PLOS | ONE

RESEARCH ARTICLE

Dramatic Number Variation of *R* Genes in Solanaceae Species Accounted for by a Few *R* Gene Subfamilies

Chunhua Wei[1,2], Jiongjiong Chen[1]*, Hanhui Kuang[1]

1 Key Laboratory of Horticultural Plant Biology, Ministry of Education, and Key Laboratory of Horticultural Crop Biology and Genetic improvement (Central Region), MOA, College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, P.R. China, 430070, **2** College of Horticulture, Northwest A&F University, Yangling, Shanxi, China, 712100

* jjchen@mail.hzau.edu.cn

Citation: Wei C, Chen J, Kuang H (2016) Dramatic Number Variation of *R* Genes in Solanaceae Species Accounted for by a Few *R* Gene Subfamilies. PLoS ONE 11(2): e0148708. doi:10.1371/journal.pone.0148708

Editor: Keqiang Wu, National Taiwan University, TAIWAN

Received: October 21, 2015

Accepted: January 20, 2016

Published: February 5, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by National Natural Science Foundation of China: No. [31221062] and [31272030]; "973" National Key Basic Research Program: no. 2009CB119000; Natural Science Foundation of Hubei Province of China: NO. [2013CFB204]; The Fundamental Research Funds for the Central Universities: No.[2014PY031].

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Most disease resistance genes encode nucleotide-binding-site (NBS) and leucine-rich-repeat (LRR) domains, and the NBS-LRR encoding genes are often referred to as *R* genes. Using newly developed approach, 478, 485, 1,194, 1,665, 2,042 and 374 *R* genes were identified from the genomes of tomato Heinz1706, wild tomato LA716, potato DM1-3, pepper Zunla-1 and wild pepper Chiltepin and tobacco TN90, respectively. The majority of *R* genes from Solanaceae were grouped into 87 subfamilies, including 16 TIR-NBS-LRR (TNL) and 71 non-TNL subfamilies. Each subfamily was annotated manually, including identification of intron/exon structure and intron phase. Interestingly, TNL subfamilies have similar intron phase patterns, while the non-TNL subfamilies have diverse intron phase due to frequent gain of introns. Prevalent presence/absence polymorphic *R* gene loci were found among Solanaceae species, and an integrated map with 427 *R* loci was constructed. The pepper genome (2,042 in Chiltepin) has at least four times of *R* genes as in tomato (478 in Heinz1706). The high number of *R* genes in pepper genome is due to the amplification of *R* genes in a few subfamilies, such as the *Rpi-blb2* and *BS2* subfamilies. The mechanism underlying the variation of *R* gene number among different plant genomes is discussed.

Background

Plants harbor a variety of disease resistance genes to protect themselves from their natural enemies, such as pests, viruses and fungi. Up to now, more than 140 resistance genes have been cloned and well characterized from flowering plants, of which approximately 80% encode nucleotide-binding-site (NBS) and leucine-rich-repeat (LRR) domains [1–3]. The NBS-LRR encoding genes belong to a large gene family, with hundreds of copies in a genome [4–6]. Based on their N-terminal structures, these R proteins can be further divided into two subclasses: TIR-NBS-LRR (TNL) that possesses a domain homologous to the Toll and interleukin-1 receptor (TIR), and non-TNL. Most non-TNL R proteins have a coiled-coil (CC) structure at

the N terminal and are often called CC-NBS-LRR (CNL) R proteins [7,8]. For convenience, all NBS-LRR encoding genes and their truncated close homologs are referred to *R* genes hereafter in this manuscript.

R genes have been identified from genomes of several plant species. For instance, 159 and 185 *R* genes were identified from the genomes of *Arabidopsis thaliana* and *A*. *lyrata*, respectively [9]; 623 and 725 from rice cultivars Nipponbare and 93–11, respectively [5]; 292 from Brachypodium [10]; 459 and 330 from woody species grape and poplar, respectively [11]; 355 from cotton [12]; 1,015 from apple [13]; 571, 289, 337 and 465 from four legume species [3]; 755, 394 and 684 from potato (*Solanum tuberosum*), tomato (*S*. *lycopersicum*) and pepper (*Capsicum annuum*), respectively [14,15]. Some species contain relative few *R* genes in genomes, for example, only 54 in the papaya genome [16] and 55–75 from an individual Cucurbitaceae species [17].

The majority of *R* genes tend to be physically clustered in plant genomes, forming gene clusters (also called *R* loci or multiple-copy *R* loci) [18]. For example, 109 of the 149 NBS-LRR genes in Arabidopsis were organized in clusters [8,18]; 119 multiple-copy loci were detected in the genome of rice cultivar Nipponbare, composing the majority (74.3%) of *R* genes (623) annotated in rice [5]; similarly, 577 out of 755 *R* genes in potato were located within a total of 92 clusters [14]. Gene duplication was considered to have played an important role in the expansion of an *R* gene family [3,10,13].

Comparative analysis revealed that presence/absence (P/A) polymorphism is prevalent between species [5,9,19,20]. An integrated map of *R* gene loci for a plant family will be helpful to understand the distribution of P/A polymorphism as well as their evolutionary mechanism, and it is also useful for future mapping and cloning of *R* genes [3,5]. *R* gene family represents the most divergent gene family in plant genome, with considerable copy number variation, P/A polymorphism as well as sequence variation caused by various evolutionary mechanisms [5,18,20,21].

R genes have been identified and compared between two closely related Solanaceae species, tomato and potato [14,22–25]. The genomes of some important Solanaceae species have been sequenced and released recently, but comparison of *R* genes and their evolution in Solanaceae species remain to be a comprehensively investigated [15,26–32].

In this study, we re-identified *R* genes from Solanaceae species and hundreds of additional *R* genes were obtained. Their distribution, organization, classification, and P/A polymorphism were analyzed. All *R* gene subfamilies were manually annotated, providing reference gene model for future studies of *R* genes in Solanaceae species. Using tomato Heinz1706 genome as a reference, an integrated *R* gene map with 427 *R* loci was constructed for Solanaceae, which may facilitate future *R* gene cloning from Solanaceae species. The mechanism for copy number variation of *R* genes among Solanaceae species was studied in detail.

Results

Identification of *R* Genes in Solanaceae Species

Using a BLASTN method (details in MM section) [5], 465, 485, 1,185, 1,665, 2,042 and 374 *R* genes were identified from the genomes of tomato cultivar Heinz1706, wild tomato LA716 (*S*. *pennellii*), potato DM1-3, pepper cultivar Zunla-1, wild pepper Chiltepin (*C*. *annuum* var. *glabriusculum*) and tobacco (*Nicotiana tabacum*) TN90, respectively (Table 1, S1 and S2 Datas). Compared with previous studies [14,15], 121 and 450 additional *R* genes were obtained from the genomes of tomato Heinz1706 and potato DM1-3 (S1 Table). Most (> 80%) of the newly identified genes do not have the NBS encoding sequences, but are highly similar to at least one NBS-LRR encoding genes, and therefore are partial *R* genes. Meanwhile, 13 (from tomato Heinz1706) and 9 *R* homologs (from potato DM1-3) identified by previous studies were missed using our method and were

Table 1. The number of *R* genes and their distribution on different chromosomes of Solanaceae species.

	Tomato (Heinz1706)	Tomato (LA716)	Potato (DM1-3)	Pepper (Zunla-1)	Pepper (Chiltepin)	Tobacco (TN90)
Chr00	7	12	226	636	1050	/
Chr01	31	38	68	104	81	/
Chr02	30	27	33	27	49	/
Chr03	11	14	15	116	115	/
Chr04	81	65	159	30	44	/
Chr05	61	58	88	141	107	/
Chr06	27	40	125	91	64	/
Chr07	17	17	21	51	57	/
Chr08	21	19	81	31	58	/
Chr09	49	55	101	177	154	/
Chr10	42	54	80	95	97	/
Chr11	54	53	124	92	97	/
Chr12	34	33	64	74	69	/
Total number	465	485	1185	1665	2042	374

Chr00 represents unanchored supper-scaffolds.

doi:10.1371/journal.pone.0148708.t001

also included for further analysis (S1 and S2 Tables). Consequently, a total of 478 and 1,194 *R* genes from Heinz1706 and DM1-3 were used for further analysis (S1 and S2 Tables).

The number of *R* genes in the genome of pepper (Chiltepin, in particular) is considerably more than that in most other plant genomes sequenced so far (S3 Table). On the other hand, some other Solanaceae species, such as tobacco and tomato, have only a moderate number of *R* genes (Table 1). The *R* gene copy number is inconsistent with the number of predicted genes or genome sizes among Solanaceae species. For example, the tetraploid tobacco has the largest genome and the largest number of predicted genes, but has low *R* gene number. The large variation of *R* gene copy number among different Solanaceae genomes is in striking contrast to the similar number of *R* genes in different Cucurbitaceae species [17].

Classification of 87 *R* Gene Subfamilies in Solanaceae

To facilitate *R* gene identification and annotation in future studies, all *R* genes identified from Solanaceae species were classified into subfamilies. Using BLASTN method (E-value $< 1e^{-10}$), the majority (more than 92%) of *R* genes in Solanaceae can be divided into 87 subfamilies, while the remaining were partial and could not be grouped into any subfamilies. A subfamily is named after a well characterized or a randomly chosen full-length gene in a group, such as subfamily *ZL-0810*. Fourteen of the 87 subfamilies have at least one *R* gene well characterized in previous studies. For the remaining 73 subfamilies, one full-length sequence was chosen to annotate manually (see MM section). They (one for each of the 87 subfamilies) are referred to as ref-genes hereafter. The 87 subfamilies include 16 TNL and 71 nTNL types (Table 2 and S4 Table). The classification of all *R* genes and their detailed annotation will help future cloning and sequence analysis of *R* genes in Solanaceae.

The Evolution of *R* gene Subfamilies

A distance tree was constructed using predicted amino acid sequences of the NBS domain of the 87 subfamilies. As expected, the 16 TNL subfamilies and the 71 nTNL subfamilies are grouped into two major lineages, respectively (Fig 1A).

Table 2. *R* gene subfamilies in Solanaceae species.

	Tomato (Heinz1706)	Tomato (LA716)	Potato (DM1-3)	Pepper (Zunla-1)	Pepper (Chiltepin)	Tobacco (TN90)	Ref-genes
Total no. of Ref-genes	81	79	83	80	83	54	87
No. of *R* genes not homologous to Ref-genes	10	9	28	61	94	28	/
No. of *R* genes homologous to Ref-genes	468	476	1166	1604	1948	346	/
No. of TNL genes	98	97	250	177	204	69	16
No.of nTNL genes	370	379	916	1427	1744	277	71
No. of subfamily with one homolog	38	31	20	16	16	13	/
No. of subfamily with ten or more homologs	14	15	23	31	30	14	/
Average No. of *R* genes in a subfamily	5.8	6.0	14.1	20.1	23.5	6.4	/

doi:10.1371/journal.pone.0148708.t002

To study the origin and evolution of the 87 *R* subfamilies from Solanaceae, they were compared with the *R* genes from Brassicaceae (*A*. *thaliana*), Poaceae (rice, maize, sorghum and *Brachypodium distachyon*) and Cucurbitaceae species (cucumber, melon and watermelon) using TBLASTX [5,9,17]. If a Solanaceae *R* gene subfamily is present in a genome, "+" is marked in corresponding position in Fig 1B. As expected, the 16 TNL subfamilies are not present in Poaceae. Interestingly, most of the 71 nTNL subfamilies are found in Poaceae (monocot) but more than half of them are absent in Cucurbitaceae and Arabidopsis (dicot), suggesting frequent loss of *R* subfamilies in these two plant families (Fig 1B).

The P/A polymorphism of *R* subfamilies also occurs among different Solanaceae species (Fig 1A and 1B). For instance, the *ZL-0810* and *ZL-1520* subfamilies are present in pepper but absent in Solanum (potato and tomato); the *DM-0233*, *DM-0846* and *R1* subfamilies are present in potato but lost in tomato; the *SL-0076* subfamily is lost in pepper; the *SL-0165* and *SL-0166* subfamilies are absent in pepper cultivar Zunla-1. The P/A polymorphism between different Solanaceae species showed that *R* gene subfamilies might be lost during a short evolutionary period.

Intron Feature Variations and Intron Gain/Loss in TNL and nTNL Subfamilies

The intron-exon boundaries and intron phase of the 87 ref-genes of Solanaceae were studied and compared. Since members in each subfamily are close homologs, it was assumed that the gene structure of each ref-gene represents that of all genes in corresponding subfamily. Surprisingly, the average number of exons for each subfamily varies dramatically between the TNL and nTNL groups, though the two groups have similar gene length. Of the 71 nTNL subfamilies, 42 have only a single exon, 24 have two exons and 5 have more than two exons, while all 16 TNL subfamilies have 4 or 5 exons (Fig 1C and S4 Table).

The distance tree and the gene structure (intron position and intron phase) suggest that the common ancestor of the nTNL lineage had no introns since none of the introns were present in two distantly related nTNL subfamilies (Fig 1C). Most (42) nTNL subfamilies have no intron, and these subfamilies with no intron spread all over the distance tree. For the 29 subfamilies with intron, they are always closely related if they have the same intron feature. For example, subfamilies *SL-0030*, *SL-0071*, *SL-0086*, *SL-0085*, *SL-0096* and *SL-0119* have the same gene structure and are closely related. Therefore, the introns in each nTNL subfamily were

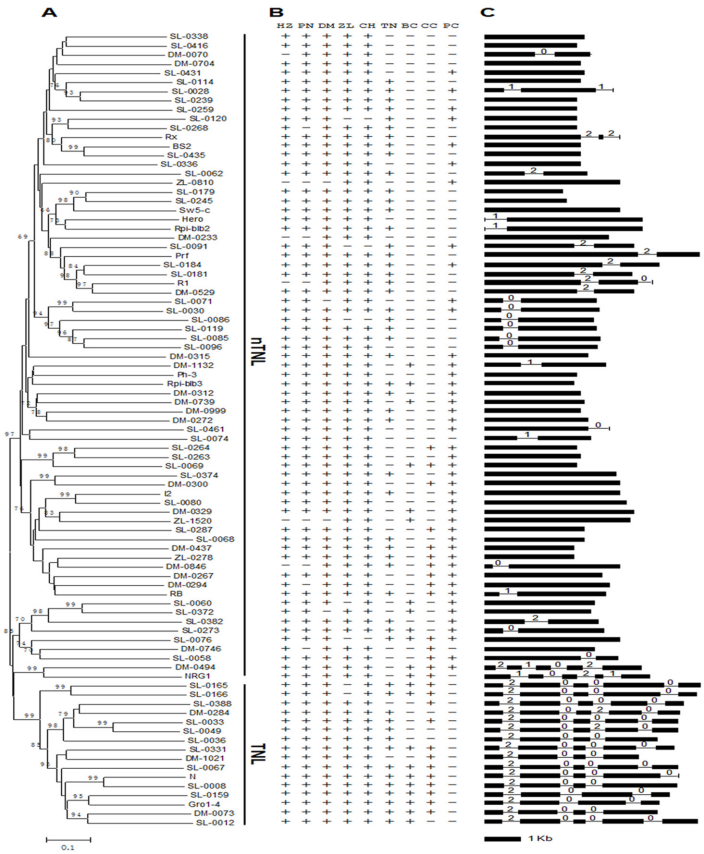

Fig 1. (A). Distance tree for the 87 ref-genes in Solanaceae. Amino acid sequences of NBS region of the 87 ref-genes were used to construct the tree. The ref-genes from tomato Heinz1706, potato DM1-3 and pepper Zunla-1 were named as *HZ*-, *DM*- and *ZL*- followed by a number, respectively. Numbers on nodes are bootstrap values, and values <65 are not shown. (B). P/A polymorphisms of the 87 subfamilies in four plant families (Solanaceae, Brassicaceae, Cucurbitaceae and Poaceae). HZ, PN, DM, ZL, CH, TN, BC, CC and PC on the top line represent tomato Heinz1706, tomato LA716, potato DM1-3, pepper Zunla-1, pepper Chiltepin, tobacco TN90, Brassicaceae, Cucurbitaceae and Poaceae, respectively. The mark "+" represent presence of the subfamily. (C). Gene models for the 87 ref-genes in Solanaceae. Black boxes represent exons, while lines linking boxes represent introns. The number indicates intron phase: 0 = intron phase 0; 1 = intron phase 1; 2 = intron phase 2. The bar represents the scale of exon, while introns are not drawn to scale.

doi:10.1371/journal.pone.0148708.g001

gained independently (such as *SL-0062*) or from the common ancestor of a few closely related subfamilies (such as the common ancestor of the aforementioned six subfamilies).

In contrast, the intron features in different subfamilies of the TNL group are highly conserved. The pattern in Fig 1C indicates that the common ancestor of the TNL group might have 3 or 4 introns (as in the *N* gene). The first three introns of the *N* gene are conserved in most TNL subfamilies, with only a few exceptions. For example, the *Gro1-4* and the *SL-0159* subfamilies might have lost intron 2. The last intron (as in the *N* gene) in the TNL subfamilies might have been subjected to frequent gain and loss since it is present in some but absent in other subfamilies.

An Integrated *R* Gene Map for Solanaceae Species Showing Prevalent P/A Polymorphism of *R* Gene Loci

R genes separated by no more than eight non-*R* genes are considered to be located at the same *R* gene locus (a multiple-copy locus) [5,33]. A total of 218, 298, 347 and 384 *R* gene loci were

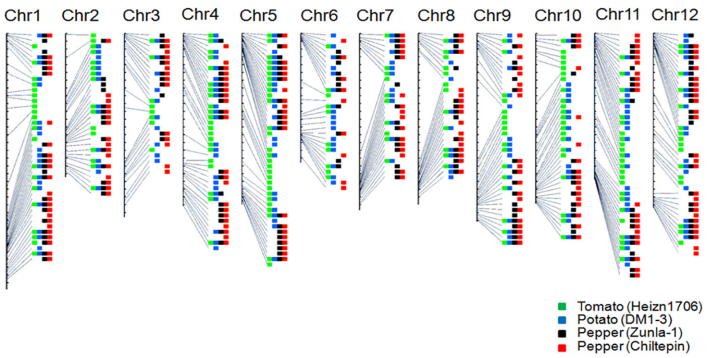

Fig 2. An integrated map of *R* loci in Solanaceae. *R* genes from different Solanaceae species were mapped onto the 12 chromosomes of tomato Heinz1706. *R* loci from tomato Heinz1706, potato DM1-3, pepper Zunla-1 and pepper Chiltepin are in green, blue, black and red, respectively.

doi:10.1371/journal.pone.0148708.g002

identified in tomato Heinz1706, potato DM1-3, pepper Zunla-1 and pepper Chiltepin (S5 Table). Of them, 238, 265 and 294 *R* loci from potato DM1-3, pepper Zunla-1 and pepper Chiltepin, which contain 761, 611 and 630 *R* homologs, respectively, were mapped into the syntenic regions of tomato chromosomes, while the remaining *R* genes failed to be mapped due to lack of enough flanking sequences. Consequently, an integrated map with 427 *R* gene loci was constructed (Fig 2). Of them, 371 *R* loci exhibit P/A polymorphism among Solanaceae species, while the other 56 *R* loci are shared by all four species (S2 Table). There are 176 loci specific to one individual species: 66 specific to tomato, 43 specific to potato, 67 specific to pepper (Zunla-1 and Chiltepin). This integrated *R* gene map will be helpful for future cloning of *R* genes from Solanaceae species.

Copy Number Variation in Solanaceae Accounted for by a Few *R* Subfamilies

Of the 87 *R* subfamilies in Solanaceae, only 81, 79, 83, 80, 83 and 54 subfamilies were found in the genomes of tomato Heinz1706, tomato LA716, potato DM1-3, pepper Zunla-1, pepper Chiltepin and tobacco TN90, respectively (Table 2). Three subfamilies (*R1*, *DM-0233* and *DM-0846*) are present in potato but absent in tomato (Fig 1 and S6 Table). Similarly, six subfamilies (*R1*, *DM-0233*, *DM-0846*, *ZL-0810*, *ZL-1520* and *CH-0768*) were found in pepper but lost in tomato. The subfamilies with P/A polymorphism among different species are usually small and it does not account for the dramatic variation of *R* gene copy number among different species (S6 Table). In other words, the number of subfamilies in a genome is not the main cause of *R* gene number variations among different species. On the contrary, the average size of *R* gene subfamilies varies dramatically in a genome. The average number of *R* genes in an *R* subfamily in pepper (20.1 and 23.5 in Zunla-1 and Chiltepin, respectively) were at least three times as many as that in tomato (5.8 and 6.0) and tobacco (6.4) (Table 2).

The five largest subfamilies (*Rpi-blb2*, *BS2*, *SL-0273*, *Sw5-c* and *I2*) in pepper Zunla-1 and Chiltepin contain 832 (50.0%) and 1,027 (50.3%) *R* genes, respectively (S6 Table). In comparison, these five subfamilies have only 87 (18.2%), 95 (19.6%) and 73 (19.5%) *R* genes in tomato Heinz1706, tomato LA716 and tobacco TN90, respectively. The 22 largest *R* subfamilies contain approximately 80% of all *R* genes in pepper Zunla-1 (1,339) and Chiltepin (1,643). Therefore, the large variation of *R* gene number among Solanaceae species is mainly accounted for by a few *R* subfamilies.

Copy Number Variation of the *Rpi-blb2* Subfamily

As the largest *R* subfamily in pepper, 298 and 366 *Rpi-blb2* homologs were found in pepper Zunla-1 and Chiltepin (S6 Table), respectively. Interestingly, the *Rpi-blb2* subfamily is Solanaceae specific, since no homologs were detected in Brassicaceae, Poaceae or Cucurbitaceae (Fig 1B). Of the chromosome-anchored *Rpi-blb2* homologs, the majority of them (> 78.0%) were located on chromosomes 5 and 6 of Solanaceae (S2 Table). Five *Rpi-blb2* loci in pepper genome have more than 10 copies. Consistent with above conclusion, the tomato Heinz1706 (28) and potato DM1-3 (86) genomes have relatively few *Rpi-blb2* genes (S6 Table). Nevertheless, the *Rpi-blb2* subfamily is the fourth and third largest subfamily in tomato Heinz1706 and potato DM1-3.

To analyze the evolution of the *Rpi-blb2* subfamily in Solanaceae, the nucleotide sequences (excluding short ones) of NBS-encoding region of 154 *Rpi-blb2* homologs (17, 29, 91 and 17 from tomato Heinz1706, potato DM1-3, pepper Zunla-1 and tobacco TN90, respectively) were aligned and a distance tree was constructed (S1 Fig). Two clades (I and II) are found on the tree and each clade has genes from all species included in this study, indicating two ancient lineages in the progenitor of Solanaceae. Many of the *Rpi-blb2* homologs of an individual species are grouped together, suggesting duplications after speciation. The majority (80%) of pairwise nucleotide identities of *Rpi-blb2* homologs of pepper in Clade I and II were lower than 90.3% and 90.5%, respectively, and there are only a few recent duplications (with nearly identical sequences). When *Rpi-blb2* homologs from all species (including wild tomato LA716 and wild pepper Chiltepin), the topology of the distance tree remains unchanged (data not shown). Ten pairs of obvious orthologs are found between tomato cultivar Heinz1706 and wild tomato LA716, and 125 pairs of orthologs between pepper cultivar Zunla-1 and wild pepper Chiltepin using bi-directional BLASTN. Sequence analysis of 59 nearly full-length (~3.5 Kb) *Rpi-blb2* homologs from pepper (both genotypes) found only 8 sequence exchanges ($P < 0.05$), consistent with independent evolution of *Rpi-blb2* homologs (data not shown).

Copy Number Variation of the *BS2* Subfamily

The *BS2* subfamily showed dramatic number variation between Solanum (7 and 3 homologs in tomato Heinz1706 and potato DM1-3, respectively) and Capsicum (271 in pepper Zunla-1 and 355 in pepper Chiltepin) (S6 Table), consistent with previous report [15]. In Solanum, the ten *BS2* homologs from tomato Heinz1706 and potato DM1-3 were located at seven *R* loci, randomly distributed on five chromosomes (Chr02, 04, 08, 11 and 12) (S2 Table). Unfortunately, a large number of the *BS2* homologs in pepper (135 in Zunla-1 and 243 in Chiltepin) were not anchored onto chromosomes. Nevertheless, the majority of the chromosome-anchored *BS2* homologs in pepper (114 of 136 in Zunla-1 and 87 of 112 in Chiltepin) were mapped on chromosomes 7 and 9. Some *R* loci on chromosome 7 and 9 of pepper were species-specific and harbored dozens of *BS2* homologs. For example, the *R* loci ZL-locus-169 and -215 are specific to pepper Zunla-1, which contain 18 and 49 *BS2* homologs, respectively; the *R* loci CH-locus-187 and -244 are specific to pepper Chiltepin and contain 11 and 47 *BS2* homologs. Therefore, unlike most other species specific *R* loci (see above), the *BS2* homologs in pepper-specific *R* loci are large and contribute to the number variation of this subfamily between Solanum and Capsicum.

To understand the relationship of the *BS2* subfamily, a distance tree was constructed using nucleotide sequences (excluding short sequences) of the NBS-encoding region of 101 homologs from tomato Heinz1706 (1), potato DM1-3 (1), pepper Zunla-1 (91) and tobacco TN90 (8) (Fig 3). The only full-length *BS2* homologs from tomato and potato belong to two different clades, clade II and V, respectively. To better understand the evolution of the *BS2* subfamily, its

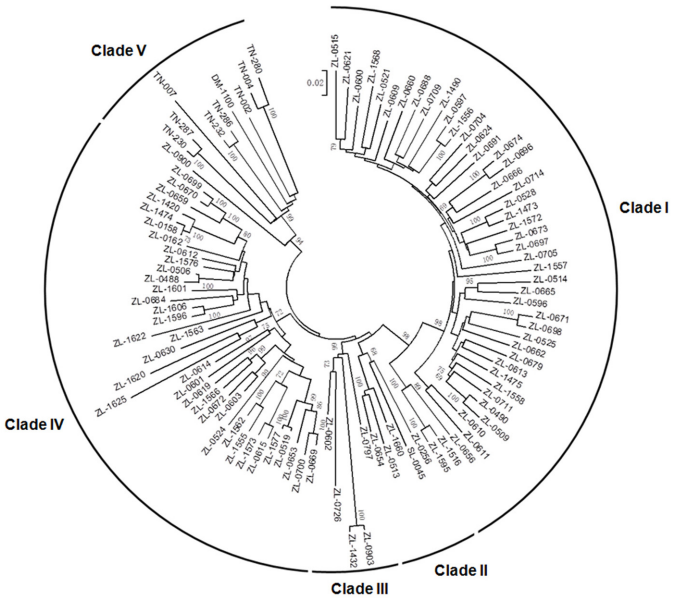

Fig 3. Distance tree of *BS2* homologs from Solanaceae species. Five clades can be found in the tree. Numbers on nodes are bootstrap values, and values <65 are not shown. Genes with name "*SL*-" are from tomato Heinz1706; genes with name "*DM*-" are from potato DM1-3; genes with name "*ZL*-" are from pepper Zunla-1; genes with name "*TN*-" are from tobacco TN90.

doi:10.1371/journal.pone.0148708.g003

expansion in pepper in particular, the eight partial *BS2* homologs from tomato (6) and potato (2), which were not used in constructing of the distance tree, were chosen to compare with all pepper *BS2* homologs in Fig 3. BLAST analysis showed that the eight partial *BS2* homologs belong to clades I, II and IV, respectively (S7 Table). Therefore, *BS2* homologs from Solanum are present in at least four of the five clades, while clade III is specific to pepper. The tree indicates that clades I-IV might have been present in the common ancestor of Capsicum and Solanum. After speciation, clade V was lost in tomato, while clade I and IV were lost in potato. In contrast, these four clades were maintained in pepper genome. Clades I and IV contain dozens of *BS2* homologs from pepper Zunla-1, obviously duplicated after speciation (Fig 3). The duplications of *BS2* homologs of clades I and IV after speciation predominantly accounted for the high copy number of the *BS2* subfamily in pepper.

To understand their evolution after duplication, sequence exchanges among 130 nearly full-length (~2.5 Kb) *BS2* homologs from pepper (both genotypes) were investigated. Only eight pairs of sequence exchanges were detected among 130 homologs ($P < 0.05$), suggesting independent evolution of most *BS2* homologs. A total of 145 pairs of orthologs were detected in all *BS2* homologs between pepper cultivar Zunla-1 and wild pepper Chiltepin (data not shown), consistent with the conclusion of independent evolution.

Discussion

The Current Method May Identify Additional *R* Genes from a Genome

R genes have been identified and analyzed from several plant genomes including Solanaceae species [3,5,8,10–15,17]. In previous studies, a species-specific NBS domain HMM profile was first constructed using hmmbuild implemented in the HMMER software. Then, the species-specific HMM profile was used to identify NBS genes in a genome with hmmsearch. Using above method, 54, 159, 435, 239, 1,015 and 327 *R* genes were detected in the genomes of

papaya, Arabidopsis, potato, Brachypodium, apple and cassava, respectively, and these identified *R* genes all contain the NBS encoding sequences [9,10,13,16,23,34]. Obviously, partial *R* genes lacking NBS encoding sequences were missed in these studies. A new approach, the *R* gene enrichment and sequencing (RenSeq) workflow, successfully increased the numbers of *R* genes from 356 and 438 to 394 and 755 in the genomes of tomato and potato, respectively [14]. However, some *R* genes, such as partial ones, would not be detected using either of above methods. We have developed a greatly improved strategy for *R* gene identification from a sequenced genome, which was successfully used to identify hundreds of additional *R* gene homologs from a genome [5]. First, a species-specific R protein database was constructed and representative sequences were used as queries to do tBLASTn, and consequently intact or partial *R* genes (including genes lacking NBS domain) would be detected from a genome. Using this method, hundreds of additional *R* genes were identified from different genomes in this study (Table 1 and S1 Table). Though most of these newly identified *R* genes are partial genes, their alleles/orthologues may be full-length and functional, and therefore their identification may provide useful reference for future studies on *R* genes in these species.

Variation of *R* Gene Copy Number among Different Species

The number of *R* genes from pepper Zunla-1 (1,665) and Chiltepin (2,042) is the largest in diploid genomes sequenced so far, which is similar to the number of *R* genes in hexaploid wheat [35]. Of the dozens of plant genomes sequenced so far, some genomes have small number of *R* genes, such as cucumber (70) and papaya (54) [16,17]. It remains unclear why the number of *R* genes can vary up to 40 times between different plant species. Theoretically, the more *R* gene homologs a genome has, the more functional resistance genes the species may harbor. However, the expansion of *R* gene homologs has fitness cost [36], and for each species, there should be a balance between resistance and fitness cost provided by all *R* gene homologs in a genome. Further studies are required to address the difference of such balance for different species.

The striking variation of *R* gene copy number may or may not exist within a plant family. The number of *R* genes from the Cucurbiteceae family is unexceptionally low [17]. On the other hand, the number of *R* genes may vary dramatically within a plant family. For example, the pepper genome (Chiltepin) has more than four times as many *R* genes as that in the tomato genome (Heinz1706) (Table 1 and S1 Table).

The Majority of *R* Genes in Plants May Be Classified into the 87 *R* Subfamilies

In this study, 87 *R* subfamilies were classified in Solanaceae species, one gene (ref-gene) from each subfamily were annotated fully. Among these 87 ref-genes (including 73 well annotated and 14 functional *R* genes), 71 were nTNL genes and 16 were TNL genes (Table 2). Except partial ones, the remaining *R* genes identified from tomato Heinz1706 (97.9%), tomato LA716 (98.1%), potato DM1-3 (97.7%), pepper Zunla-1 (96.3%), pepper Chiltepin (95.4%) and tobacco TN90 (92.5%) were categorized into 81, 79, 83, 80, 83 and 54 of the 87 *R* subfamilies (Table 2). For comparative analysis in this study, 1,934, 159 and 147 *R* genes identified from Poaceae (rice, maize, sorghum and brachypodium), Brassicaceae (Arabidopsis) and Cucurbitaceae species (cucumber, melon and watermelon) were objected to compare with the 87 ref-genes using TBLASTX method [5,9,17]. As a result, 1,645 (85.1%), 159 (100%) and 146 (99.3%) of *R* genes in Poaceae, Brassicaceae and Cucurbiteceae were classified to one of the 87 Solanaceae *R* subfamilies (data not shown). In other words, the 87 *R* subfamilies contain the majority ($>$ 85%) of *R* genes in the four plant families. In future studies, it will be very useful to identify additional *R* gene subfamilies from other plant species and construct a database with

comprehensive *R* gene subfamilies, which represent all ancient *R* gene lineages. A comprehensive database of *R* gene subfamilies will facilitate studies on annotation, cloning and evolution of *R* genes in future studies.

Extensive Amplification of a Few *R* Subfamilies May Increase *R* gene Copy Number Considerably

Although the number of *R* genes in pepper Zunla-1 and Chiltepin is considerably more than that in other Solanaceae species, the number of *R* subfamilies in different Solanaceae species is similar (Tables 1 and 2). Further analysis showed that subfamilies with P/A polymorphism between species are usually small and make little contribution to the copy number variation of *R* genes among species. More than 80% of all *R* genes were contributed by the top 24, 23, 20, 22, 22 and 27 largest *R* subfamilies in tomato Heinz1796, tomato LA716, potato DM1-3, pepper Zunla-1, pepper Chiltepin and tobacco TN90, respectively (S6 Table). As expected, functional *R* genes are more likely from the large subfamilies: 12 of the 14 cloned *R* genes from Solanaceae belong to these top largest *R* subfamilies. For instance, subfamilies *I2* and *Hero*, which harbor functional genes conferring resistance to *Fusarium oxysporum* f sp *lycopersici* and *Globodera rostochiensis*, respectively [37,38], are the third and ninth largest one in tomato; subfamilies *RB* and *Rpi-blb3*, which provide resistance to *Phytophthora infestans* [39,40], are the second and tenth largest one in potato, respectively; similarly, subfamily *N* is the fifth largest one in tobacco, conferring resistance to tobacco mosaic virus (TMV) [41]. The copy number of a subfamily may also vary dramatically. For example, subfamily *BS2* and *Rpi-blb2*, as the top two largest *R* subfamilies in pepper, have a total of 569 homologs in Zunla-1 and 721 in Chiltepin but they have no more than 90 homologs in Solanum and tobacco, respectively. The top five largest *R* subfamilies in pepper genomes contain nearly half of all *R* genes (832 of 1,665 and 1,027 of 2,042 in Zunla-1 and Chiltepin, respectively), while these five subfamilies harbor only 87 (18.2%), 95 (19.6%) and 73 (19.5%) *R* genes in tomato Heinz1706, tomato LA716 and tobacco TN90, respectively. In summary, the expansion of a few *R* gene subfamilies may substantially increase the copy number of *R* genes in a genome.

Materials and Methods

Identification of *R* Genes in Solanaceae Species

The genome data and gene models of tomato cultivar Heinz1706 (version 2.50), *S. pennelli* LA716 (version 1.0), potato DM1-3 (version 4.03), cultivar pepper Zunla-1 (version 2.0), wild pepper Chiltepin (version 2.0) and tobacco TN90 (SRP029183) were downloaded from corresponding web sites [26,27,29,30,32].

To genome-wide identify *R* genes in genomes, both Hidden Markov Model (HMM) and BLAST methods were used in this study [5]. Firstly, amino acid sequences of each genome were used to search against the HMM profile of NB-ARC domain (Pfam PF00931) using software HMMER3.0 with default parameter settings. Secondly, using key words, such as "NBS-LRR, NB ARC", "ATP binding cassette" and "LRR kinase", related protein sequences were retrieved from NCBI and a validating database was constructed, containing 3,146 NBS-LRR proteins, 3,098 ABC-transporter proteins and 6,758 kinase proteins. Then, the amino acid sequences identified from HMM searching were used as queries to search against the validating database, using BLASTP program with the E-value setting to $1e^{-10}$. Sequences with best hit of NBS-LRR or NB-ARC protein were used as R protein seeds to identify partial or intact *R* gene homologs from a genome using TBLASTN method (E-value cutoff of $1e^{-10}$). Finally, in order to confirm the results, following the rationale described above, all *R* homologs

were validated again using non-redundant protein database from NCBI with BLASTX method (E-value cutoff of $1e^{-10}$). Once again, only sequences with best hit of NBS-LRR protein were considered as candidate *R* genes and used in further study.

Candidate *R* genes identified from tomato (Heinz1706), wild tomato (LA716), potato (DM1-3), pepper (Zunla-1 and Chiltepin) and tobacco (TN90) were named as *HZ-*, *PN-*, *DM-*, *ZL-*, *CH-* and *TN-* followed by a number, respectively.

Annotation and Classification of *R* Genes

The following workflow was used to classify *R* genes into subfamilies. To simplify the calculations, all *R* genes of tomato Heinz1706 were first divided into subfamilies using BLASTN method with E-value setting to $1e^{-10}$. If a subfamily has a close homolog that has been well characterized in previous studies, the well characterized *R* gene was used as the reference gene and its gene model as the reference gene model for this *R* subfamily. In cases of subfamily with no well known *R* genes, full-length genes from the subfamily were chosen for manual annotation, including their intron/exon structure and intron phase. Gene model for full-length genes from the same subfamily are unexceptionally conserved. One of the manually annotated full-length genes was randomly chosen as the reference gene (ref-gene). Then, all *R* genes from other five Solanaceae genomes were compared with the ref-genes and classified to a subfamily using BLASTN method with E-value setting to $1e^{-10}$. *R* genes from the five genomes that have no similarity with any *R* gene subfamilies in tomato, were used to repeat above process: new subfamily classification and ref-gene annotation. If a *R* gene in a subfamily has different gene structure (large missing sequences or premature stop codon), it was considered as a partial gene or/and pseudogene.

Pfam and COILS were used to investigate if an R protein has TIR or CC domain, respectively [42,43]. All TNL and nTNL encoding genes were checked for intron position and intron phase. Three different intron phases in spliceosomal introns are defined: phase-0 as intron located before the first nucleotide of a codon, phase-1 as intron located before the second nucleotide of a codon and phase-2 before the third nucleotide of a codon [8,44,45]. The exon number, exon length and intron phase of ref-genes were annotated manually [5,17].

Phylogenetic Analysis of *R* Genes

The coding sequences of *R* genes were translated into amino acid with a web server AUGUSTUS [46]. All the nucleotide (when all homologs from the same subfamily) or protein sequences (when homologs from different subfamilies) were aligned using program Muscle [47] and edited in GeneDoc. Mega 5.0 was used to construct Neighbor-joining (NJ) tree with Kimura two-parameter substitution model for nucleotide sequences and p-distance model for amino acid sequences [48]. Bootstrap value was calculated using 1,000 replicates, and a claimed clade mostly has a bootstrap value higher than 65. Sequence exchanges were detected using software Geneconv [49] with default parameters and confirmed manually.

P/A Polymorphsim and An Integrated Map of *R* Genes

An *R* gene locus is defined as a locus with two or more *R* genes separated by no more than eight non-*R* genes [5,33]. If one or several *R* genes are present at a locus in a genome but none in another genome at the syntenic region, this locus is considered as P/A polymorphic locus [5]. All *R* gene loci were mapped onto tomato chromosomes according to the synteny of their flanking regions, resulting in an integrated map for Solanaceae [5].

Supporting Information

S1 Data. A fasta-file for all *R* gene sequences identified from tomato and potato genomes.
(TXT)

S2 Data. A fasta-file for all *R* gene sequences identified from pepper and tobacco genomes.
(TXT)

S1 Fig. Distance tree of *Rpi-blb2* homologs from Solanaceae species. Two clades, I and II, are in the tree. Numbers on nodes are bootstrap values, and values <65 are not shown. Genes with name "*SL*-" are from tomato Heinz1706; genes with name "*DM*-" are from potato DM1-3; genes with name "*ZL*-" are from pepper Zunla-1; genes with name "*TN*-" are from tobacco TN90.
(TIF)

S1 Table. Details of *R* gene identified with our method compared with that reported in Jupe et al. (2013).
(XLSX)

S2 Table. Information of all *R* genes identified from Solanaceae species in this study.
(XLSX)

S3 Table. *R* gene numbers in some plants.
(XLSX)

S4 Table. Annotation of 87 reference genes in Solanaceae.
(XLSX)

S5 Table. Organization of *R* genes in Solanaceae species.
(XLSX)

S6 Table. Copy number of homologs in each subfamily in Solanaceae species.
(XLSX)

S7 Table. Information of best hits of eight partial *BS2* homologs from Solanum.
(XLSX)

Author Contributions

Conceived and designed the experiments: JC. Performed the experiments: CW. Analyzed the data: CW. Contributed reagents/materials/analysis tools: CW. Wrote the paper: CW HK JC. Analysis and interpretation of the data and draft of the manuscript: CW. Critical review of the manuscript: HK. Designed the study and revised the manuscript: JC.

References

1. Liu J, Liu X, Dai L, Wang G (2007) Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants. J Genet Genomics 34: 765–776. PMID: 17884686
2. Yang S, Li J, Zhang X, Zhang Q, Huang J, Chen JQ, et al. (2013) Rapidly evolving *R* genes in diverse grass species confer resistance to rice blast disease. Proc Natl Acad Sci U S A 110: 18572–18577. doi: 10.1073/pnas.1318211110 PMID: 24145399
3. Shao ZQ, Zhang YM, Hang YY, Xue JY, Zhou GC, Wu P, et al. (2014) Long-term evolution of nucleotide-binding site-leucine-rich repeat genes: understanding gained from and beyond the legume family. Plant Physiol 166: 217–234. doi: 10.1104/pp.114.243626 PMID: 25052854
4. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. (2006) The genome of black cottonwood, *Populus trichocarpa*. Science 313: 1596–1604. PMID: 16973872

5. Luo S, Zhang Y, Hu Q, Chen J, Li K, Lu C, et al. (2012) Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family. Plant Physiol 159: 197–210. doi: 10.1104/pp.111.192062 PMID: 22422941
6. He L, Du C, Covaleda L, Xu Z, Robinson AF, Yu JZ, et al. (2004) Cloning, characterization, and evolution of the NBS-LRR-encoding resistance gene analogue family in polyploid cotton (Gossypium hirsutum L.). Mol Plant Microbe Interact 17: 1234–1241. PMID: 15553248
7. Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND, et al. (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J 20: 317–332. PMID: 10571892
8. 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. PMID: 12671079
9. Guo YL, Fitz J, Schneeberger K, Ossowski S, Cao J, Weigel D, et al. (2011) Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis. Plant Physiol 157: 757–769. doi: 10.1104/pp.111.181990 PMID: 21810963
10. Tan S, Wu S (2012) Genome Wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*. Comp Funct Genomics 2012: 418208. doi: 10.1155/2012/418208 PMID: 22693425
11. Yang S, Zhang X, Yue JX, Tian D, Chen JQ (2008) Recent duplications dominate NBS-encoding gene expansion in two woody species. Mol Genet Genomics 280: 187–198. doi: 10.1007/s00438-008-0355-0 PMID: 18563445
12. Wei H, Li W, Sun X, Zhu S, Zhu J (2013) Systematic analysis and comparison of nucleotide-binding site disease resistance genes in a diploid cotton *Gossypium raimondii*. PLoS One 8: e68435. doi: 10.1371/journal.pone.0068435 PMID: 23936305
13. Arya P, Kumar G, Acharya V, Singh AK (2014) Genome-wide identification and expression analysis of NBS-encoding genes in Malus x domestica and expansion of NBS genes family in Rosaceae. PLoS One 9: e107987. doi: 10.1371/journal.pone.0107987 PMID: 25232838
14. Jupe F, Witek K, Verweij W, Sliwka J, Pritchard L, Etherington GJ, et al. (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. Plant J 76: 530–544. doi: 10.1111/tpj.12307 PMID: 23937694
15. Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, et al. (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nat Genet 46: 270–278. doi: 10.1038/ng.2877 PMID: 24441736
16. Porter BW, Paidi M, Ming R, Alam M, Nishijima WT, Zhu YJ, et al. (2009) Genome-wide analysis of *Carica papaya* reveals a small NBS resistance gene family. Mol Genet Genomics 281: 609–626. doi: 10.1007/s00438-009-0434-x PMID: 19263082
17. Lin X, Zhang Y, Kuang H, Chen J (2013) Frequent loss of lineages and deficient duplications accounted for low copy number of disease resistance genes in Cucurbitaceae. BMC Genomics 14: 335. doi: 10.1186/1471-2164-14-335 PMID: 23682795
18. Meyers BC, Kaushik S, Nandety RS (2005) Evolving disease resistance genes. Current opinion in plant biology 8: 129–134. PMID: 15752991
19. Luo S, Peng J, Li K, Wang M, Kuang H (2011) Contrasting evolutionary patterns of the *Rp1* resistance gene family in different species of Poaceae. Molecular biology and evolution 28: 313–325. doi: 10.1093/molbev/msq216 PMID: 20713469
20. Wei C, Kuang H, Li F, Chen J (2014) The *I2* resistance gene homologues in Solanum have complex evolutionary patterns and are targeted by miRNAs. BMC Genomics 15: 743. doi: 10.1186/1471-2164-15-743 PMID: 25178990
21. Kuang H, Woo SS, Meyers BC, Nevo E, Michelmore RW (2004) Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. Plant Cell 16: 2870–2894. PMID: 15494555
22. Jupe F, Pritchard L, Etherington GJ, Mackenzie K, Cock PJ, Wright F, et al. (2012) Identification and localisation of the NB-LRR gene family within the potato genome. BMC Genomics 13: 75. doi: 10.1186/1471-2164-13-75 PMID: 22336098
23. Lozano R, Ponce O, Ramirez M, Mostajo N, Orjeda G (2012) Genome-wide identification and mapping of NBS-encoding resistance genes in *Solanum tuberosum* group phureja. PLoS One 7: e34775. doi: 10.1371/journal.pone.0034775 PMID: 22493716
24. Andolfo G, Sanseverino W, Rombauts S, Van de Peer Y, Bradeen JM, Carputo D, et al. (2013) Overview of tomato (Solanum lycopersicum) candidate pathogen recognition genes reveals important

Solanum R locus dynamics. New Phytol 197: 223–237. doi: 10.1111/j.1469-8137.2012.04380.x PMID: 23163550

25. Andolfo G, Sanseverino W, Aversano R, Frusciante L, Ercolano MR (2014) Genome-wide identification and analysis of candidate genes for disease resistance in tomato. Mol Breeding 0.1007/s11032-013-9928-7.
26. Potato Genome Sequencing C, Xu X, Pan S, Cheng S, Zhang B, Wu D, et al. (2011) Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195. doi: 10.1038/nature10158 PMID: 21743474
27. Tomato Genome C (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641. doi: 10.1038/nature11119 PMID: 22660326
28. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, et al. (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345: 1181–1184. doi: 10.1126/science.1255274 PMID: 25190796
29. Qin C, Yu C, Shen Y, Fang X, Chen L, Min J, et al. (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc Natl Acad Sci U S A 111: 5135–5140. doi: 10.1073/pnas.1400975111 PMID: 24591624
30. Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A, et al. (2014) The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun 5: 3833. doi: 10.1038/ncomms4833 PMID: 24807620
31. Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, et al. (2014) Draft genome sequence of eggplant (Solanum melongena L.): the representative Solanum species indigenous to the old world. DNA Res.
32. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, et al. (2014) The genome of the stress-tolerant wild tomato species *Solanum pennellii*. Nat Genet 46: 1034–1038. doi: 10.1038/ng.3046 PMID: 25064008
33. Richly E, Kurth J, Leister D (2002) Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution. Mol Biol Evol 19: 76–84. PMID: 11752192
34. Lozano R, Hamblin MT, Prochnik S, Jannink JL (2015) Identification and distribution of the NBS-LRR gene family in the Cassava genome. BMC Genomics 16: 360. doi: 10.1186/s12864-015-1554-9 PMID: 25948536
35. Gu L, Si W, Zhao L, Yang S, Zhang X (2014) Dynamic evolution of NBS-LRR genes in bread wheat and its progenitors. Mol Genet Genomics.
36. Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J (2003) Fitness costs of *R*-gene-mediated resistance in *Arabidopsis thaliana*. Nature 423: 74–77. PMID: 12721627
37. Simons G, Groenendijk J, Wijbrandi J, Reijans M, Groenen J, Diergaarde P, et al. (1998) Dissection of the fusarium *I2* gene cluster in tomato reveals six homologs and one active gene copy. Plant Cell 10: 1055–1068. PMID: 9634592
38. Ernst K, Kumar A, Kriseleit D, Kloos DU, Phillips MS, Ganal MW, et al. (2002) The broad-spectrum potato cyst nematode resistance gene (Hero) from tomato is the only member of a large gene family of NBS-LRR genes with an unusual amino acid repeat in the LRR region. Plant J 31: 127–136. PMID: 12121443
39. Song J, Bradeen JM, Naess SK, Raasch JA, Wielgus SM, Haberlach GT, et al. (2003) Gene *RB* cloned from *Solanum bulbocastanum* confers broad spectrum resistance to potato late blight. Proc Natl Acad Sci U S A 100: 9128–9133. PMID: 12872003
40. Park TH, Gros J, Sikkema A, Vleeshouwers VG, Muskens M, Allefs S, et al. (2005) The late blight resistance locus *Rpi-bib3* from *Solanum bulbocastanum* belongs to a major late blight *R* gene cluster on chromosome 4 of potato. Mol Plant Microbe Interact 18: 722–729. PMID: 16042018
41. Whitham S, Dinesh-Kumar SP, Choi D, Hehl R, Corr C, Baker B, et al. (1994) The product of the tobacco mosaic virus resistance gene *N*: similarity to toll and the interleukin-1 receptor. Cell 78: 1101–1115. PMID: 7923359
42. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. (2013) Pfam: the protein families database. Nucleic Acids Res 42: D222–230. doi: 10.1093/nar/gkt1223 PMID: 24288371
43. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science 252: 1162–1164. PMID: 2031185
44. Sharp PA (1981) Speculations on RNA splicing. Cell 23: 643–646. PMID: 7226224
45. Nguyen HD, Yoshihama M, Kenmochi N (2006) Phase distribution of spliceosomal introns: implications for intron origin. BMC Evol Biol 6: 69. PMID: 16959043
46. Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33: W465–467. PMID: 15980513

47. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797. PMID: 15034147

48. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739. doi: 10.1093/molbev/msr121 PMID: 21546353

49. Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6: 526–538. PMID: 2677599