until the use of next generation sequencing (NGS) approaches, which have highly facilitated the rapid identification of resistance (*R*) genes, PCR-based approaches implementing degenerated primers were extensively applied in order to isolate such loci. Most of these approaches targeted the family of NBS (nucleotide-binding site) / LRRs-containing genes, the most highly expanded group of genes linked directly to *R* genes functions (Debener and Byrne, 2014). Currently, a huge number of such genes have been revealed in plant genomes, however without assigning an exact functional confirmation to individual genes (Dangl et al., 2013). The family of NBS / LRRs-containing genes mediates resistance to a wide range of phytopathenic fungi (Wan et al., 2010). Resistance gene analogs (*RGA*s) are a large class of potential *R*-genes and are a valuable resource for the discovery and development of molecular markers, enhancing numerous disease resistance breeding programs (Ameline-Torregrosa et al., 2008; Perazzolli et al., 2014). Moreover, LRR domains, across their expansion in plant species, commonly evolve novel ligand-binding specificities under diversifying selection *via* a number of different mechanisms, like point mutations at variable residues, variations in repeat numbers, tandem and segmental gene duplications, gene conversions, unequal crossing-over events and other rearrangements (Friedman and Baker, 2007; Meyers et al., 2003; Yang et al., 2008; Zambounis et al., 2012). Therefore, the evolution patterns of NBS / LRR-containing genes, as of their *RGAs* counterparts, appear to be a complex process (Sekhwal et al., 2015; Zhou et al., 2004). The evidence above underlies the basis of this survey, whose aim was to investigate if it is likely that positive selection may be employed as an evolutionary force, with signatures acting on NBS-LRR-encoding genes and particularly on their *RGAs* in *P. avium*. Furthermore, the location of these positively selected amino acid residues is crucial for obtaining novel gene functions (Mondragon-Palomino et al., 2002). Previous studies have reported that solvent-exposed regions of the LRRs domains or even repeats are under positive selection; reversibly such evidence has been interpreted as a sign of the involvement of these regions in pathogen recognition (Parniske et al., 1997; Zambounis at al., 2012). Therefore, the objective of this study was to gain insights of the evolutionary profiles of the *P. avium* NB-ARC / LRR-encoding *RGAs* genes, hypothesizing that successive episode of diversifying selection might contribute in the acquisition of novel pathogens recognition repertoires in *P. avium*. Overall, our findings suggest that *P. avium RGAs* in fact exhibit signs of such selection. These results could provide a critical foundation for the ongoing *P. avium* disease resistance breeding efforts in future.

Results and Discussion

RGAs abundance and phylogeny

Yield and fruit quality of *P. avium* crops can be substantially decreased by harmful fungal pathogens. In order to combat pathogens, plant species generally rely on their innate immunity mediated mainly *via* PAMPs at infection sites (Chisholm et al., 2006; Perazzolli et al., 2014). For example, pathogen diverse effectors are recognized by plant receptor proteins, which in turn activate the downstream defense reactions (Glazebrook, 2005; Sekhwal et al., 2015). The LRRs class of *RGA*s genes is comprehensively surveyed in terms of sequence evolution and genome distribution across plant genomes, and therefore selection upon them for durable disease resistance is a crucial component of nearly all plant breeding programs (Mace et al., 2014).

An abundant number of *RGA*s from *P. avium* was mined, although its genome is still not publicly available (https://genomics.wsu.edu/sweet-cherry-genome-project/). This is absolutely in line with *RGAs* common abundance and diversity in other plant genomes (Chen et al., 2010; Sekhwal et al., 2015, Zhou et al., 2004). We assume that the putative existence of *R* pseudogenes would suggest that the *P. avium RGAs* might be subjected to an ongoing birth and death evolution. In our dataset, all 173 *RGAs* genes from *P. avium* were sharing Pfam and InterProScan hits with the NB-ARC domain (PF00931.18) and with the LRR-containing proteins (PTHR23155), respectively. To gain insight into the origin of the expansion of the 173 *P. avium RGAs*, a RAxML-based phylogenetical analysis was conducted (Stamatakis, 2014) based on a Muscle alignment (Edgar, 2004) at the amino acid level. A significant portion of the phylogenetic clades contained only a limited number of *RGA*s, while the backbone topology of the tree was very well resolved. It contained 344 branches, reflecting a series of consecutive gene duplications up to the most recent bursts of duplication events at terminal branches (Fig 1). The overall pairwise identity among the 173 amino acid sequences was 54.9%, implying a rather high degree of divergence among them.

Assignment of PGGs

All-against-all blast(p) searches (E-value cutoff < 10^{-10}) were applied to assess the extent to which these 173 *RGAs* genes are grouped in MCL (Markov Cluster Algorithm) tribes (similarity cutoff of 50%), allowing their classification into respective paralogous genes groups (PGGs). All but one of these 173 *RGAs* candidates were clustered in 12 PGGs, ranging from two up to 37 per tribe. The majority (84%) of the *RGAs* genes were clustered in five different PGGs (PGG-1, PGG-2, PGG-6¸ PGG-7 and PGG-11). In contrast, three PGGs contained only two *RGAs* sequences (Table 1). Similar results were obtained when the SCPS (spectral clustering of proteins) software was employed and a hierarchical clustering parameter was applied (data not shown) (Nepusz et al., 2010). Hence, the robustness and accuracy of the *RGAs* clustering was confirmed. It was noted that, in some cases, *RGAs* which have been clustered in the same PGG, were almost identical in their amino acid composition, supporting the case of various duplication events in the genome (Zambounis et al., 2012). Our results indicate also that the topologies and compositions of the phylogenetic clades were consistent with those of the MCL-based *RGAs* clusters, indicative of the accuracy of both analyses.

Signatures of positive selection acting across PGGs lineages

Positive selection is defined as “the rapid fixation of beneficial non-synonymous mutations” and is an important driving force for the evolution of a number of paralogous genes towards their functional diversification which may be selectively favored in plants under biotic pressures (Delph and Kelly, 2014). There are two modes of positive selection pressures in genes level: the directional and the diversifying selection (Zhang and Rosenberg, 2002). The first selective mode promotes a substantial functional change, often resulting in an altered biochemical activity. In opposite, diversifying selection promotes the rapid alteration of genes sequences in different alleles or species by a non-directional mode, which often increases the reservoir of the ligands that can be recognized under fungal attacks, but rarely alters the main protein functions. Therefore, diversifying selection is evident mainly in plant host defense genes families (Delph and Kelly, 2014). Besides, balancing and positive selection are likely to act simultaneously, creating diversifying selection (Miyake et al., 2009). In contrast, human-based selection is in general a purifying selection (net negative effect on genes diversity) rather than diversifying, because sweeps the genome due to genetic drift effects by means of

Table 1. PGGs under positive selective pressures using the counting YN00 and the CODEML methods. Nine PGGs were tested for evolutionary signatures, containing more than two clustered (according to the MCL-based approach) *RGA*s sequences. The numbers of the *RGA*s sequences of each PGG are also indicated.

PGGs	*Number of RGAs* sequences	Signs of positive selective pressures (YN00 method [1])	Positive selection acting in branches (CODEML method)	Positive PGGs (both methods [2])
PGG-1	19	1.34, ∞	YES	YES
PGG-2	37	3.10, ∞	YES	YES
PGG-3	3	1.20	YES	YES
PGG-4	2		NT	
PGG-5	2		NT	
PGG-6	34	3.43, ∞	YES	YES
PGG-7	24	2.37, ∞	YES	YES
PGG-8	8	1.29, ∞	YES	YES
PGG-9	2		NT	
PGG-10	3		YES	NO
PGG-11	32	9.71, ∞	YES	YES
PGG-12	6	∞	YES	YES

[1] The numbers indicate the highest omega values among all *RGA*s pairwise comparisons for each PGG; ∞ refers to pairwise comparisons in which dN > 0 and dS = 0.

[2] YES indicate PGGs under positive selection using both the YN00 and CODEML methods.

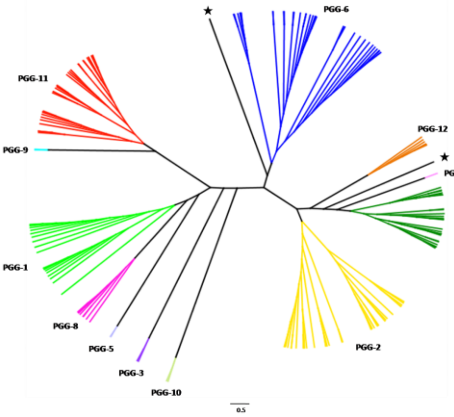

Fig 1. RAxML phylogeny of the *RGAs* amino acid sequences, which were aligned using the Muscle program. Phylogenetical clades in different colors are corresponding to *RGA*s sequences which were assigned to the 12 distinct PGGs using the MCL-based clustering approach. With asterisks are indicated the *RGA*s sequences which did not being clustered at all using the above approach.

Table 2. Statistical parameters as being calculated with CODEMLSITES program among the positively selected sites for each PGG. The total numbers of positively selected sites (amino acid residues) as well their types and exact positions in the relevant alignments are indicated.

PGGs	ℓ (M7) [1]	ℓ (M8) [2]	LRT [3]	Statistical significance (*P*)	Number of sites [4]	Numbers of positively selected sites	Positions and types of positively selected sites [5]
PGG-1	-742.31	-740.34	3.94	<0.5	160	0	
PGG-2	-2660.73	-2655.22	11.02	<0.01	157	1	**8 C**
PGG-3	-711.09	-706.19	9.8	<0.01	170	1	**61 I**
PGG-4	NT						
PGG-5	NT						
PGG-6	-1264.43	-1255.66	17.54	<0.001	156	2	31 S, **153 V**
PGG-7	-2644.30	-2630.08	15.22	<0.001	155	14	**8 K**, 12 R, 27 K, 39 P, 40 A, 54 S, **67 C**, 68 Q, 72 D, 98 R, 104 D, **107 F**, 120 C, **141 E**
PGG-8	-1395.25	-1383.55	23.4	<0.001	160	16	**11 I, 13 N**, 25 K, **28 S, 38 V, 40 T**, 43 D, **50 K, 56 A, 63 Q, 85 F**, 87 S, **113 P**, 133 R, **144C, 155 V**
PGG-9	NT						
PGG-10	-651.96	-645.12	13.68	<0.001	158	4	37 V, 42 R, 111 M, 130 S
PGG-11	3374.73	-3356.37	36.72	<0.001	248	21	9 P, 10 R, **13 P,** 14 M, **27 P**, **45 G**, 46 M, 53 Y, 55 A, **57 E,** 60 C, 61 N, 107 L, **123 K, 143 V, 156 A,** 167 D, 169 N, 175 P, **183 I**, 184 E
PGG-12	-898.25	-890.05	16.4	<0.001	160	3	79 L, **126 Y, 145 D**

[1,2] Log-likelihood values for the two evolutionary scenarios using models M7 and M8.

[3] Likelihood ratio test (LRT) among M8 and M7 models.

[4] These numbers are referring to the amino acid lengths of the respective PGGs alignments feeding the CODEMLSITES program.

[5] Positions are accordingly to PGGs alignments; Types of positively selected sites (amino acid residues) with posterior Bayesian probabilities greater than 0.95 are shown in lightface, whilst greater than 0.99 values are shown in boldface.

96iu34removing rare and randomly occurred alleles which support diversification (Delph and Kelly, 2014; Mace et al., 2014).

In our study, we investigated whether accelerated evolution and signatures of positive selection might have also contributed to the divergence of these *RGAs* genes in *P. avium*. Analyses were performed separately in each of the nine PGGs consisting from more than two *RGA*s (Tables 1, 2). Initially, the average codon-based evolutionary divergences were calculated over all sequence pairs, using the MEGA 5 software (Tamura et al., 2011). In all instances, the average numbers of synonymous (dS) mutations per synonymous sites over all sequence pairs were below the value of "2". This is a crucial cut-off value above which the corresponding sequences would have to be excluded from further analyses, in order to bypass any saturation effects based on nucleotide substitutions (Yang and Nielsen, 2000).

Subsequently, all nine PGGs were subjected to several tests of positive selection using the approximate counting (YN00) and the ML (CODEML / CODEMLSITES) programs implemented in the PAML version 4.8 package (Yang 2007). Using the YN00 method, non-synonymous (dN) and dS mutation values across the entire *RGA*s sequences were calculated, with observed omega (ω) values being significantly greater than "1" in eight PGGs (Table 1). Thus, highly ω values were revealed, such as 9.71 for PGG-11 and 3.43 for PGG-6, among all *RGA*s pairwise comparisons for each PGG (Table 1 and Supp Table 1). It was considered that this approach provided sufficient evidence that extensive signatures of positive selection are present across the *P. avium* PGGs. Additionally, we applied the CODEML (Yang, 2007) program to identify PGGs under positive selection and to validate the results obtained using the YN00 counting method. Statistically significant evidence of positive selection (ω values higher then 1) were detected in numerous branches of the PGGs evolutionary trees in all nine PGGs datasets (Table 1). These selective signs were spreading across the entire tree lineages acting with a non-general rule, as positive selection episodes were evident both in the terminal branches (reflecting a series of recently successive bursts of gene duplications) and in the ancient ones. Therefore, CODEML analysis allowed us to hypothesize that recent episodes of positive selection have occurred, overlapping similar more ancient events among these *RGA*s. Other studies have previously addressed the existence of positive selection driving the rapid diversification of plant *RGAs* and their *R* genes counterparts such as in *Arabidopsis* (Chen at al., 2010), in annual, perennial ryegrass (Li et al., 2006) and in *Malus domestica* (Perazzolli et al., 2014). In parallel, Khan et al. (2015) in order to address the evolution of NBS-LRR-encoding and *RGA*s genes within *Gossypium hirsutum* derived from *G. arboretum* and *G. raimondii*, postulated that the evolution of these sequences occurred by continuous accumulation of mutants that led to positive selection, after separating from its diploid parents and simultaneously altering its susceptibly to fungal diseases as compared to donor ancestors. Recently, Mace et al., (2014) demonstrated that the evolutionary plasticity of NBS-encoding resistance genes in sorghum is driven by multiple and contrasting processes through both natural and human-mediated selection, including purifying selection towards fixation of beneficial alleles and selective removal of deleterious ones, positive directional selection, and balancing selection, in which multiple alleles were remained at intermediate frequencies both in wild (ancestral) and cultivated (descendant) populations. In the same study, the reported omega values were consistently diminished for non- NBS encoding genes throughout the genome, either under selection or neutral expectations, whilst nucleotide diversity in the wild genotypes were constantly higher against the cultivated ones. In our case, we assume that the selective pressures at *RGA*s in *P. avium*, with signs of positive selection pressures acting upon them as they were found, are associated mainly with natural selection being influenced by tree species longevity in the face of continual selective demand for fungal disease resistance and less with human-mediated genotypes improvement by breeding efforts for selective traits. Recently, in five natural wild cherry populations in Greece there were detected evolutionary signatures under negative frequency-dependent selection, a case of balancing selection, acting at the self-incompatibility (S) locus promoting a higher within population genetic diversity (Ganopoulos et al., 2012).

Positive selective pressures acting among amino acid residues across the RGAs

In order to test if there is positive selection at individual amino acid residues level, the CODEMLSITES program was employed (Table 2). CODEMLSITES compared various models (M0 against M3, M1 against M2, and finally M7 against the stringiest M8) for all PGG dataset alignments (Yang, 2007). All three models (M2, M3 and M8) allowing for selection, were significantly supported over the other models (Table 2). Extensive signs of adaptive selection (P<0.01, P<0.001) were found in eight out of the nine PGGs, acting rather widely upon amino acid residues across the respective tree branches (Table 2). Only Pa_RGA-1 did not reflect any evidence of positive selection, since the divergence among the 19 *RGAs* comprising this PGG was quite low (only 20 SNPs along the 483 alignment). The 62 in total positively selected sites (with ω values up to 10), along with the relevant CODEMLSITES statistics, are depicted in Table 2. These sites were selected with high posterior probabilities using the empirical Bayesian analysis and after the implementation of M8 model. Recently, comparative analyses of trees regarding *R* genes from Rosaceae species have revealed that solvent-exposed residues of the LRRs domains are hyper-variable, with intensive diversifying selective pressures acting on them (Perazzolli et al., 2014). Moreover, such evidence of positive selection is consistent with host-pathogen co-evolution processes leading to acquisition of novel resistance functional specificities (Perazzolli et al., 2014; Zambounis et al., 2012). In turn, these findings would, for instance, facilitate the development of informative *P. avium* RGAs-derived sequence characterized amplified region (SCAR) markers, based both on *RGAs* genes with extensive positive selection acting upon them, and on the localization of the positively selected residues. Such markers would be particularly useful for identifying fungal resistant *P. avium* genotypes in routine marker-assisted selection (MAS) breeding programs using segregating progenies.

Materials and Methods

RGAs mining, structural and functional evaluation

Based on their NB-ARC domain profile (PF00931.18/ CL0023), we downloaded from the NCBI protein database by keyword searches, all 173 *P. avium RGAs* partial proteins. Using these amino acid sequence IDs, we have also retrieved by a python script the respective ORF nucleotide sequences. All amino acid sequences had to fulfill the following criteria in order to be evaluated for evolutionary acting signatures and profiles: (a) existence of a valid structure in means of protein length up to 100 residues, positions of LRRs or NB-ARC domain motifs, blast(p) hits against known plant resistance genes, and (b) confirmation of functional domains predictions using the Pfam database (http://pfam.sanger.ac.uk/search#tabview=tab1) and InterProScan 5 software package (Jones et al., 2014).

Phylogenetics of the RGAs and clustering assignment of PGGs

The phylogenetic relationships among the 173 *P. avium RGAs* amino acid sequences were revealed by performing a Muscle alignment (Edgar, 2004) and a tree reconstruction, using the RAxML program (Stamatakis, 2014) with a gamma model of rate heterogeneity and an ML estimate of the alpha parameter. All the above analyses were performed using the Geneious R7 platform (Kearse at al., 2012). At a next step, a MCL-based tribing approach was applied. Unique PGGs, technically termed as MCL tribes, were identified by formatting the amino acid *RGAs* dataset, performing all-against-all blast(p) searches and by feeding the similarity results to the MCL algorithm, using an inflation value of "2". The validity of these, 12 in total, PGGs was checked, using three criteria: (a) overall similarity throughout the majority of the paralogous coding sequences; (b) no or few gaps across the aligned sequences in each PGG; and (c) at least 50% amino acid identity between the paralogous in each PGG. The application of these criteria resulted in robust and reliable alignments, eliminating the impact of drawbacks for the downstream positive selection analyses, such as gap-induced misalignments and elevating divergence among aligned amino acid sequences.

Evolutionary analyses using maximum likelihood approaches

Only nine out of 12 PGGs, counting up to three *RGAs*, (Table 1), were assigned separately for positive selection signatures, using the YN00, CODEML and CODEMLSITES programs from the PAML version 4.8 package (Yang, 2007). All these tests compare the rates of dS and dN mutations, both among the nucleotide codon sequences and along the branch lineages in the respective evolutionary trees. The substitution saturation effects were estimated by calculating dS rates between the aligned nucleotide sequences using an altered Nei-Gojobori (1986) method presented by Yang and Nielsen, (2000) and inferred in MEGA 5 software (Tamura et al., 2011).

Initially, the rates of dN/dS nucleotide substitutions per site across all possible pairwise comparisons within each of the nine PGGs were calculated using the approximate counting methods (Yang and Nielsen, 2000) implemented in the YN00 program, using the PAML version 4.8 package (Yang, 2007). The total dN, dS, and ω values for each of the nine PGGs pairwise comparisons are shown in Supp Table 1.

NJ trees and nucleotide alignments were feeding the CODEML and CODEMLSITES programs according to Lynn et al., (2004a) and Lynn et al., (2004b); whereas, the methods for these analyses were used according to Zambounis et al., (2012). Amino acid sequence alignments were performed using the MUSCLE program (Edgar 2004), whilst the construction of phylogenetic trees was performed using the MEGA 5 software (Tamura et al., 2011). CODEML was tested separately in all nine PGG *RGAs* datasets for estimating the variable selective pressures acting among lineages in their phylogenies, whilst CODEMLSITES was employed to test for site-specific codon substitution models allowing detection of selective signatures among amino acid residues (Yang, 2007). In both programs, log-likelihood calculations were performed for each model and comparisons between each other were conducted by likelihood ratio tests (LRTs). The empirical Bayesian method further allowed the detection of codon residues subjected to positive selection by posterior probabilities. All data sets, including alignments and outputs of the YN00, CODEML and CODEMLSITES programs are available upon request.

Conclusion

RGAs in *P. avium*, could potentially be useful *R* genes - linked candidates. These *RGAs* to be applied to evolutionary bioinformatics approaches were found to share conserved domains and structural features. These genes were found to occur in similarity clusters, implying possible frequent tandem duplication and ectopic translocation events in the genome. This could also explain the extensive and intensive episodes of positive selective pressures acting across their lineages and amino acid residues. In principle, this reservoir of genetic variation could drive the evolution of novel *R* genes specificities in *P. avium*. It is known that typically, numerous selective episodes happen from the birth of a paralogous locus by a duplication event at an ancestral locus. This is in accordance with our findings, as positively selected *RGA* branches among the tested PGGs were also ancestral, besides the terminal ones (pointing of a more recent positive selection acting upon them). Overall, the existence of positive selection acting on *RGAs* is in coincidence with several other surveys. Furthermore, numerous positively selected amino acid residues were found to be distributed rather equally across the *RGA* sequences. These residues might have originally conferred specificity to a hypothetical ligand, which was being apparently altered repeatedly to provide novel binding functions. These results could be exploited in the *P. avium RGAs*-based ongoing breeding programs. Finally, once the genome is publicly available, we plan to systematically annotate all *RGAs* and *R* genes in *P. avium*.

Acknowledgments

The co-authors have not a conflict of interest to declare. 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

Ameline-Torregrosa C, Wang B-B, O'Bleness MS, Deshpande S, Zhu H, Roe B, Young ND, Cannon SB (2008) Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. Plant Physiol. 146: 5-21.

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep. 9: 208-218.

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.

Chisholm ST, Coaker G, Day B, Staskawicz BJ (2006) Host-microbe interactions: shaping the evolution of the plant immune response. Cell. 124: 803-814.

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

Debener T, Byrne DH (2014) Disease resistance breeding in rose: current status and potential of biotechnological tools. Plant Sci. 228: 107-117.

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

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

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

Ganopoulos I, Tsaballa A, Xanthopoulou A, Madesis P, Tsaftaris A (2013) Sweet cherry cultivar identification by high-resolution-melting (HRM) analysis using gene-based SNP markers. Plant Mol Biol Rep. 31: 763-768.

Ganopoulos I, Aravanopoulos F, Argiriou A, Tsaftaris A (2012) Genome and population dynamics under selection and neutrality: an example of *S*-allele diversity in wild cherry (*Prunus avium* L.) Tree Genet Genomes. 8: 1181-1190.

Ganopoulos IV, Kazantzis K, Chatzicharisis I, Karayiannis I, Tsaftaris AS (2011) Genetic diversity, structure and fruit trait associations in Greek sweet cherry cultivars using microsatellite based (SSR/ISSR) and morpho-physiological markers. Euphytica. 181: 237-251.

Glazebrook J (2005) Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. Annu Rev Phytopathol. 43: 205-227.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics. 30: 1236-1240.

Kappel F, Granger A, Hrotkó K, Schuster M (2012) Cherry. In: Fruit Breeding. USA, Springer. 459-504.

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.

Li J, Xu Y, Fei S, Li L (2006) Isolation, characterization and evolutionary analysis of resistance gene analogs in annual ryegrass, perennial ryegrass and their hybrid. Physiol Plantarum. 126: 627-638.

Lynn DJ, Higgs R, Gaines S, Tierney J, James T, Lloyd AT, Fares MA, Mulcahy G, O'Farrelly C (2004a) Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken. Immunogenetics. 56: 170-177.

Lynn DJ, Lloyd AT, Fares MA, O'Farrelly C (2004b) Evidence of positively selected sites in mammalian α-defensins. Mol Biol Evol. 21: 819-827.

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.

Marti AF, Athanson B, Koepke T, Forcada CF, Dhingra A, Oraguzie N (2012) Genetic diversity and relatedness of sweet cherry (*Prunus avium* L.) cultivars based on single nucleotide polymorphic markers. Front Plant Sci. 3: 116.

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.

Miyake T, Takebayashi N, Wolf DE (2009) Possible diversifying selection in the imprinted gene, MEDEA, in *Arabidopsis*. Mol Biol Evol. 26: 843-857.

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.

Nepusz T, Sasidharan R, Paccanaro A (2010) SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. BMC Bioinformatics. 11: 120.

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.

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.

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.

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, 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.

Zhang J, Rosenberg HF (2002) Diversifying selection of the tumor-growth promoter angiogenin in primate evolution. Mol Biol Evol. 19: 438-445.

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.