Mol. Cells *33*, 385-392, April 30, 2012
DOI/10.1007/s10059-012-0003-8

Molecules
and
Cells

A Genome-Wide Comparison of NB-LRR Type of Resistance Gene Analogs (RGA) in the Plant Kingdom

Jungeun Kim[1,2], Chan Ju Lim[1], Bong-Woo Lee[1], Jae-Pil Choi[1], Sang-Keun Oh[1], Raza Ahmad[1], Suk-Yoon Kwon[1], Jisook Ahn[1,2], and Cheol-Goo Hur[1,2,*]

Plants express resistance (*R*) genes to recognize invaders and prevent the spread of pathogens. To analyze nucleotide binding site, leucine-rich repeat (NB-LRR) genes, we constructed a fast pipeline to predict and classify the *R* gene analogs (RGAs) by applying in-house matrices. With predicted ~37,000 RGAs, we can directly compare RGA contents across entire plant lineages, from green algae to flowering plants. We focused on the highly divergent NB-LRRs in land plants following the emergence of mosses. We identified entire loss of Toll/Interleukin-1 receptor, NB-LRR (TNL) in *Poaceae* family of monocots and interestingly from *Mimulus guttatus* (a dicot), which leads to the possibility of species-specific TNL loss in other sequenced flowering plants. Using RGA maps, we have elucidated a positive correlation between the cluster sizes of NB-LRRs and their numbers. The cluster members were observed to consist of the same class of NB-LRRs or their variants, which were probably generated from a single locus for an *R* gene. Our website (http://sol.kribb.re.kr/PRGA/), called plant resistance gene analog (PRGA), provides useful information, such as RGA annotations, tools for predicting RGAs, and analyzing domain profiles. Therefore, PRGA provides new insights into *R*-gene evolution and is useful in applying RGA as markers in breeding and or systematic studies.

INTRODUCTION

Immunity protects eukaryotes against pathogens or exotic molecules that are released into host cells. Both animals and plants have developed innate immunities to recognize pathogen-associated molecular patterns (PAMPs), such as bacterial flagellins, lipopolysaccharides, fungal chitin, oomycete Pep-13 and heptaglucosides (Zipfel and Felix, 2005). Animals contain several classes of pattern-recognition receptors (PRRs) that are capable of recognizing distinct PAMPs and directly activating immunity-related cells (Ausubel, 2005; Zipfel and Felix, 2005). Similar to animals, plants contain PRRs to mediate innate immunity, called PAMP triggered immunity (PTI) (Jones and Takemoto, 2004; Zipfel and Felix, 2005). To combat innate immunity, pathogens have developed various strategies to interfere with PAMP-related signal transduction, such as the direct secretion of effectors into host cells (McDowell and Simon, 2008). Plants, which have no specialized immune cell, should be able to autonomously recognize effectors. In response to effector proteins released from pathogens, plants maintain a large number of *R* genes that directly or indirectly recognize effectors and initiate effector triggered immunity (ETI). In addition, plants have evolved *R* genes carrying hyper-variable genetic diversities to recognize a plethora of effectors via gene duplication, sequence exchanges and diversifying selection (McDowell and Simon, 2008). This hyper-variability of *R* genes contributes to recognize effector proteins. Because of the strong relationship between *R* genes functions and their complexities in the plant genome, a comparison of the *R* gene contents among different plants species will provide useful information for practical applications and important discoveries.

Despite disease resistances that is imparted by R genes, these gene products can be categorized into five main classes on the basis of domain organization, except for a few *R* genes (Martin et al., 2003). There are seven domains that are involved in R-proteins: Toll/Interleukin-1 receptor (TIR), coiled-coil (CC), leucine zipper (LZ), nucleotide-binding site (NBS), leucine rich repeat (LRR), transmembrane (TM) and serine-threonine kinase (STK) domains (Martin et al., 2003; van Ooijen et al., 2007). As described above, R-protein and PRR are determined by their recognition molecules of effectors and PAMPs, respectively. Therefore, we classified all structured resistance gene analogs (RGA), including major groups i) CNL (CC-NBS-LRR), ii) TNL (TIR-NBS-LRR), iii) receptor-like kinases (RLK, LRR-TM-STK), iv) receptor-like proteins (RLP, LRR-TM), v) Pto (cytosolic STK) and seven NB-LRR variants (see method). NB-LRRs (CNL and TNL groups) are known to have several variants such as TIR-NBS (TN) and CC-NBS (CN), both of which exhibit losses of the LRRs from the TNL and CNL structures, respectively (Dangl and Jones, 2001; Meyers et al., 2002). Despite the currently unknown functions of these NB-LRR vari-

[1]Green Bio Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea, [2]Bioinformatics, University of Science and Technology, Daejeon 305-350, Korea
*Correspondence: hurlee@kribb.re.kr

Received January 4, 2012; revised February 8, 2012; accepted February 9, 2012; published online March 26, 2012

Keywords: NB-LRR, NB-LRR cluster, plant genome, RGA, RGA map

ants, their expressions have been verified to occur at low levels, indicating their functionality (Meyers et al., 2002). Therefore, genome-wide discoveries of RGAs would help facilitate a better understanding of complex recognition systems against various ranges of effectors and/or PAMPs that invade plants.

Today, the majority of cloned *R* genes encode NB-LRR proteins, over seventy NB-LRR coding genes have been reported with known resistance specificities from 5 major RGA groups (Moffett, 2009). The NB-LRRs are evolved to recognize more specific effector proteins and mediate a "high impact" defense responses, a type of programmed cell death known as the hypersensitive response (HR); whereas membrane bounded PRRs recognize MAMP and mediate "low impact" defense response (reviewed in Moffett, 2009). In addition, NB-LRR-encoding genes are one of the largest and most variable gene families found in plants. Their copy numbers also vary between species as well as within species (Yang et al., 2006; Zhang et al., 2010). This variability enables NB-LRR to recognize broad spectra of effector proteins. Therefore, it is important to construct database to identify majority of NB-LRRs including their variants in entire plant kingdom. It would be helpful to characterize and compare NB-LRR proteins in whole plant lineage.

Thanks to current sequencing technologies, genomes of 22 plants have been published or released, and, consequently, approximately 800,000 putative genes needed to be annotated. Therefore, an efficient pipeline is required to automatically identify and classify RGAs that are obtained from high-throughput sequences. There are two specialized *R*-gene databases available at the moment. First, the NIBLRR (http://niblrrs.ucdavis.edu/) was developed to predict all NB-LRRs and variants and show their genomic distributions in *Arabidopsis thaliana* (Ath) (Meyers et al., 1999; 2003). In contrast, the PRGdb (http://prgdb.cbm.fvg.it/) was constructed with experimentally validated *R*-genes and predicted *R*-genes that originated from public databases (Sanseverino et al., 2010). Unfortunately, these databases do not consider plant genomes, not compare R-gene distributions in plants by considering whole genomic levels.

To provide genome-wide RGA information of plant kingdom, we constructed the PRGA database (http://sol.kribb.re.kr/PRGA/), which can be used to promptly identify RGAs and classify them into 12 groups and/or subgroups considering NB-LRR variants. The PRGA database provides 27,875 RGAs for the entire plant kingdom ranging from green algae to flowering plants. These RGA data show the dynamic distribution of CNL and TNL groups, including NB-LRR creations, expansions and contractions in plant lineage, whereas other classes (RLK, RLP, and Pto etc.) are widely distributed. RGA maps show a positive relationship between *R*-gene abundance and genomic organization, specific to NB-LRR clusters. In addition, the PRGA also stores 8,912 RGAs that have been predicted using 47 UniGene sets for the *Streptophyta* lineage, which provide snapshots of *R*-gene distributions for those plants whose genomic sequences are not available yet. In addition, the PRGA provides web-based RGA prediction and sequence comparisons. Therefore, the PRGA website is helpful for analyzing conserved and variable regions of *R*-genes and for designing probes for *R*-gene marker candidates in various plants.

MATERIALS AND METHODS

Sequence resources

The protein sequences and general feature format (gff) files for 22 genome sequenced plants were downloaded from each genome sequencing consortium or joint genome institute (JGI) database (Supplementary Table S1). Forty-seven UniGene sets for *Streptophyta* were downloaded from the NCBI repository (http://www.ncbi.nih.gov/).

The construction of hidden markov model (HMM) matrices for RGA-coding domains

To expedite domain identification, we constructed HMM matrices that are specific to RGA domains, such as NBS, LRR, TIR, LZ and STK, by using manually verified domain sequences. Each domain was aligned with clustalW (ver. 2.0.11) and manually verified conserved domains were compared with already published domain profiles. Afterwards, aligned domains were used to build HMM matrices using HMMER (version 2.3.2) with "-f" option for maximum search sensitivities (Eddy, 1998). To reflect different domain profiles of NBS domain, we separately constructed NBS matrices according to their N-terminal signal domains (TIR and non-TIR; CC or LZ) (Meyers et al., 2003; Yang et al., 2008). Therefore, NBS matrices in PRGA system enable to distinguish different types of NBS domains, generated from CNL or TNL, by choosing higher score of NBS profiles (NB_CC, NB_TIR).

RGA prediction and classification

The overall process of RGA prediction and classification is represented in Fig. 1. PRGA system analyzes protein sequences to facilitate domain search against HMM matrices. To achieve this, PRGA automatically translates DNA sequence into 6 frames and returns longest protein sequence as query, for corresponding DNA sequence. For bulk analysis, PRGA reduces RGA candidates by implementing BLASTp against known RGAs, manually identified from UniProt, with an expect value of 1e-4. The sequences meeting this criterion are used as queries to identify RGA-coding domains by implementing hmmPfam against in-house HMM matrices. To guarantee the significance of each domain, we applied different thresholds that were experimentally determined by considering the domain length, coverage and degree of conservation: "1e-20" for NBS, "1e-10" for TIR/LZ, "1e-5" for STK, and "1e-1" for LRR. We used the COILS (Lupas et al., 1991) and TMHMM (Krogh et al., 2001) programs for identification the CC and TM domains by using default option, respectively. In the second step, putative RGAs were grouped into twelve RGA groups and/or subgroups based on presence or absence of specific domains. Two pairs of NB-LRR variants (NL_{TIR} and NL_{CC}, N_{TIR} and N_{CC}) contain the same domain organization but are differently classified by score when searched against NBS matrices.

Accuracy test

To represent proportion of actual positives, we statistically measured sensitivity of PRGA classification by comparing classification of curated proteins. We downloaded protein sequences and their RGA class that code for R proteins from the PRGdb (Sanseverino et al., 2010). With 94 queries, as gold-standard, we implemented the web-based RGA prediction and compared RGA classes. The equation for sensitivity is as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Where, TP and FN are true positives and false negatives, respectively. The TP and FN measure the number of predicted RGAs in this study which are correctly or differently classified comparing with the RGA class in PRGdb, respectively.

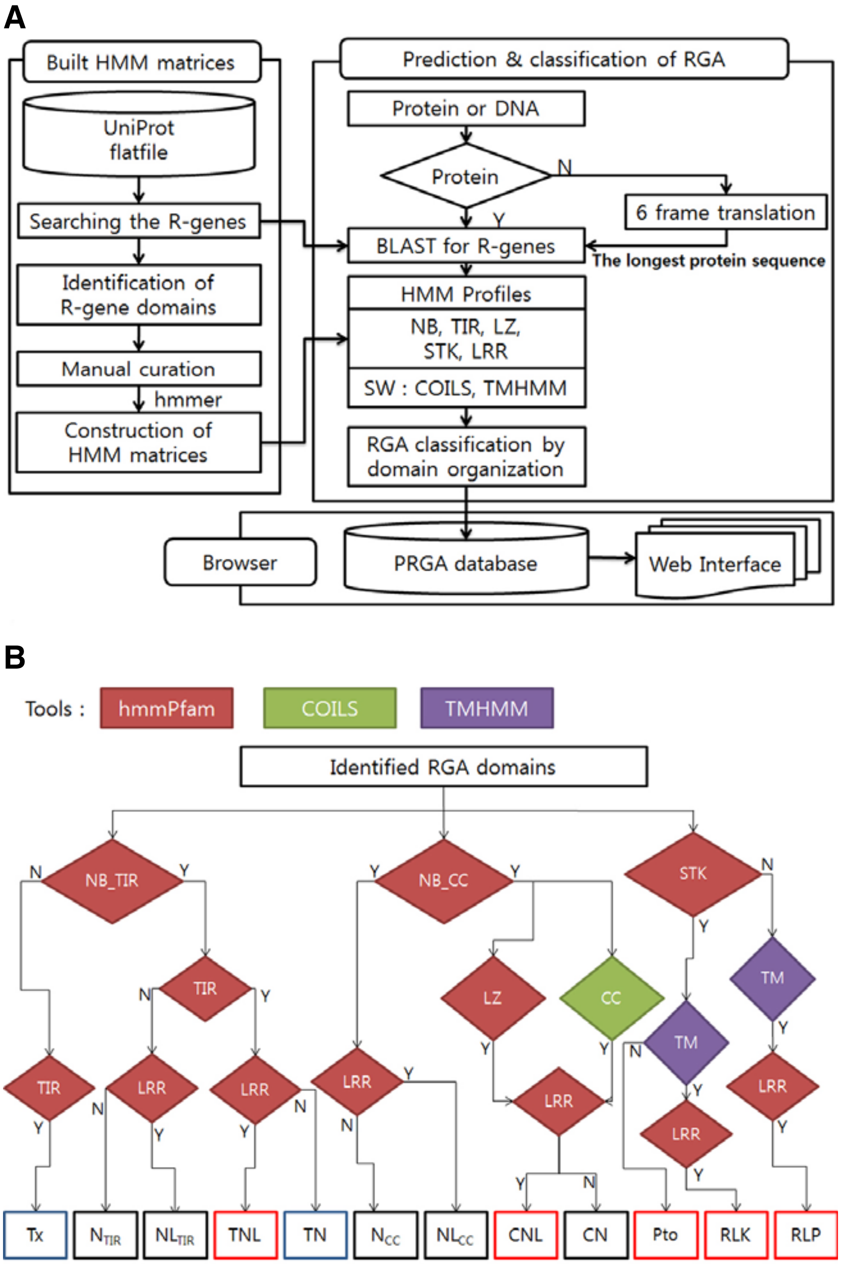

Fig. 1. The overall processes of RGA prediction and classification. (A) The PRGA system is composed of 3 parts: the building of HMM matrices, RGA prediction and classification, and web browser. "SW" denotes software. (B) RGAs were classified on the basis of presence or absence of their functional domains, such as TIR, CC, LZ, NBS, LRR, STK and TM domains. The red box was predicted *via* hmmPfam by searching against in-house HMM matrices (Eddy, 1998). The green and purple boxes were predicted by COILS (Lupas et al., 1991) and TMHMM (Krogh et al., 2001), respectively. We separately built the NBS domains using their N-terminal TIR and CC domains, which are called as "NBS_TIR" and "NBS_CC," respectively. The NBS domain is classified into a different class by higher scoring between "NBS_TIR" and "NBS_CC." "Y" and "N" denotes the presence and absence of the *R*-gene coding domain, respectively. Among the 12 RGA classes, the red boxes correspond to the five major RGA classes, the blue box represents RGAs that have been analyzed at the expression level, and the black box denotes RGAs whose results have not been experimentally confirmed by any other researcher.

RGA map

The RGA map represents the physical location of the R-proteins with protein coordinates in the gff files, while subgroups of *R*-genes are represented with different lines and color codes.

RESULTS

Construction of RGA prediction system and database

We constructed a fast system that is suitable for predicting (Fig. 1A) and classifying RGAs (Fig. 1B) from high-throughput sequences or web-based services. We constructed a database that stores classified RGAs by retrieving several databases having genomic resources (Supplementary Table 1). A total of 27,875 RGAs were predicted from 22 whole genome sequenced plants, representing a global RGA distribution landscape among plant species (Supplementary Fig. 1; Table 2). In addition, 8,912 RGAs were predicted using 47 UniGene sets from the *Streptophyta* lineage, which gives a snapshot of RGAs for those plant species whose full genome sequence is not available yet (Supplementary Table 3).

To validate accuracy of PRGA, we compared the classification of 94 known RGAs downloaded from the PRGdb (Sanseverino et al., 2010). Eighty-four of the RGAs were classified into the same class as in the PRGdb, 7 of the RGAs were classified into their subgroups, and 3 could not match with PRGdb due to low domain conservation (Supplementary Table 4). Most of differences were found in CNL group because its CC domain has been modified a lot (Supplementary Table 4). It has been reported that many of CC domains cannot be characterized with coiled-coil prediction programs because of modifications of few motifs within this domain excepting the conserved "EDVID" motif or combined them with other motifs such as *Solanaceous*

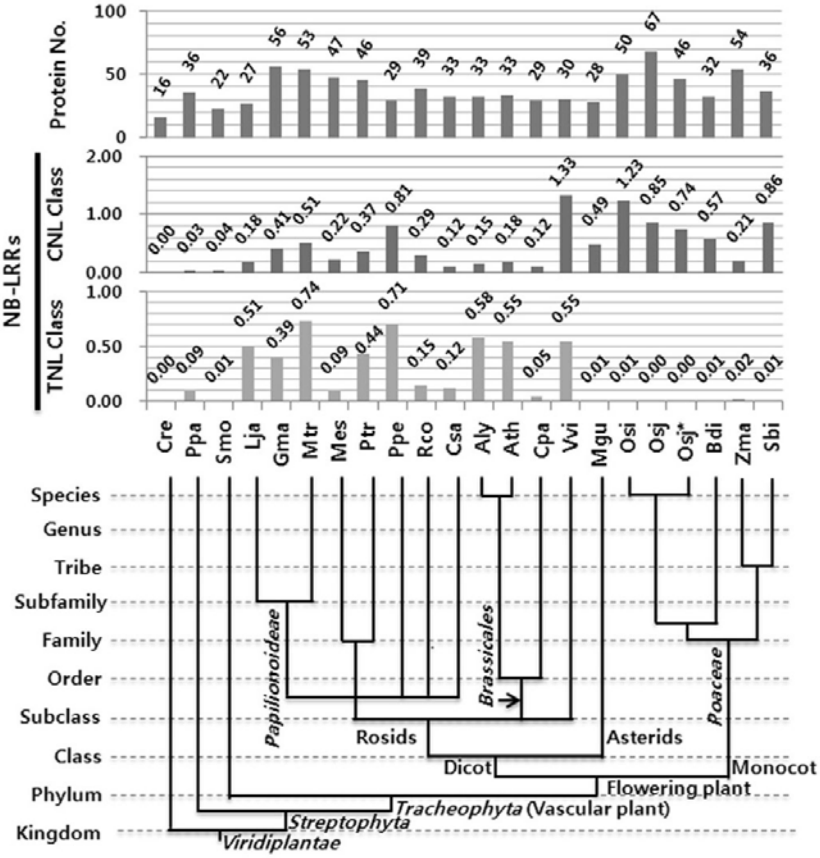

Fig. 2. The NB-LRR distribution in genome-sequenced plants. As a major *R*-gene class, we represented the distribution of the CNL and TNL classes that belongs to the NB-LRRs. The numbers of y-axis represent total number of proteins, which is shown in thousands, in each genome (top) and percentage of CNL and/or TNL class in their genome. The frequencies of NB-LRRs in plant kingdom distributed with rage from 0.0% (Cre) to 1.3% (Vvi) and from 0.0% (Cre) to 0.74% (Mtr) for CNL and TNL, respectively. Other classes are represented in Fig. S1. The abbreviation of each species name can be found in Supplementary Table S1. Osj* represents *Oryza sativa japonica* genome assembled by Syngenta. The taxonomic relationship is as followed by the NCBI taxonomy database.

domain (SD) or "BED" DNA binding domain (Moffett, 2009; van Ooijen et al., 2007). A comparative analysis of RGAs in *Arabidopsis thaliana* (Ath) whole genome yielded 30 NB-LRRs and 58 RLKs more RGAs than previously reports (Meyers et al., 2003; Shiu and Bleecker, 2001); however, we were unable to compare RLP and Pto because of the lack of previous reports. In detail, we compared the number of RGAs in each NB-LRRs subclass with those of classifications in reported by Meyers et al. (2003) and identified 7, 15, 3, 2 and 12 more CNLs, NL_{TIR}, N_{TIR}, TN and Tx proteins, respectively. These results indicated that PRGA identifies TNL subclasses more sensitive than CNL subclasses. Moreover, all of newly identified RGAs were structurally verified. We also found several RGAs were not identified in this study because of low coverage of conserved domain. Moreover, the variations arise in numbers of NB-LRR in this study may be due to the variations in Arabidopsis (Ath) sequence available to previous researchers and to us. Therefore, PRGA system is useful to analyze NB-LRR type of RGA. However, we couldn't compare RLK class because RLK protein sequences or IDs were not reported by Shiu and Bleecker (2001).

NB-LRR evolution in the plant lineage

To directly compare NB-LRR distributions among plants, we evaluated NB-LRR frequencies for each class using the percentage of genomic content in each species (Fig. 2). This comparison demonstrated the dynamic distributions of the CNL and TNL classes, whereas the other classes (RLP, RLK and Pto) were evenly distributed among plants (Supplementary Fig. 1). Both the CNL and TNL classes have been involved with NB-LRRs as major *R*-genes, which increases the relevance of focusing on the genetic architectures of NB-LRRs (Young, 2000). In addition, these dynamic patterns of NB-LRRs require in-depth examination to understand NB-LRR evolution and recognition mechanisms in different plant species. In this study, NB-LRRs were not identified from green algae (*Chlamydomonas reinhardtii*, Cre) but identified from moss, *Physcomitrella patens* (Ppa) and *Selaginella moellendorffii* (Smo) (Fig. 2) and other flowering plants. We could identify less numbers of NB-LRRs in moss as compared to flowering plants; however, there are significant variations in numbers of NB-LRRs even in closely related species (Fig. 2). TNL distributions were especially more variant than those of CNLs (Fig. 2, Supplementary Table 2). TNLs are frequently expanded in the Rosids but did not appear in the *Poaceae* family and *Mimulus guttatus* (Mgu, belonging to the Asterids). These dynamic frequencies of TNLs are visible in closely related species (Fig. 2). *Carica papaya* (Cpa) contains 10-times less TNLs in comparison to Ath or *Arabidopsis lyrata* (Aly), while both belong to *Brassicales* order. Additively, two-fold difference in TNLs was observed in *Glycine max* (Gma) and *Medicago truncatula* (Mtr), while both are in the *Papilionoideae* subfamily.

The NB-LRR distribution in plants lacking whole genome sequences

The currently available genomic resources lack gymnosperm sequences and have weak evidence for Asterids (only Mgu sequence is available); hence, a conclusive discussion of NB-LRR evolution in the entire plant lineage is challenging. To overcome these limitations, we added 8,912 RGAs that were

Table 1. The *R*-gene distribution that was predicted using the UniGene set*

Resources	Species	Abb.	No. of seq[d]	CNL (%)	TNL (%)	RLK (%)	RLP (%)	Pto (%)
UniGene	[a]*Picea glauca*	Pgl	22,472	11 (0.05)	25 (0.11)	2 (0.01)	34 (0.15)	142 (0.63)
UniGene	[a]*Picea sitchensis*	Psi	19,828	13 (0.06)	46 (0.23)	10 (0.05)	28 (0.14)	93 (0.47)
UniGene	[a]*Pinus taeda*	Pta	18,079	2 (0.01)	25 (0.14)	2 (0.01)	16 (0.09)	104 (0.58)
UniGene	[b]*Oryza sativa*	Os	40,971	371 (0.91)	1 (0.00)	268 (0.65)	137 (0.33)	510 (1.24)
Genome	[b]*Oryza sativa indica*	Osi	49,710	611 (1.23)	5 (0.00)	382 (0.77)	165 (0.33)	781 (1.57)
Genome	[b]*Oryza sativa japonica*	Osj	67,393	575 (0.85)	3 (0.00)	456 (0.68)	182 (0.27)	390 (1.38)
Genome	[b]*Oryza sativa japonica*	Osj*	45,824	337 (0.74)	2 (0.00)	263 (0.57)	103 (0.23)	521 (1.14)
Genome	[b]*Brachypodium distachyon*	Bdi	32,255	183 (0.57)	2 (0.00)	249 (0.77)	86 (0.27)	733 (2.27)
Genome	[b]*Zea mays*	Zma	53,764	112 (0.21)	10 (0.00)	289 (0.54)	93 (0.17)	1,045 (1.94)
Genome	[b]*Sorghum bicolor*	Sbi	36,338	314 (0.86)	3 (0.00)	258 (0.71)	110 (0.30)	686 (1.89)
Genome	[c]*Mimulus guttatus*	Mgu	27,501	136 (0.50)	2 (0.00)	204 (0.74)	88 (0.32)	634 (2.31)
UniGene	[c]*Artemisia annua*	Aan	9,461	0 (0.00)	2 (0.03)	0 (0.00)	4 (0.04)	87 (0.92)
UniGene	[c]*Capsicum annuum*	Can	8,987	11 (0.12)	0 (0.00)	2 (0.02)	12 (0.13)	0 (0.00)
UniGene	[c]*Coffea canephora*	Cca	4,148	3 (0.07)	1 (0.02)	0 (0.00)	3 (0.07)	1 (0.02)
UniGene	[c]*Helianthus annuus*	Hv	23,594	43 (0.18)	0 (0.00)	20 (0.08)	31 (0.0)	203 (0.86)
UniGene	[c]*Solanum lycopersicum*	Les	36,455	29 (0.08)	3 (0.01)	42 (0.12)	57 (0.16)	252 (0.69)
UniGene	[c]*Lactuca sativa*	Lsa	7,939	1 (0.01)	2 (0.03)	2 (0.03)	19 (0.24)	99 (1.25)
UniGene	[c]*Nicotiana tabacum*	Nta	19,752	4 (0.02)	7 (0.04)	8 (0.04)	33 (0.17)	193 (0.93)
UniGene	[c]*Solanum meloongena*	Sml	8,218	9 (0.11)	4 (0.05)	0 (0.00)	36 (0.19)	148 (0.79)
UniGene	[c]*Solanum tuberosum*	Stu	18,783	18 (0.10)	15 (0.08)	4 (0.02)	36 (0.19)	148 (0.79)

*In this table, we present some of the RGAs that were predicted using the UniGene set. All RGA distributions are represented in Supplementary Table S3 and take into account their subgroups using 47 UniGene sets for plants.

[a]These species are from the gymnosperms.

[b]The representative species is part of the *Poaceae* family and exhibits a loss of the R-protein in the TNL group.

[c]These species are part of the Asterids.

[d]The number of UniGene sequences that were used in this research.

predicted using 47 UniGene sets (Supplementary Table 3). These data were used to estimate NB-LRR losses in each plant under study. To analyze the existence or loss of NB-LRR in each plant, we compared each class of NB-LRRs to TNLs in *Oryza sativa* (Os) (Table 1). The number of TNLs in Os can provide criteria to determine CNL or TNL losses in each species because there are enough UniGenes to represent entire transcripts, moreover, we already know about the complete loss of TNLs in this species (Monosi et al., 2004; Zhou et al., 2004). We also represented RGAs from different plants using genomic or UniGene data (Table 1). This data give an insight of TNL loss from *Poaceae* families (monocots) and only in Mgu, which is dicot. The rare TNL's presented in these species cannot be considered as true TNL's because they have lost either one or more domains compared to full length TNL (Supplementary Table 2). Keeping this in view, we analyzed RGA distribution from UniGene sets of three gymnosperms (Table 1; Supplementary Table 3). Interestingly, there was higher number of TNL in gymnosperms than Os or in other TNL-loss plants having genome-wide sequence data (Table 1). In addition, full-length CNLs and TNLs were also identified in *Picea sitchensis* (Psi) and *Pinus taeda* (Pta) (Supplementary Table 3), whereas all of the TNLs identified in TNL loss-plants are partial fragments of TNLs (Supplementary Tables 2 and 3), analyzed in this study. These data led us to conclude that both CNLs and TNLs may have functions in gymnosperms.

Considering the expansion of NB-LRRs in all land plants, the *Poaceae* family exhibits a missing TNL class that is specific to their lineage (Monosi et al., 2004; Zhou et al., 2004). In addition, we identified loss of TNLs in Mgu that belong to the Asterids subclass (Fig. 2; Table 1; Supplementary Table 2) and a member of dicots. Therefore, we focused on TNL loss in Asterids using UniGene sets (Table 1; Supplementary Table 3). We identified seven full-length TNLs in the *Solanaceae* family; however, other Asterids have similar patterns of TNL numbers as found in Os (Supplementary Table 3). Therefore, some Asterids, such as *Solanaceae*, contain functional TNLs, whereas others do not. These findings cannot be considered conclusive because UniGene sets couldn't represent the entire genome. However, we investigated 9 Asterids and identified full length TNLs only in *Solanaceae* families. We identified 19 and 13 full length CNLs in *Helianthus annuus* (HV) and *Lactuca sativa* (Les) with 23,594 and 36,455 UniGenes, respectively, whereas even a single full length TNL was not identified (Supplementary Table 3). Therefore, we hypothesize that TNLs are relatively poorly distributed or have been species (or family) specifically lost in Asterids. However, additional sequence data from this family is required to validate this hypothesis.

Contribution of the NB-LRR cluster to R-gene evolution

In RGA maps, we presented the physical location of RGAs on chromosomes with different color codes for RGA groups (http://sol.kribb.re.kr/PRGA/Genomic_protein.php). *Vitis vinifera* (Vvi) was selected as a model and its map has been presented because it has highest RGA numbers among the plants that were studied in this report (Fig. 3). The RGA map for Vvi depicted that the NB-LRR clusters are concentrated in specific chromosomal regions (Chromosome 9, 12, 13, 18) whereas the other

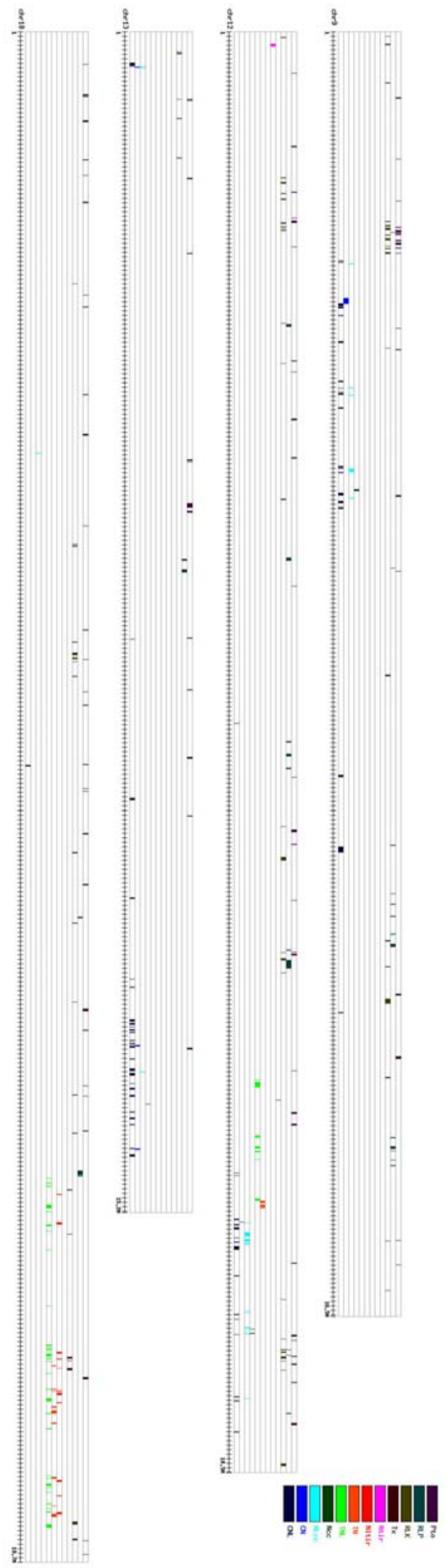

Fig. 3. Chromosomes mapping of RGAs for grape. This map represents the physical locations of the RGAs in grape on the chromosome classifying domain organizations. The chromosomes appear as anchored and oriented super contigs based on reference genetic markers.

classes (RLK, RLP and Pto) are widely distributed throughout the chromosomes (Fig. 3, Supplementary Results). Among NB-LRR clusters, CNL and TNL clusters are distantly appeared and their variants are co-occurred with their full-length CNL or TNLs. Therefore, we calculated the correlation of co-occurrence between full-length NB-LRRs and their subgroups (Supplementary Table 5). Therein, we observed a high correlation, with an average (avg.) of 0.81, and observed stronger correlation in TNLs (avg. 0.97) than CNLs (avg. 0.66). We also analyzed the relationship between the NB-LRRs cluster size and their abundance (Supplementary Results, Supplementary Table 6). A cluster size is defined as the number of NB-LRRs that the neighboring NB-LRRs appear within 200 Kbp. We found that the plants that had larger numbers of NB-LRRs exhibited larger cluster sizes. For example, Mtr has the largest TNL rate (0.74%) and 50 (13%) TNL members or their variants are located on Chr. 6.

DISCUSSION

PRGA database

Advances in sequencing technologies have increased the number of available whole genome sequences (Supplementary Table 1). This wealth of data challenges researchers who are interested in screening R proteins from high-throughput sequences (Chen et al., 2010; Kohler et al., 2008; Li et al., 2010; Meyers et al., 2003; Porter et al., 2009; Yang et al., 2008; Zhou et al., 2004). Despite the existence of two currently available *R* gene databases, we constructed a new robust pipeline to predict and classify RGAs from high-throughput genomic and transcriptomic data (Figs. 1A and 1B). This pipeline divides RGAs into 5 classes and 7 subclasses of NB-LRR variants which facilitate the analysis of RGAs using segmental sequences, such as ESTs (Fig. 1A). We constructed PRGA website (http://sol.kribb.re.kr/PRGA) to provide RGA information (Supplementary Fig. 2), as well as services for RGA prediction and domain profiling (Supplementary Fig. 3). It is unique system classifying RGAs from user's inputs as well as providing domain profiles of sequences assembled in this database. Especially, domain profiling of user's query can provide domain features suitable to design the degenerate primers to amplify potential *R*-genes (Supplementary Fig. 3). Therefore, the PRGA can be utilized to analyze various features of RGAs and can assist to design studies based on RGA markers, map-based cloning, and RGA-based breeding.

The contribution of NB-LRR clusters to their expansion

To recognize various effectors, it is important to consider the hyper-variability of R-proteins. Many biological processes can confer variability to NB-LRRs, including mutation, duplication, recombination, unequal crossing over and gene conversion (Michelmore and Meyers, 1998). In this context, NB-LRR clusters, wherein allelic series are generated from one locus, have been studied to explain the increment of recognition capacity of ever changing pathogen effectors (Michelmore and Meyers, 1998; Parniske et al., 1997). To identify the genome wide relationship of NB-LRR clusters and variability, we analyzed the genomic distribution of each type of NB-LRRs from more than 20 genome-sequenced plants. In RGA maps, CNL and TNL clusters are separated from one another, whereas CNL/TNL variants are co-clustered with their full-length CNLs/TNLs in grape (Fig. 3). In RGA map for Vvi, the rates of the co-occurrences of full-length CNLs/TNLs and their variants are highly correlated at an avg. of 0.66 and 0.97, respectively (Supplementary Result, Supplementary Table 5). In addition, it has been demonstrated that frequent unequal crossing-over events had mediated genetic variation within the *R* gene clusters in lettuce (Kuang et al., 2004). The NB-LRR variants might have emerged by the deletion of one or more domains during the unequal cross-over. This development suggests that domain rearrangements may frequently occur during unequal cross-over; therefore, NB-LRR variability has been increased. In addition, we found positive correlation of NB-LRR abundance with cluster size (Supplementary Table 6). NB-LRR abundant plants exhibited larger number of clusters or larger number of NB-LRRs in the clusters. In this study, we investigated representative RGA maps that were generated from 22 plants and analyzed the relationship between NB-LRR abundance and cluster size/number. These data demonstrate that sequence crossing-over may be the primary reason of increased *R*-gene variability in individual species.

The appearance of NB-LRRs in the plant kingdom

Previous genome wide R-protein studies have demonstrated different distributions of CNL and TNL classes (Li et al., 2010; Meyers et al., 2003; Zhou et al., 2004) with major focus on TNL abundance in dicots while little attention was paid to their existence in monocots (Li et al., 2010; Zhou et al., 2004). Rare presence of TNL such as TN or Tx in monocots supports the idea that TNL has been lost specifically during monocot evolution (Meyers et al., 2002). Therefore, it is important to identify full-length TNLs from ancestor lineages to explain loss of TNL from present day monocots (Zhou et al., 2004). To find any clue, we identified more than 37,000 RGAs in the entire plant lineage, from green algae to flowering plants. This large-scale analysis demonstrated the appearance of full-length CNLs and TNLs in mosses (which were not identified in green algae) that evolved into seed plants (gymnosperms and angiosperms) (Fig. 2; Table 1). As a common ancestor of the seed plants, presence of full-length CNLs and TNLs in mosses clearly demonstrates the evolutionary inheritance of TNLs and CNLs to common ancestor of angiosperms and gymnosperms. However, TNLs were lost by *Poaceae* family (monocots) during the long evolutionary process while, it retained the CNLs evolved from common ancestor of angiosperms (Table 1, Supplementary Tables 2 and 3). Thus, these data lead to the possibility of loss or degeneration of the TNLs in monocots after differentiation of gymnosperms and angiosperms. In addition to monocots, we identified entire loss of TNLs from *Mimulus guttatus* (a dicot belonging to Asterids; Fig. 2), which leads to the possibility of species-specific TNL loss in dicot plants. By analyzing NB-LRR distributions in the entire plant lineage, we can suggest that full-length CNLs and TNLs evolved and firstly appeared in mosses and that there is a possibility of a complete loss of TNLs at the species level.

Dynamic evolutionary patterns of NB-LRRs in flowering plants

We analyzed the dynamic patterns of NB-LRRs among flowering plants (Fig. 2). We found that NB-LRR contents are independent of the plant's evolutionary relationships even among related plants. These dynamic distributions highlight how plants defend against various pathogens with limited numbers of NB-LRRs. In a gene-for-gene concept, NB-LRRs interact with corresponding effector proteins that are released from pathogens and inhibit the pathogen's growth (Bonas and Van den Ackerveken, 1999). For this hypothesis to be correct, plants are required to have as many NB-LRR repertories as the number of pathogen diversities. Researchers have reported evidence of R protein expansions and divergences using various methods, including phylogenetic analysis (Chen et al., 2010; Li et al., 2010; Yang et al., 2008), frequent substitutions (Li et al., 2010; Yang et al., 2008), copy number variations (Zhang et al., 2010) and the presence/absence of genes among closely related species (Chen et al., 2010; Li et al., 2010). This NB-LRR variability easily explains the evolution of R-proteins when defending against diverse pathogens; however, some plants (Mes, Rco, Csa, Mgu and Zma), approximately one-fourth of the flowering plants that were investigated in this study, contain smaller numbers of NB-LRRs, in the range of 0.2-0.4%, whereas NB-LRRs make up more than 1% of the genomic content of other plants (Mtr, Ppe, Vvi and Osi) (Fig. 2). These decreases in the numbers of NB-LRRs are not consistent with the gene-for-gene concept because plants have far less NB-LRRs to cope every effector protein. Porter et al. (2009) discussed these small numbers of NB-LRR in Cpa with a preference of indirect recognition mechanisms, which can recognize more diverse pathogen repertories (Porter et al., 2009); however, it is still unknown if differentially preferential defense mechanisms exist in R-protein abundant and R-protein limited plants. Li et al. (2010) have suggested that it is a more cost efficient mechanism to preserve the least number of *R*-gene copies, as seen in *Zea mays* (Zma), and maintain a balance in the presence and absence of natural enemies (Li et al., 2010). It is still not known whether having many R-proteins or least number of R-protein in the presence of new infecting pathogens is most efficient. Therefore, dynamic NB-LRRs distributions in plants provide a new challenge to understanding the mechanisms of different *R* gene contents.

In this study, we developed pipeline to predict RGAs from the high-throughput sequencing data applying BLAST and HMM. This database stores RGAs from 22 genome sequenced plants and 47 UniGene sets. In spite of their usefulness to understand global landscape of plant RGAs, these bioinformatics based methods are significantly influenced by the genome sequence coverage, sequence assembly accuracy and/or the source of inquiry sequences. As an example of this study, the numbers of the identified RGAs are significantly different; although both Osj and Osj* sequences came from the same genotype, Nipponbare, suggesting that the results are significantly affected by the above mentioned factors (Fig. 2; Supplementary Fig. 1). The PRGA is the only database arisen with whole genome sequenced plants. This amount of data set is useful in inferring the evolutionary history of NB-LRRs based on copy number variation and genomic distributions of NB-LRRs. However, the current data could not compare copy number variations between variants or cultivars; because the available genomic sequences are from single genotype or cultivar. In fact, a recent report showed the number of NB-LRRs varied among different cultivars of a species by several folds (Zhang et al., 2010). These variations of NB-LRRs appeared not only among the species native to different geographical regions or different ecotypes but also between wild species and cultivated ones (Zhang et al., 2010). It requires NB-LRR comparison within species in the near future.

In summary, the PRGA has limitation that prediction results are affected by sequencing qualities or query sequences. And current version of PRGA cannot compare NB-LRRs between variants and cultivars. Because of these limitations, user needs to consider the current status of genome sequencing and annotation which is used in this study. However, the PRGA is the only database having strength to see just at a glance the variation of NB-LRRs in whole plant lineage and easily analyze NB-LRRs from high-throughput sequencing data. These strengths may make it an attractive tool in circumstances where many sequences are being released quite rapidly for analyzing genomic and transcriptomic data.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This work was supported by grants from the Crop Functional Genomics Center (CFGC) and the Microbial Genomics & Applications Center (MGAC), which is one of the 21st Century Frontier Research Programs of the Ministry of Education, Science, and Technology (MEST), South Korea.

REFERENCES

Ausubel, F.M. (2005). Are innate immune signaling pathways in plants and animals conserved? Nat. Immunol. *6*, 973-979.

Bonas, U., and Van den Ackerveken, G. (1999). Gene-for-gene interactions: bacterial avirulence proteins specify plant disease resistance. Curr. Opin. Microbiol. *2*, 94-98.

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

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

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

Jones, D.A., and Takemoto, D. (2004). Plant innate immunity - direct and indirect recognition of general and specific pathogen-associated molecules. Curr. Opin. Immunol. *16*, 48-62.

Kohler, A., Rinaldi, C., Duplessis, S., Baucher, M., Geelen, D., Duchaussoy, F., Meyers, B.C., Boerjan, W., and Martin, F. (2008). Genome-wide identification of NBS resistance genes in Populus trichocarpa. Plant Mol. Biol. *66*, 619-636.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567-580.

Kuang, H., Woo, S.S., Meyers, B.C., Nevo, E., and Michelmore, R.W. (2004). Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. Plant Cell *16*, 2870-2894.

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

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science *252*, 1162-1164.

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

McDowell, J.M., and Simon, S.A. (2008). Molecular diversity at the plant-pathogen interface. Dev. Comp. Immunol. *32*, 736-744.

Meyers, B.C., Dickerman, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sobral, B.W., and Young, N.D. (1999). Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J. *20*, 317-332.

Meyers, B.C., Morgante, M., and Michelmore, R.W. (2002). TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in Arabidopsis and other plant genomes. Plant J. *32*, 77-92.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell *15*, 809-834.

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

Moffett, P. (2009). Mechanisms of recognition in dominant R gene mediated resistance. Adv. Virus Res. *75*, 1-33.

Monosi, B., Wisser, R.J., Pennill, L., and Hulbert, S.H. (2004). Full-genome analysis of resistance gene homologues in rice. Theor. Appl. Genet. *109*, 1434-1447.

Parniske, M., Hammond-Kosack, K.E., Golstein, C., Thomas, C.M., Jones, D.A., Harrison, K., Wulff, B.B., and Jones, J.D. (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.

Porter, B.W., Paidi, M., Ming, R., Alam, M., Nishijima, W.T., and Zhu, Y.J. (2009). Genome-wide analysis of Carica papaya reveals a small NBS resistance gene family. Mol. Genet. Genomics *281*, 609-626.

Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., Frusciante, L., and Ercolano, M.R. (2010). PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic Acids Res. *38*, D814-821.

Shiu, S.H., and Bleecker, A.B. (2001). Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. Proc. Natl. Acad. Sci. USA *98*, 10763-10768.

van Ooijen, G., van den Burg, H.A., Cornelissen, B.J., and Takken, F.L. (2007). Structure and function of resistance proteins in solanaceous plants. Annu. Rev. Phytopathol. *45*, 43-72.

Yang, S., Feng, Z., Zhang, X., Jiang, K., Jin, X., Hang, Y., Chen, J.Q., and Tian, D. (2006). Genome-wide investigation on the genetic variations of rice disease resistance genes. Plant Mol. Biol. *62*, 181-193.

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

Young, N.D. (2000). The genetic architecture of resistance. Curr. Opin. Plant Biol. *3*, 285-290.

Zhang, M., Wu, Y.H., Lee, M.K., Liu, Y.H., Rong, Y., Santos, T.S., Wu, C., Xie, F., Nelson, R.L., and Zhang, H.B. (2010). Numbers of genes in the NBS and RLK families vary by more than four-fold within a plant species and are regulated by multiple factors. Nucleic Acids Res. *38*, 6513-6525.

Zhou, T., Wang, Y., Chen, J.Q., Araki, H., Jing, Z., Jiang, K., Shen, J., and 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.

Zipfel, C., and Felix, G. (2005). Plants and animals: a different taste for microbes? Curr. Opin. Plant Biol. *8*, 353-360.