Mol Breeding (2014) 33:227–233
DOI 10.1007/s11032-013-9928-7

SHORT COMMUNICATION

Genome-wide identification and analysis of candidate genes for disease resistance in tomato

G. Andolfo · W. Sanseverino · R. Aversano · L. Frusciante · M. R. Ercolano

Received: 18 January 2013 / Accepted: 13 July 2013 / Published online: 30 July 2013

Abstract Tomato (*Solanum lycopersicum* L.) has served as an important model system for studying the genetics and molecular basis of resistance mechanisms in plants. An unprecedented challenge is now to capitalize on the genetic and genomic achievements obtained in this species. In this study, we show that information on the tomato genome can be used predictively to link resistance function with specific sequences. An integrated genomic approach for identifying new resistance (R) gene candidates was developed. An R gene functional map was created by co-localization of candidate pathogen recognition genes and anchoring molecular markers associated with resistance phenotypes. In-depth characterization of the identified pathogen recognition genes was performed. Finally, in order to highlight expressed pathogen recognition genes and to provide a first step in validation, the tomato transcriptome was explored and basic molecular analyses were conducted. Such methodology can help to better direct positional cloning, reducing the amount of effort required to identify a functional gene. The resulting candidate loci selected are available for exploiting their specific function.

Keywords *Solanum lycopersicum* · R loci · Physical map · Predicted proteins · Genomic approach

Electronic supplementary material The online version of this article (doi:10.1007/s11032-013-9928-7) contains supplementary material, which is available to authorized users.

G. Andolfo · W. Sanseverino · R. Aversano ·
L. Frusciante · M. R. Ercolano (✉)
Department of Agricultural Sciences, University of Naples 'Federico II', Via Università 100, 80055 Portici, Italy
e-mail: ercolano@unina.it

Introduction

Tomato (*Solanum lycopersicum* L.) is one of the most important horticultural crops worldwide. It is subject to numerous pathogen attacks that can significantly reduce yields. Many tomato breeding projects have aimed to introduce resistance genes through classical and molecular genetic approaches (Foolad 2007). The immune response governed by resistance (R) genes has been investigated in depth in this species, contributing to elucidating important R gene molecular and genetic mechanisms in plants (Ercolano et al. 2012). Comprehensive knowledge of genomic R loci architecture in this species could help explain gene arrangement and diversification, as well as design a new genomic breeding strategy.

The recent sequencing of the tomato genome (Tomato Genome Consortium 2012) would appear very useful for improving the identification of disease resistance genes or genomic regions harboring them. Recently, a 769 pathogen recognition gene tomato dataset was categorized according to the presence and order of protein domains, phylogenetic analysis and physical arrangement within the genome (Andolfo et al. 2013).

Currently, there is tremendous interest in the advanced use of genome-wide data for identifying new resistance genes. In order to speed up R loci tagging and pathogen recognition gene identification, several strategies have been explored (Pan et al. 2000; Riely and Martin 2001; Caicedo and Schaal 2004; Mazourek et al. 2009; McHale et al. 2006). Genomic approaches can enhance the identification of genes that encode for resistance traits. After a genomic interval underlying a disease resistance trait has been identified, there are various possibilities for tracking down the gene responsible. Traditional approaches can be extremely costly, tedious, and time-intensive, given the difficulty of marker development and the size and complexity of R gene clusters (McDowell and Simon 2008). Annotation data and genetic map information represent an invaluable resource for performing this task. A better understanding of tomato R gene genomic architecture could streamline cloning efforts.

In this study, we identified strong pathogen recognition gene candidates linking predicted pathogen recognition proteins with previously mapped R loci, characterized in detail the identified pathogen recognition genes, highlighting peculiar pathogen recognition domain arrangements, and finally provided molecular validation of our predictions, both exploring the tomato transcriptome and performing experimental validation. Puzzling information were collected and combined in order to obtain a synergy between different approaches. Our strategy was constructed to reduce the time required for R gene identification and to make easier their cloning, a critical step towards modern genome breeding. In many cases, a predicted protein was narrowed down to a small region, allowing the identification of one or few candidates, now available for exploiting their specific function. We believe our attempt captured fundamental aspects of data integration contributing to pinpointing key steps in genetic, genomic, and phenotypic data synthesis for a better R gene isolation.

Results and discussion

Constructing the physical map of tomato R genes

In tomato, several resistance phenotypes have been genetically mapped, delineating genome regions harboring causative resistance loci. Examples include the *Py1* gene for resistance to corky root rot (Doganlar et al. 2002) on chromosome 3, the *Pto* gene conferring resistance to *Pseudomonas syringae* resistance (Martin et al. 1991) on chromosome 5, the root-knot nematode resistance locus *Mi* (Kaloshian et al. 1998), the *Ty1* gene for tomato yellow leaf curl virus resistance (Hanson et al. 2000) on chromosome 6, and the *Sw5* gene for tomato spotted wilt virus resistance (Stevens et al. 1995) on chromosome 9. While some of these tomato disease resistance genes have been cloned using genetic map-based methods, many more have been mapped but not cloned to date (Foolad 2007; Ercolano et al. 2012). An integrated genomic approach could help to find a specific function for predicted genes. For this reason we built up a detailed R loci physical map, based on identifying predicted pathogen recognition genes located in the proximity of marker sequences associated with previously identified R loci (Andolfo et al. 2013). By performing a literature search we selected a pool of 82 markers flanking the target loci (R genes and quantitative trait loci). The markers used were chosen on the basis of a very important factor, namely, their close association with genes not yet cloned (Supplemental Table S1). Once the markers were placed on chromosomes, we were able to select putative pathogen recognition genes, which permitted us to discriminate those which fell between at least one pair of markers, allowing us to focus on a restricted set of genes. Out of 769 predicted pathogen recognition protein sequences, about 368 corresponding genes (48 %) were localized among markers linked with functionally defined and mapped, but uncloned, R genes.

The functional R gene map based on the co-localization of pathogen recognition putative proteins with R loci linked markers is shown in Fig. 1. The map shows that the 368 candidate pathogen recognition genes are distributed on 12 chromosomes. The visualization of information linked to a locus is a fundamental step in interpreting data and in suggesting correlations between genetic and genomic data, even if markers delimit regions that can include a variable number of genes (ranging from 1 to 50) which can belong to the same class of pathogen recognition genes or otherwise. Indeed, several markers co-localize with a large group of putative pathogen recognition genes, making identification of individual candidate genes difficult. Moreover, in some cases, the chromosomal

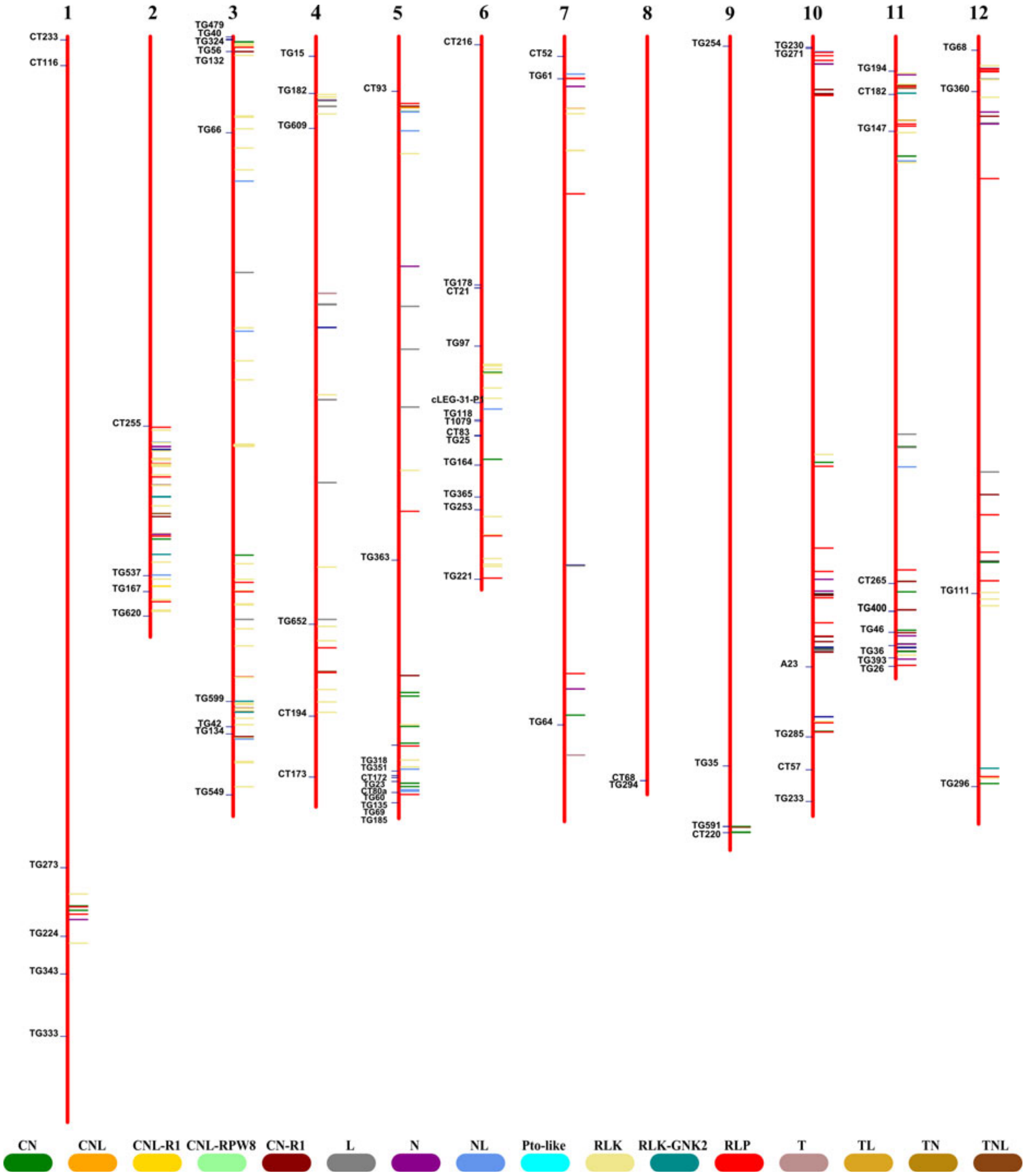

Fig. 1 Overview of predicted pathogen recognition genes localized among markers linked with functionally established R loci. The *color* used for each gene indicates the structural class to which the encoded protein has been assigned. *RLP* receptor-like protein, *RLK* receptor-like kinase, *CNL* (coiled coil/ nucleotide-binding site/leucine-rich repeat) protein, *TNL* Toll interleukin resistance/nucleotide-binding site/leucine-rich repeat, *Gnk2* ginkbilobin-2, *RPW8* domain was identified in two proteins isolated in *A. thaliana* that confer resistance against a broad range of powdery mildew races; *R1* domain characteristic of the R1 protein, *PTO-like* genes encoding the typical serine threonine domain characteristic of the Pto protein

regions bounded by markers of different R loci overlapped, delimiting a common chromosome area, as in the case of markers Lb3, EB7 and Xv-4.

Characterization of putative tomato pathogen recognition genes

In order to better classify and label tomato proteins in the functional map, phylogenetic investigations were conducted. Three separate phylogenetic trees were produced, designated the Nucleotide-Binding Site (NBS) group, extracellular Leucine-Rich Repeat-Serine/Threonine (eLRR-Ser/Thr) group, and Kinase (KIN) group (Supplemental Fig. S1). These analyses were useful for obtaining additional information on chromosome regions under study for a specific resistance trait. We defined as a genomic region-specific sub-clade a phylogenetic clade containing at least five sequences situated on the same genomic region with bootstrap support greater than 70 %.

The NBS phylogenetic tree shown in Supplemental Fig. S1, containing 86 tomato predicted proteins and 23 reference proteins involved in the resistance process, allowed us to identify orthologs to functional proteins and to detect three interesting genomic region sub-clades. Proteins grouped in the region-specific sub-clade on chromosome 11 showed an average identity of 56 % and those on chromosome 5 an average identity of 32 and 67 %. Interestingly, the Toll-Interleukin-Resistance/Nucleotide-Binding Site (TIR-NBS) clade included a protein with a Resistance to Powdery Mildew (*RPW8*) domain at the N terminal. This domain was identified in two proteins isolated in *Arabidopsis thaliana* that confer resistance against a broad range of powdery mildew races (Xiao et al. 2001). The association of the *RPW8* domain with NBS-LRR domains could help to shed light on the mechanism of action of both *RPW8* and genes of similar architecture in the *Solanum* species.

The eLRR-Ser/Thr group comprised 83 sequences and 15 reference proteins, including receptor-like proteins (RLP), involved in defense as well as in the development process. Phylogenetic analysis highlighted sub-clades that identified specific chromosomal regions with potential candidate genes for resistance. In particular, two sub-clades that included proteins located on chromosomes 12 and 7 were identified (Supplemental Fig. S1).

The KIN group contained 143 predicted sequences and 12 reference proteins. The relative phylogenetic tree in Supplemental Fig. S1 revealed two interesting super-clades (Pto and Gnk2 superclades). The first super-clade included proteins with a kinase domain similar to that of the protein Pto, while the second included proteins that possess a Ginkbilobin-2 (Gnk2) domain. The serine/threonine kinase Pto protein confers immunity to *Pseudomonas syringae pv. tomato* (Pedley and Martin 2003) and its overexpression has been shown to confer broad resistance (Tang et al. 1999). The Gnk2 is an antifungal protein found in the endosperm of *Ginkgo* seeds, which inhibits the growth of phytopathogenic fungi such as *Fusarium oxysporum* (Sawano et al. 2007; Miyakawa et al. 2009).

To date, all cloned tomato pathogen recognition genes with known resistance have been found to exist in gene clusters or arrays within the genome (Andolfo et al. 2013). On looking at the genome distribution of pathogen recognition genes linked to R markers, 224 genes were identified (about 60 %) that reside either in a gene cluster or in an array of 2–3 genes. Of these, some resistance loci, inherited in Mendelian fashion, were analyzed in greater detail to perform initial screening of potential resistance genes. Table 1 reports seven loci containing potential pathogen recognition genes and Supplemental Table S2 shows the list of gene IDs of all the candidate genes. The markers linked to *Xv4* discriminated a region on chromosome 3 that comprised nine genes, including a receptor-like kinase (RLK) protein with an extracellular Gnk2 domain. On chromosome 6 the markers of *Ol1* identified a single coiled coil/nucleotide-binding site/leucine-rich repeat (CNL) gene that has a peculiar domain (PTHR23155: SF94-Panther). Eight genes were located on chromosome 6 between the markers of *Ty1*, and a *Pto*-like gene was also included between markers of *Ty3*. On chromosome 9 the markers linked to *Ph3* included three CNL and one Toll interleukin resistance/nucleotide-binding site/leucine-rich repeat (TNL) protein. The TNL sequence showed a tyrosine-protein kinase active site (IPR008266-Prosite) corresponding to the LRR domain. The markers linked to the *Ty2* locus allowed 14 genes to be discriminated, located on chromosome 11, including genes belonging to cluster *I2*. Sixteen genes were located on chromosome 12 between the markers of *Lv*. This analysis allowed us to fine-tune the search for candidate genes.

Table 1 Selected resistance loci inherited in Mendelian fashion harbouring predicted pathogen recognition genes, showing the locus name, the pathogen to which it gives resistance and the number of candidate genes identified subdivided by class

Locus name	Pathogen	Reference	No. candidate genes/protein class				
			CNL[a]	RLK	RLP	TNL	Unknown
Lv	*Leveillula taurica*	Foolad (2007)	1	2	11	–	2
Ty2	Tomato yellow leaf curl virus	Foolad (2007)	3	1	1	–	8
Ph3	*Phytophthora infestans*	Foolad (2007)	3	–	–	1	–
Ol1	*Oidium lycopersicum*	Foolad (2007)	1	–	–	–	–
Ty3	Tomato yellow leaf curl virus	Foolad (2007)	–	–	–	–	1
Ty1	Tomato yellow leaf curl virus	Foolad (2007)	1	6	–	–	1
Xv4	*Xanthomonas campestris pv. vesicatoria* Race T3	Foolad (2007)	–	7	1	–	1

[a] *CNL* coiled coil/nucleotide-binding site/leucine-rich repeat, *RLK* receptor-like kinase, *RLP* receptor-like protein, *TNL* toll interleukin resistance/nucleotide-binding site/leucine-rich repeat, *Unknown* genes that encode novel domain associations or single domains

Putative pathogen recognition gene functionality tests

A first prerequisite for testing the functionality of a gene is the identification of its transcript. In order to ascertain that the predicted genes derive from functional sequences, we categorized the expressed tomato predicted pathogen recognition sequences close to R loci. Table 2 reports the number of expressed predicted pathogen recognition genes co-localizing with R loci for each chromosome and the number of expressed genes falling in a cluster or array. On average, 80 % of the genes examined proved to have a transcript in the tomato genome ranging from 63 % (chromosome 10) to 100 % (chromosomes 3, 6 and 9, Supplementary Table S3). Of them, 197 genes are located in clusters or arrays that might have a resistance function.

To verify that the predicted genes were actually present in tomato and there were neither inaccuracies of the predictor, nor prediction-distorted inaccuracies related to alignment defects, molecular analysis was carried out. All the genes tested were found in the tomato genome. Out of 37 gene sequences tested, 34 were shown to be also transcribed. Supplemental Table S4 reports the genomic ID for each gene tested, the chromosomal location, the expected amplicon length and the class to which belongs. Sequence identity scores found between SL2.40 reference sequences and those of 10 selected DNA amplicons obtained in our experiment is reported in Supplemental Table S5.

Materials and methods

Physical location

To construct the physical maps, the predicted pathogen recognition genes were collected in an SQL database and catalogued with the information on their characteristics and their location. A custom PERL script, connecting the database and transforming the information into vector graphics (SVG) images, was written to design each chromosome. The sequences of the 82 chromosome markers linked with R genes not yet cloned and reported by Foolad (2007) were taken from the SGN database (Supplemental Table S1).

Phylogenetic analysis

Evolutionary analyses were conducted using MEGA5 (Tamura et al. 2011). The phylogenetic relationships of predicted pathogen recognition proteins were inferred separately (e.g., NBS, eLRR-Ser/Thr and KIN groups) using the maximum likelihood method based on the WAG model (Whelan and Goldman 2001). The bootstrap consensus tree, inferred from 100 replicates, was taken to represent the evolutionary history of the sequences analyzed (Felsenstein 1985). All the amino acid sequences were aligned using MUSCLE 3.6 (Edgar 2004).

Validation of prediction results

The expression data of *S. lycopersicum* variety HEINZ 1706 were obtained from the Tomato Genome

Table 2 Results on predicted pathogen recognition genes co-localizing with R loci for each chromosome, the number of total and expressed pathogen recognition genes, as well as the number of expressed genes falling in the clusters or arrays, is shown

Chromosome	Predicted pathogen receptor genes (no.)	Expressed pathogen receptor genes (no.)	Identified cluster or array (no.)	Expressed gene in cluster or array (no.)[a]
1	7	6	1	2
2	54	54	8	29
3	55	55	11	28
4	31	25	8	20
5	38	32	6	19
6	16	16	4	8
7	25	22	7	21
9	4	4	1	4
10	46	28	11	15
11	50	43	13	35
12	40	39	8	25

[a] The number reported is calculated from the number of total genes identified in cluster/arrays for each chromosome

Consortium (2012) and used to extract a pathogen recognition gene expressed dataset. A pool of 37 predicted R genes was used to perform molecular validation. DNA and RNA were extracted from leaf tissue of genotype *S. lycopersicum* variety HEINZ 1706 using DNeasy and RNeasy Plant mini kits (Qiagen, Valencia, CA, USA), respectively. PCR was executed with 25 ng of genomic or complementary DNA, 10 pmol primers, 1 U of *Taq* DNA polymerase Kit (Invitrogen, Carlsbad, CA, USA), 10 pmol dNTPs, and 2 mM $MgCl_2$ in 25 µl reaction volumes. Amplification was performed using the following cycling conditions: 1 min at 94 °C, followed by 30 cycles of 1 min at 94 °C, 1 min 30 s at 60 °C and 2 min at 72 °C, with a final extension for 7 min at 72 °C. Amplicons were separated by electrophoresis on agarose gel (1.5 %), and photographed by a GelDoc apparatus. Primers were designed with Primer3 (http://frodo.wi.mit.edu), with a length between 18 and 27 bp. The length of the amplified fragments ranged from 300 to 1,000 bp, and the T_m of the specific primers was 59 °C for all pairs of primers (Online Resource S4). Amplicons were sequenced using the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) and run on automated DNA sequencers (ABI PRISM 3100 DNA Sequencer, Applied Biosystems). Sequence data deriving from SL2.40 reference were aligned with corresponding sequences originated from amplicons using MUSCLE 3.6 (Edgar 2004).

Acknowledgments We thank Dr Mark Walters for English language editing of the manuscript. This research was carried out within the GenoPOM-PRO Project funded by Italian Ministry of Education, University and Research.

References

Andolfo G, Sanseverino W, Rombauts S, Van der Peer J, Bradeen JM, Carputo D, Frusciante L, Ercolano MR (2013) Overview of tomato (*Solanum lycopersicum*) candidate pathogen recognition genes reveals important *Solanum* R locus dynamics. New Phytol 197(1):223–237

Caicedo AL, Schaal BA (2004) Heterogeneous evolutionary processes affect R-gene diversity in natural populations of *Solanum pimpinellifolium*. Proc Natl Acad Sci USA 101:17444–17449

Doganlar S, Frary A, Daunay MC, Lester RN, Tanksley SD (2002) A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the Solanaceae. Genetics 161:1697–1711

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Ercolano MR, Sanseverino W, Carli P, Ferriello F, Frusciante L (2012) Genetic and genomic approaches for R-gene mediated disease resistance in tomato: retrospects and prospects. Plant Cell Rep 31:973–985

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Foolad MR (2007) Genome mapping and molecular breeding of tomato. Int J Plant Genomics 64358:52

Hanson PM, Bernacchi D, Green S, Tanksley SD, Muniyappa V, Padmaja AS, Chen H, Kuo G, Fang D, Chen J (2000) Mapping a wild tomato introgression associated with

Tomato yellow leaf curl virus resistance in a cultivated tomato line. J Am Soc Hort Sci 125:15–20

Kaloshian I, Yaghoobi J, Liharska T, Hontelez J, Hanson D, Hogan P, Jesse T, Wijbrandi J, Simons G, Vos P, Zabel P, Williamson VM (1998) Genetic and physical localization of the root-knot nematode resistance locus *Mi* in tomato. Mol Gen Genet 257:376–385

Martin GB, Williams JG, Tanksley SD (1991) Rapid identification of markers linked to a *Pseudomonas* resistance gene in tomato by using random primers and near-isogenic lines. Proc Natl Acad Sci USA 88:2336–2340

Mazourek M, Cirulli ET, Collier SM, Landry LG, Kang BC, Quirin EA, Bradeen JM, Moffett P, Jahn MM (2009) The fractionated orthology of *Bs2* and *Rx*/*Gpa2* supports shared synteny of disease resistance in the Solanaceae. Genetics 182:1351–1364

McDowell JM, Simon SA (2008) Molecular diversity at the plant-pathogen interface. Dev Comp Immunol 32:736–744

McHale L, Tan X, Koehl P, Michelmore RW (2006) Plant NBS-LRR proteins: adaptable guards. Genome Biol 7:212

Miyakawa T, Miyazono K, Sawano Y, Hatano K, Tanokura M (2009) Crystal structure of ginkbilobin-2 with homology to the extracellular domain of plant cysteine-rich receptor-like kinases. Proteins 77:247–251

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

Pedley KF, Martin GB (2003) Molecular basis of Pto-mediated resistance to bacterial speck disease in tomato. Annu Rev Phytol Pathol 41:215–243

Riely B, Martin G (2001) Ancient origin of pathogen recognition specificity conferred by the tomato disease resistance gene *Pto*. Proc Natl Acad Sci USA 98:2059–2064

Sawano Y, Miyakawa T, Yamazaki H, Tanokura M, Hatano K (2007) Purification, characterization, and molecular gene cloning of an antifungal protein from *Ginkgo biloba* seeds. Biol Chem 388:273–280

Stevens MR, Lamb EM, Rhoads DD (1995) Mapping the Sw-5 locus for tomato spotted wilt virus resistance in tomatoes using RAPD and RFLP analyses. Theor Appl Genet 90: 451–456

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

Tang X, Xie M, Kim YJ, Zhou J, Klessig DF, Martin GB (1999) Overexpression of *Pto* activates defense responses and confers broad resistance. Plant Cell 11:15–29

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Xiao S, Ellwood S, Calis O, Patrick E, Li T, Coleman M, Turner JG (2001) Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by *RPW8*. Science 291:118–120