A THESIS FOR THE DEGREE OF MASTER OF SCIENCE

NLR-Finder: An Easy and Efficient Annotation Tool for the NLR Superfamily in Plant Genomes

식물 유전체에서 병 저항성 유전자군을 동정하기 위한 생물정보 프로그램 개발

FEBRUARY, 2017

JIEUN PARK

INTERDISCIPLINARY PROGRAM IN AGRICULTURAL GENOMICS

COLLEGE OF AGRICULTURE AND LIFE SCIENCES

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

NLR-Finder: An Easy and Efficient Annotation Tool for the NLR Superfamily in Plant Genomes

UNDER THE DIRECTION OF DR. DOIL CHOI
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF
SEOUL NATIONAL UNIVERSITY

BY
JIEUN PARK

MAJOR IN HORTICULTURAL CROP GENOMICS
INTERDISCIPLINARY PROGRAM IN AGRICULTURAL GENOMICS
THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

FEBRUARY, 2017

APPROVED AS A QUALIFIED THESIS OF JIEUN PARK
FOR THE DEGREE OF MASTER OF SCIENCE
BY THE COMMITTEE MEMBERS

CHAIRMAN _____
Jin Hoe Huh, Ph.D.

VICE-CHAIRMAN _____
Doil Choi, Ph.D.

MEMBER _____
Yong-Hwan Lee, Ph.D.

NLR-Finder: An Easy and Efficient Annotation Tool for the NLR Superfamily in Plant Genomes

JIEUN PARK

INTERDISCIPLINARY PROGRAM IN AGRICULTURAL GENOMICS

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

ABSTRACT

Gene annotation is an essential process to identify gene structures and define biological functions. It is an important step for subsequent analyses including gene cloning and identification of genes for agricultural traits. However, current gene annotation misrepresents the whole gene repertoire due to biased gene model construction. Nucleotide-binding and leucine-rich repeat (NLR) superfamily is one of the poorly annotated gene families in plants. The NLR family tends to be clustered in genomes by segmental and tandem duplications, which makes the gene annotation challenging. The NLR-Finder was developed for unbiased genome-wide identification of the NLR superfamily in assembled plant genomes. The NLR-Finder firstly detects candidate NLR gene regions by extending 30 kb to both sides of all the identified NB-ARC domain

regions. Secondly, evidence-based NLR genes are predicted by aligning published proteins and transcriptome sequences to the candidate gene regions. Thirdly, additional NLR genes are extracted using an *ab initio* prediction approach. Lastly, final NLR gene models are generated by integration of the evidence- and *ab initio*-based NLR genes. The re-annotation was performed using the NLR-Finder on 17 different plant genomes. On average, public annotation tools identified about 310 genes, whereas the NLR-Finder annotated about 497 genes. In *Gossypium hirsutum* and *Vigna radiata*, the number of re-annotated genes tripled compared to that of publicly available data. The re-annotated genes were successfully validated by comparing with high-quality annotations of *Arabidopsis thaliana, Brachypodium distachyon,* and *Solanum lycopersicum*. This study demonstrated that the NLR-Finder provides an easy-to-use and efficient method to annotate the NLR superfamily in plant genomes.

Keywords**:** annotation, *ab initio* gene prediction, HMMER, nucleotide-binding and leucine-rich repeat (NLR) gene, protein mapping

Student number: 2015-21804

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ETI	Effector-triggered immunity
EVM	Evidence modeler
HMMER	Hidden Markov model (HMM) search method
HR	Hypersensitive response
NLR	Nucleotide-binding and leucine rich repeat
PAMP	Pathogen-associated molecular pattern
PCD	Programmed cell death
PTI	PAMP-triggered immunity

INTRODUCTION

A mass of genome sequences from prokaryote to eukaryote have been accumulated as consequence of improvement in DNA sequencing technologies. However, the gene annotation is still inaccurate and challenging. Many studies have pointed out that annotated genes have been released prematurely and misrepresented the whole gene repertoire (Devos and Valencia, 2001; Gilks *et al.*, 2002; van den Berg *et al.*, 2010; Gotoh *et al.*, 2014). In particular, gene annotation in plant genomes is challenging due to the large genome size and repetitive sequences. Furthermore, the gene contents are also complex, as shown by the presence of large gene families and abundant pseudogenes which are nearly identical sequences derived from recent whole genome duplication events and transposon activity (Schatz *et al.*, 2012). A previous study analyzed annotation quality of 47 plant genomes and reported that 50-60% of annotated gene structures include errors such as inherently fragmented genes in incomplete sequencing regions, and pseudogenes (Gotoh *et al.*, 2014).

Nucleotide-binding and leucine-rich repeat (NLR) superfamily is one of the poorly annotated gene families in plants due to the repetitive nature of the genes (Meyers *et al.*, 2003). Previous studies have shown that public gene prediction software does not detect up to 40% of the total NLR genes (Jupe *et al.*, 2013; Andolfo *et al.*, 2014). The genes are the most representative type of disease resistance genes (Meyers *et al.*, 2003), and contain leucine-rich repeat (LRR) domains in their C-terminal and NB-

ARC (nucleotide-binding adaptor shared by APAF-1, Resistance proteins, and CED-4) domains in central regions (van Ooijen *et al.*, 2008; Seo *et al.*, 2016). NLR genes can be classified into two types, CC-NLR and TIR-NLR, based on the presence of an N-terminal Coiled-coil (CC) motif or Toll/interleukin-I receptor-like (TIR) domain (Eitas and Dangl, 2010). NB-ARC domains are highly conserved and play roles in ATP binding and hydrolysis (Lukasik and Takken, 2009). For the LRR and TIR/CC domains, they are involved in activation and interaction with signaling partners, respectively (Lukasik and Takken, 2009).

In this study, the NLR-Finder was developed as a high-accuracy tool for a NLR superfamily annotation. In order to test the performance of the NLR-Finder, the tool was run with 17 plant genomes. The re-annotated genes were compared to public annotation data. On average, public annotation tools identified about 310 genes, whereas the NLR-Finder annotated about 497 genes. In some species, the number of re-annotated genes tripled compared to that of publicly available data. Annotated genes were validated with proven high-quality gene annotations including *Arabidopsis thaliana, Brachypodium distachyon,* and *Solanum lycopersicum*. This study demonstrated that the NLR-Finder provides an easy-to-use and efficient method to annotate the NLR gene family.

LITERATURE REVIEWS

Gene annotation

Gene annotation is the process of finding the location of genes and determining the biological functions of the genes in assembled genomes. Genome sequencing has become easy to perform, and costs have dramatically fallen by technological advances. A major challenge in the post-genome sequencing era is to obtain reliable annotations of these genomes (Jupe *et al.*, 2013). However, gene annotation is still challenging and complex (Yandell and Ence, 2012). The annotation of eukaryotic genomes is especially more complicated than that of prokaryotic genomes due to larger genome size and the complexity of gene structure. In order to annotate more easily, a lot of public annotation pipelines such as EVM (Haas *et al.*, 2008), GLEAN (Elsik *et al.*, 2007), Maker (Cantarel *et al.*, 2008), and JIGSAW (Allen and Salzberg, 2005) have been developed.

The annotation has multiple steps. The first step in most of annotation pipelines is repeat masking (Yandell and Ence, 2012). The repeat sequences including transposable elements (TEs) as well as simple repeats are identified and masked in this step. Repeat masking is divided into two methods, hard-masking and soft-masking. In hard-masking, complex repeats are completely removed from any further consideration of future phases in the annotation process (Kielbasa *et al.*, 2011), and replaced with the letter N. In soft-masking, repeats with low-complexity are transformed to lowercase

letters so that this prevents alignment programs from using repeats as seeds but allows alignments in non-masked regions to extend into the soft-masked regions (Cantarel *et al.*, 2008). In the second step, proteins, ESTs, and transcriptome are aligned to the assembled genome using alignment tools such as Exonerate (Slater and Birney, 2005), GeneWise (Birney and Durbin, 2000), and Bowtie (Langmead *et al.*, 2009). The results of alignments are filtered by percent identity or percent similarity (Yandell and Ence, 2012). In the step of *ab initio* gene prediction, construction of training sets is the most important because the gene predictor depends on the number and the variety of genes in the training sets to find genes in assembled genomes (Goodswen *et al.*, 2012).

Annotation errors

Annotation quality is one of the most important factors for many subsequent analyses since most analyses are performed based of the annotation. However, a lot of studies have argued that majority of published gene annotations are still low in quality. A previous study reported that more than half of annotated genes in plants include variety of annotation errors (Gotoh *et al.*, 2014). The gene models contain pseudo- and fragmented genes annotated in low quality areas of sequencing (Gotoh *et al.*, 2014). Additionally, the annotation error rate is at least 8% in *Mycoplasma genitalium* (Brenner, 1999). However, the rate was calculated via comparison of annotations generated by three different groups. This indicates that the error rate was estimated without consideration of innate error in the annotation. Most importantly, there are errors in

public databases which are used when annotating genes. These errors will exacerbate the issue since these databases will inevitably lead to further errors, making the post-annotation analysis unreliable. Even one of the most commonly used databases, Gene Ontology (GO) sequence database (GOSeqLite), contains errors ranging from a rate of 28% to 30% (Jones *et al.*, 2007).

Even though annotation accuracy is steadily improving, there are many critical errors that still need to be corrected. To reduce the error of annotation, re-annotation should be performed using the updated data (van den Berg *et al.*, 2010). It should be avoided to use databases including errors for protein mapping or *ab initio* prediction. However, if an errorless database is unavailable, one should be sure to always use the most up-to-date data in order to achieve the best results (van den Berg *et al.*, 2010).

Nucleotide-binding and leucine-rich repeat (NLR) genes

In nature, plants are attacked by diverse pathogens such as fungi, bacteria, viruses, and insects. To protect themselves from these harmful pathogens, plants use defense mechanisms they have developed over millions of years of evolution. The defense systems can be divided into two layers (Dodds and Rathjen, 2010). In the first layer, pathogen-associated molecular patterns (PAMPs) of pathogens are recognized by pattern-recognition receptors (PRRs) located in plasma membranes (Dangl *et al.*, 2013). In consequence, PAMP-triggered immunity (PTI) is induced to limit microbial colonization. To suppress PTI, some pathogens deliver virulence proteins known as

effectors into host cells (Dangl *et al.*, 2013). In response, plants that have the second layer of defense recognize the effectors and induce a strong immune response called effector-triggered immunity (ETI) to suppress pathogen growth. As a result of ETI responses, programmed cell death called hypersensitive response (HR) are often observed (Fei *et al.*, 2016). The second layer is governed by intracellular resistance (R) genes. Most of the R genes belong to the nucleotide-binding and leucine-rich repeat (NLR) gene family (Glowacki *et al.*, 2011).

As the name suggests, these genes include nucleotide-binding (NB) domains and leucine-rich repeat (LRR) domains. NLR genes are divided into two subclasses, CC-NLR and TIR-NLR, based on the presence of an N-terminal Coiled-coil (CC) motif or Toll/interleukin-I receptor-like (TIR) domain (Eitas and Dangl, 2010). Plant NLR genes are generally known to detect pathogens through direct or indirect interaction, although the precise mechanism of the genes remains an open question (DeYoung and Innes, 2006). In NLR genes, different domains have specific roles. For example, a NB domain possess ATP binding and hydrolysis capabilities, while CC/TIR plays a role in interaction with signaling partners, and LRR domains are implicated in the activation of partners (Lukasik and Takken, 2009).

The repetitive structure of the gene causes difficulty to annotate the genes (Steuernagel *et al.*, 2015). Additionally, the genes tend to cluster in genomes by segmental and tandem duplications (McHale *et al.*, 2006), which makes the gene annotation more challenging. The number of NLR genes in many genomes tends to be

underestimated, and the genes need to be improved through re-annotation (Jupe *et al.*, 2013).

MATERIALS AND METHODS

Plant genomes, protein data, and transcriptome collection

In order to represent the whole clades of the plants, the genomes were evenly chosen from monocot to dicot. Eight plant genomes (*Arabidopsis thaliana, Brachypodium distachyon, Glycine max, Gossypium raimondii, Oryza sativa, Solanum lycopersicum, Vitis vinifera,* and *Zea mays*) were selected for re-annotation first by the selection criteria; 1) pseudomolecules and transcriptome data are available, 2) scaffold N50 is more than 1 Mb, 3) contig N50 is around 30 kb. To test performance of the pipeline, nine other genomes were additionally chosen (*Ananas comosus, Brassica oleracea, Capsicum annuum, Citrullus lanatus, Citrus sinensis, Gossypium hirsutum, Solanum tuberosum, Vigna angularis,* and *Vigna radiata*). The species were annotated with EVM (Haas *et al.*, 2008), GLEAN (Elsik *et al.*, 2007), and Maker (Cantarel *et al.*, 2008), which are commonly used as annotation pipelines, and compared their annotations with gene models of the NLR-Finder after re-annotation. Genome fasta files and protein data were downloaded from Phytozome (https://phytozome.jgi.doe.gov/) except for *C. annuum* and *C. lanatus*. In-house genome and protein data were used for *C. annuum*. For *C. lanatus*, data were downloaded from the Cucurbit Genomics Database (http://www.icugi.org/cgi-bin/ICuGI/index.cgi) (Table 1).

Transcriptome data were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). These were paired sequence data, and the read lengths were over 80 bp. Bacterial

Table 1. Genomic resources used to test the performance of the NLR-Finder.

Species	Genome	Protein		Reference
	Assembled size (Mb)	Version	# of genes	
Ananas comosus	526	V3.0	27,024	(Ming *et al.*, 2015)
Arabidopsis thaliana	115	TAIR10	27,416	(Arabidopsis Genome Initiative, 2000)
Brachypodium distachyon	270	V3.1	34,310	(International Brachypodium Initiative, 2010)
Brassica oleracea	630	V1.0	35,400	(Liu *et al.*, 2014)
Capsicum annuum	3,060	V1.6	34,897	(Kim *et al.*, 2014)
Citrullus lanatus	354	V1.0	23,440	(Guo *et al.*, 2013)
Citrus sinensis	367	V1.0	29,406	(Xu *et al.*, 2013)
Glycine max	955	Wm82.a2.v1	56,044	(Schmutz *et al.*, 2010)
Gossypium hirsutum	2,500	V1.1	70,478	(Zhang *et al.*, 2015)
Gossypium raimondii	775	V2.1	37,505	(Wang *et al.*, 2012)
Oryza sativa	390	V7.0	42,189	(Goff *et al.*, 2002)
Solanum lycopersicum	782	iTAGv2.3	34,727	(Sato *et al.*, 2012)
Solanum tuberosum	727	V3.4	35,119	(Potato Genome Sequencing Consortium, 2011)
Vigna angularis	612	V3.0	26,857	(Kang *et al.*, 2015)
Vigna radiata	543	V6.0	22,368	(Kang *et al.*, 2014)
Vitis vinifera	498	Genoscope.12X	26,346	(Jaillon *et al.*, 2007)
Zea mays	2,048	6a	63,480	(Schnable *et al.*, 2009)

sequences, duplicated short reads, and low-quality sequences below Q20 (quality score) were filtered out in the preprocessing step. Bacterial genomes from GenBank were used for reference, and Bowtie2 v2.0.0-beta7 (--local -D 15 -R 2 -N 0 -L 20 -i S,1,0.65) was used for mapping sequences to reference bacteria genomes. To eliminate low-quality sequences, in-house Perl scripts were used. After preprocessing, sequences were assembled using TopHat v2.0.12 and Cufflinks v2.2.1 (Ghosh and Chan, 2016) with a default parameter (Table 2).

Workflow of the NLR-Finder

Gene annotation was carried out using the NLR superfamily annotation pipeline (NLR-Finder). The NLR-Finder consists of five steps: 1) identification of candidate gene regions, 2) domain search, 3) identification of candidate NLR gene regions, 4) structural annotation, 5) gene model integration/filtering (Figure 1).

Step 1: Identification of candidate gene regions

The NLR-Finder performed a six-frame translation from a genomic FASTA data to identify candidate gene regions. After the translated sequences were cut by stop-codon, the fragments were generated in FASTA format.

Step 2: Domain search

HMMERv.3 (Finn *et al.*, 2011), was used to search all NB-ARC domains in

Table 2. Properties of transcriptome used for gene annotation.

Species	Raw data (Gb)	Preprocessed data (Gb)	Tissue	SRR ID
Ananas comosus	18	11	Whole plant, leaf	SRR1165179, SRR2983923, SRR2984781
Arabidopsis thaliana	66	42	Flower, leaf, root, seed, stem	SRR633726, SRR633727, SRR1773569, SRR1773570, SRR1773572, SRR1773573, SRR2187609
Brachypodium distachyon	10	7	Shoot, stem	SRR1635409, SRR1797575
Brassica oleracea	13	7	Flower, leaf, root, stem	SRR630923, SRR630924, SRR630925, SRR630927
Capsicum annuum	33	14	Fruit, root, stem	In-house resources
Citrullus lanatus	20	7	Flower, bud, stem	SRR494474, SRR494479, SRR2033940, SRR2033941, SRR2033942, SRR2033943
Citrus sinensis	20	8	Fruit, leaf	SRR867166, SRR867397, SRR867425, SRR867435
Glycine max	36	14	Flower, leaf, root, seed, stem	SRR1174205, SRR1174207, SRR1174216, SRR1174219, SRR1174226
Gossypium hirsutum	23	16	Leaf, root, stem	SRR2081039, SRR2081040, SRR2081042, SRR2081045
Gossypium raimondii	17	7	Seed, leaf, ovule	SRR389181, SRR389182, SRR389183, SRR959890, SRR959899
Oryza sativa	64	22	Ear, leaf, panicle, root	SRR1179192, ERR855945, ERR855947, DRR013723
Solanum lycopersicum	14	8	Fruit, root, leaf, bud, flower	SRR1514810, SRR3031978, SRR3031982
Solanum tuberosum	14	3	Flower, leaf, root, tuber	SRR122109, SRR122122, SRR122124, SRR1207290
Vigna angularis	40	24	Flower, leaf, root, stem	DRR031872, DRR031873, DRR031876, DRR031877
Vigna radiata	18	8	Whole plant, seed	SRR1653637, SRR1867748
Vitis vinifera	36	20	Flower, leaf, pooled RNA	SRR519455, SRR519456, SRR520374, SRR522298, SRR2043222
Zea mays	28	10	Leaf, ovule, pollen, root, shoot	SRR254171, SRR255405, SRR445651, SRR445656, SRR2886947

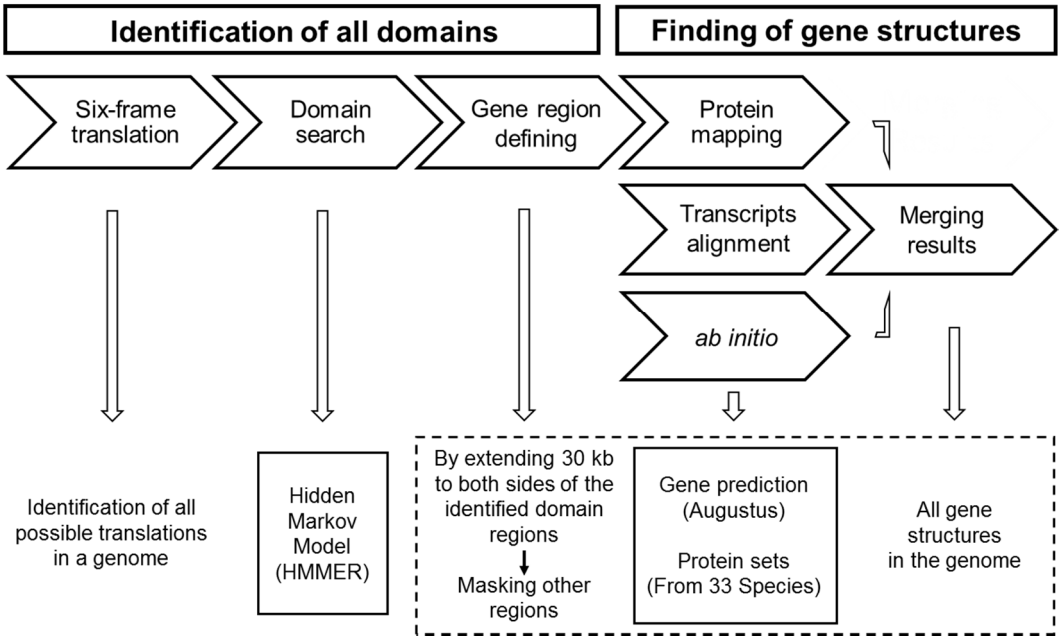


Figure 1. Workflow of the NLR-Finder.

the results of Step 1. To build training sets, thirty NB-ARC domains with the least e-values were extracted from public protein data of the genome using Pfam database (PF00931). Alignment of the nucleotide sequences of the domains was generated with MEGA 6 (default parameter) (Tamura *et al.*, 2013) using ClustalW algorithm (Larkin *et al.*, 2007), and the results were converted from FASTA format to Stockholm format in IBIVU (http://www.ibi.vu.nl/). The aligned NB-ARC domain sequences were used as the training sets to run HMMERv.3. The domains identified in this step were not filtered by any other methods such as e-value cut-off to avoid missing any candidates.

Step 3: Identification of candidate NLR gene regions

The average NLR gene length is 3.2 kb and the longest one is approximately 20 kb. To mask unnecessary parts in genomes and reduce computing time, the NLR-Finder defined candidate NLR gene regions by extending 30 kb to both sides of the identified NB-ARC domain regions, and mask other regions without the domain.

Step 4: Structural annotation

This step is divided into three parts: 1) protein mapping, 2) transcripts alignment using Integrated Structural Gene Annotation Pipeline (ISGAP) (Kim *et al.*, 2015), 3) *ab initio* gene prediction. From 33 plant genomes, 9,557 NLR genes (average length: 882 bp) were collected to perform the protein mapping. To find gene structures, the NLR genes were aligned to the masked genome of Step 3 using Exonerate v.2.2.0

with parameters –percent 50 and –maxintron 20000 (Slater and Birney, 2005). Gene models having early stop-codons were filtered. In the gene models without stop-codons, 3 bp from the 3'-end region were extended to find the stop-codons. Consensus sequences were then constructed from the extended and partial gene models by merging same sequences. To remove redundant gene models in the same locus of the genome, the longest full-type genes among the consensus gene models were extracted as representative gene models. If a full-type gene was not found, the longest partial gene was selected as the representative model. For transcripts alignment, reference assembly was performed using masked genomes of Step 3 and followed the ISGAP method of previous study (Kim *et al.*, 2015). To build a training set of Augustus, an *ab initio* gene prediction program, NLR genes of the species identified by Pfam database (PF00931) were integrated with full-type genes generated by the protein mapping and transcripts alignment. After then, the Augustus was run using the training set.

Step 5: Gene model integration/filtering

After the structural annotation, the gene models without the NB-ARC domains identified in Step 2 were filtered out, and then integrated into the final gene model. The merging order is as followed: 1) full-type NLR genes of the protein mapping, 2) NLR genes annotated in the transcripts alignment, 3) the results of the Augustus, 4) partial-type NLR genes of the protein mapping. Finally, the final gene model was filtered with HMMERv.3 by using the same training sets of Step 2 to determine surely the gene

model containing the NB-ARC domains.

RESULTS

Transcriptome raw data preprocessing and reference assembly

Transcriptome sequences were obtained from 17 species, and the sequence size ranged from 10 to 66 Gb (Table 2). This data were collected from diverse tissues such as flower, fruit, leaf, root, seed, and stem. The data were paired-end reads, and their read lengths were more than 80 bp. After preprocessing, the sequences ranging from 3 to 42 Gb remained and were used for reference assembly (Table 2). For details, see the materials and methods section.

Re-annotation of NLR genes with the NLR-Finder

In order to test the performance of the NLR-Finder, the pipeline was run on 17 plant genomes. Among the 17 species, genomes of *Arabidopsis thaliana, Brachypodium distachyon,* and *Solanum lycopersicum* were used for validation (see below). A six-frame translation was conducted in the genomes to find candidate gene regions. All candidate NB-ARC domains were then identified in the translated sequences using HMMERv.3. *Citrullus lanatus* contained the smallest number of the domains with 88 domains, whereas *Gossypium hirsutum* had the largest number of the domains with 1,861 domains (Table 3). On average, there were 858 NB-ARC domains in the 14 genomes.

After defining candidate gene regions using the position of all the NB-ARC

Table 3. Numbers of identified NB-ARC domains and annotated NLR genes.

Species	# of NB[a] domains	# of NB[a] domains		# of NLR genes	
	HMMERv.3	NLR-Finder	Public data[b]	NLR-Finder	Public data[b]
Ananas comosus	299	280	233	218	174
Brassica oleracea	530	429	203	301	155
Capsicum annuum	1,856	1,391	1,071	910	766
Citrullus lanatus	88	78	71	50	46
Citrus sinensis	1,288	1,082	673	778	516
Glycine max	938	747	668	528	476
Gossypium hirsutum	1,861	1,355	434	994	316
Gossypium raimondii	803	677	418	471	303
Oryza sativa	953	893	820	595	539
Solanum tuberosum	921	732	449	556	343
Vigna angularis	401	351	150	286	104
Vigna radiata	544	468	143	369	96
Vitis vinifera	1,201	984	455	704	326
Zea mays	329	289	270	194	177

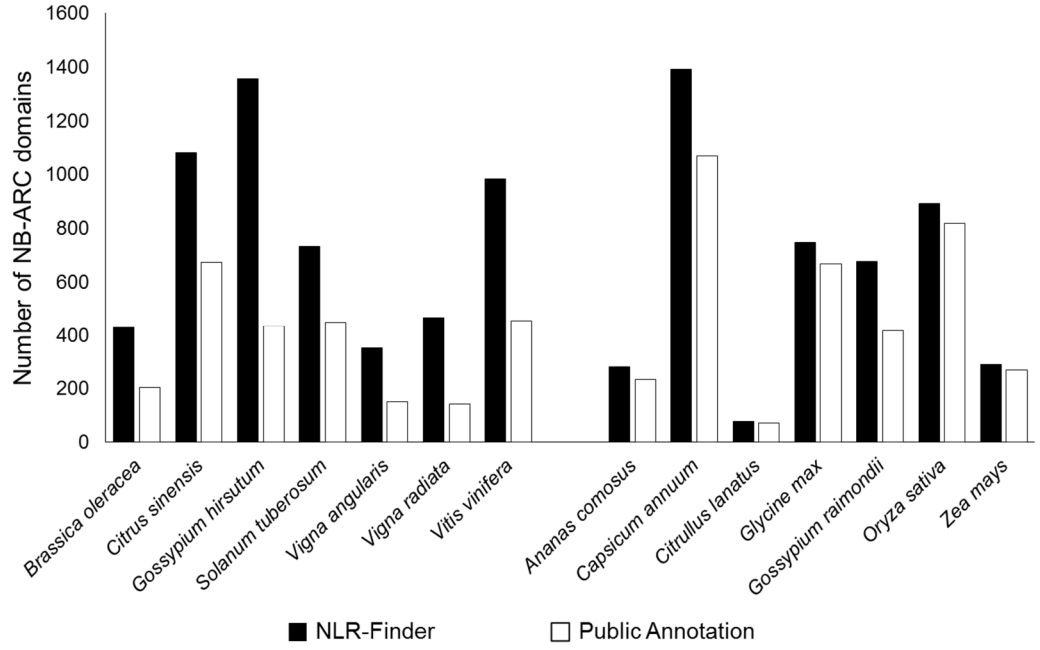
[a]NB: NB-ARC domain.

[b]Public data: NLR genes including NB-ARC domains among publicly available data.

domains, the structural annotation was performed with protein mapping, transcripts alignment, and *ab initio* gene prediction. All gene models were merged into final gene models and filtered by HMMERv.3 to remove genes which have no NB-ARC domains. To compare the gene models with publicly available data in equal conditions, public annotation data without the NB-ARC domain identified in the domain search step were filtered out. Afterwards, the genes having no NB-ARC domains were filtered out once again by HMMERv.3 using the same training sets used in the domain search step. The filtering methods of public annotation data were same for Step 5 of the NLR-Finder in materials and methods.

The number of NB-ARC domains identified by the NLR-Finder ranged from 78 to 1,391, and 71 to 1,071 in the public data (Table 3). The NLR-Finder identified an average of 264 domains more than the domains of the publicly available data (Figure 2A). For NLR genes, the number annotated by the NLR-Finder was immensely diverse from 50 to 994 (Table 3). In the publicly available data, the number of identified NB-ARC domains ranged from 46 to 766. The NLR-Finder found more annotated genes ranging from 4 to 678 (Figure 2B), and on average, annotated about 187 more genes. In *G. hirsutum*, *Vigna angularis*, and *Vigna radiata*, the number of re-annotated genes tripled compared to that of publicly available data (Figure 2B). Additionally, the NLR-Finder could generally detect longer NLR genes compared to the pre-existing annotated genes, even though the length of genes annotated by the NLR-Finder was shorter than public data in some species (Table 4).

(A)


Number of NB-ARC domains
1600
1400
1200
1000
800
600
400
200
0
Brassica oleracea
Citrus sinensis
Gossypium hirsutum
Solanum tuberosum
Vigna angularis
Vigna radiata
Vitis vinifera
Ananas comosus
Capsicum annuum
Citrullus lanatus
Glycine max
Gossypium raimondii
Oryza sativa
Zea mays
NLR-Finder
Public Annotation

(B)

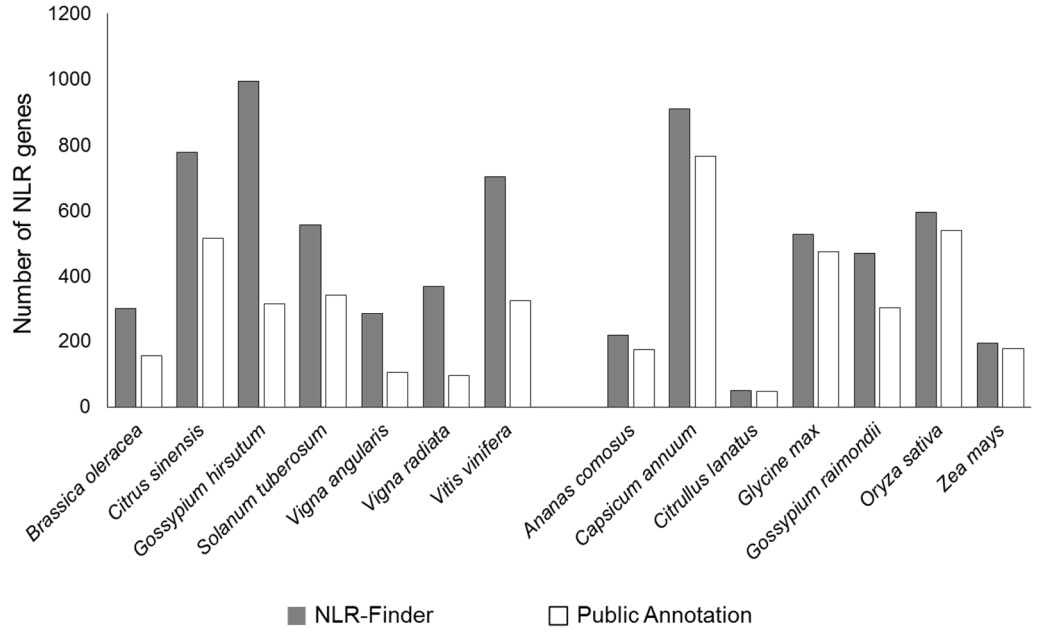

Number of NLR genes
1200
1000
800
600
400
200
0
Brassica oleracea
Citrus sinensis
Gossypium hirsutum
Solanum tuberosum
Vigna angularis
Vigna radiata
Vitis vinifera
Ananas comosus
Capsicum annuum
Citrullus lanatus
Glycine max
Gossypium raimondii
Oryza sativa
Zea mays
NLR-Finder
Public Annotation

Figure 2. The NLR-Finder identifies more NB-ARC domains and NLR genes compared to those of publicly available annotations. (A) The bar graph labeled in black indicates the number of identified NB-ARC domains by the NLR-Finder. The white bars show the number of NB-ARC domains in publicly available data. (B) The grey bar graph shows the number of annotated NLR genes by the NLR-Finder. The white bars indicate the number of NLR genes including NB-ARC domains from publicly available data.

Table 4. Lengths and numbers of annotated NLR genes.

Species	NLR-Finder (Average length, bp)	Public data[a] (Average length, bp)
Ananas comosus	218 (1,052)	174 (1,018)
Brassica oleracea	301 (764)	155 (662)
Capsicum annuum	910 (615)	766 (588)
Citrullus lanatus	50 (905)	46 (888)
Citrus sinensis	778 (882)	516 (927)
Glycine max	528 (892)	476 (898)
Gossypium hirsutum	994 (938)	316 (947)
Gossypium raimondii	471 (984)	303 (1,000)
Oryza sativa	595 (899)	539 (880)
Solanum tuberosum	556 (715)	343 (698)
Vigna angularis	286 (978)	104 (1,027)
Vigna radiata	369 (1,017)	96 (1,007)
Vitis vinifera	704 (989)	326 (848)
Zea mays	194 (752)	177 (745)

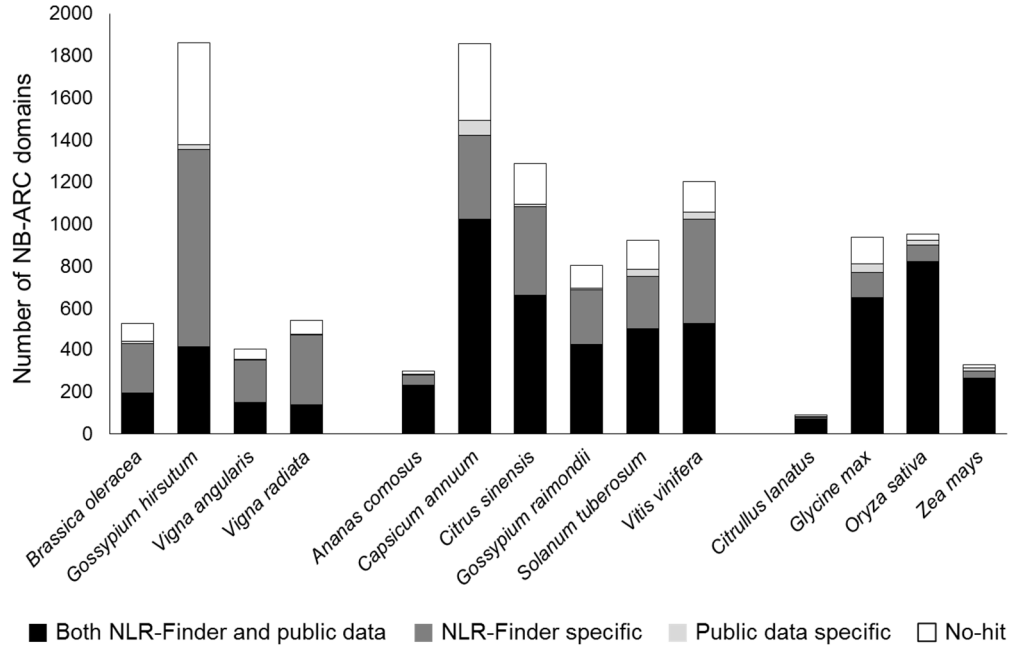
[a]Public data: NLR genes including NB-ARC domains among publicly available data.

For a detailed comparison, the number of the domains identified by the NLR-Finder and public annotation pipelines was calculated in each species with the basis of all candidate domains (100%) identified by HMMERv.3 (Figure 3). Next, overlapped domains discovered by both the NLR-Finder and public annotation tools were confirmed, and specific domains identified by either the NLR-Finder or public annotation pipelines were found among all the candidates. In the 14 species, the number of domains in the "NLR-Finder specific" were greater than those of "Public data specific" (Figure 3A). Regarding *Brassica oleracea, G. hirsutum, V. angularis*, and *V. radiata*, there were even a greater number of domains found in "NLR-Finder specific" than those of "Both NLR-Finder and public data" (Figure 3A). On average, the NLR-Finder identified approximately 707 (83%) of NB-ARC domains among 858 (100%) total domains, while the public annotation tools identified approximately 454 (53%) domains (Figure 3B). The number of the domains searched by both the NLR-Finder and public annotation tools was 433 (51%) among all 858 domains (100%). The number of specifically identified domains was 273 (32%) and 20 (2%) in the NLR-Finder and public annotation tools, respectively.

Validation of the NLR-Finder using high-quality plant genomes

To evaluate the annotations performed by the NLR-Finder, gene models annotated by the NLR-Finder were compared with the public annotations of *A. thaliana*, *B. distachyon,* and *S. lycopersicum*, which are commonly considered as high-quality

(A)


Number of NB-ARC domains
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Brassica oleracea
Gossypium hirsutum
Vigna angularis
Vigna radiata
Ananas comosus
Capsicum annuum
Citrus sinensis
Gossypium raimondii
Solanum tuberosum
Vitis vinifera
Citrullus lanatus
Glycine max
Oryza sativa
Zea mays
Both NLR-Finder and public data
NLR-Finder specific
Public data specific
No-hit

(B)

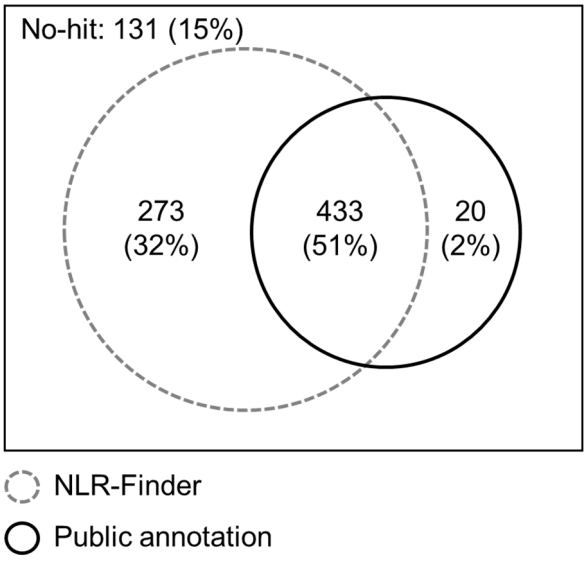

No-hit: 131 (15%)
273
(32%)
433
(51%)
20
(2%)
NLR-Finder
Public annotation

Figure 3. The number of NB-ARC domains found only by the NLR-Finder is greater than the number identified only by public annotation pipelines. (A) "Both NLR-Finder and public data" indicates the number of NB-ARC domains identified by both the NLR-Finder and public annotation pipelines. "NLR-Finder specific" represents domains found only by the NLR-Finder, while "Public data specific" means domains identified only by public annotation pipelines. "No-hit" shows domains unidentified by neither the NLR-Finder nor public annotation pipelines. (B) The Venn diagram shows the average number of the domains from public annotations and the NLR-Finder results in 14 plant genomes.

plant genomes. The number of the NLR genes and NB-ARC domains identified by the NLR-Finder in three species were nearly identical to the genes and domains of public annotation tools (Table 5).

In a more detailed comparison, the number of domains found by the NLR-Finder and public annotation tools were calculated on the basis of all the NB-ARC domains (100%) identified by HMMERv.3 (Figure 4). In *A. thaliana*, 227 domains were identified by both NLR-Finder and public annotation tools. “NLR-Finder specific” contained five domains, and “Public data specific” contained four. “No-hit” domains contained 21. For *B. distachyon*, 680, 25, and 27 were contained in each of “Both NLR-Finder and public data”, “NLR-Finder specific”, and “Public data specific”. “No-hit” domains were 33. For *S. lycopersicum*, the number of domains found by both NLR-Finder and public annotation tools were 355. “NLR-Finder specific”, “Public data specific”, and “No-hit” were 25, 31, and 99 in each. Therefore, even in comparison with the high-quality genomes, the performance of the NLR-Finder was comparable with those of public annotation tools.

In *A. thaliana*, one NLR gene was annotated only by public annotation tools among 171 annotated NLR genes (Figure 5A). The NLR-Finder annotated two novel NLR genes. For *B. distachyon*, seven NLR genes were identified only by public annotation pipelines, whereas 18 novel genes were annotated only by the NLR-Finder (Figure 5B). Regarding *S. lycopersicum*, public annotation tools found specifically 14 NLR genes, and the NLR-Finder annotated 20 novel NLR genes (Figure 5C).

Table 5. Numbers of identified NB-ARC domains and annotated NLR genes in high-quality plant genomes.

Species	# of NB[a] domains	# of NB[a] domains		# of NLR genes	
	HMMER	NLR-Finder	Public data[b]	NLR-Finder	Public data[b]
Arabidopsis thaliana	257	229	229	170	171
Brachypodium distachyon	765	702	701	390	382
Solanum lycopersicum	510	373	374	268	267

[a]NB: NB-ARC domain.

[b]Public data: NLR genes including NB-ARC domains among publicly available data.

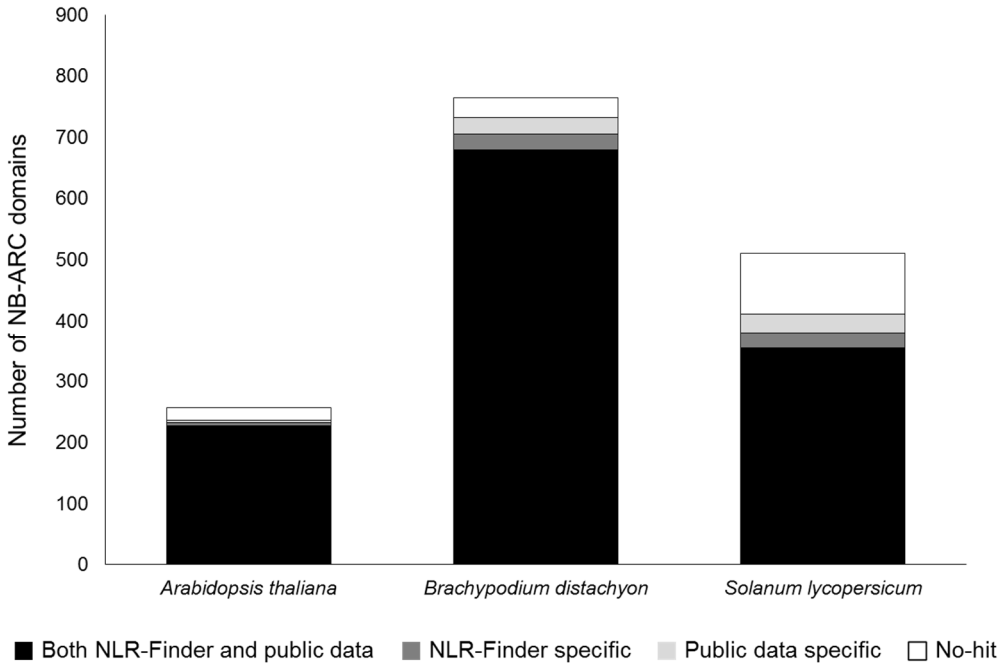
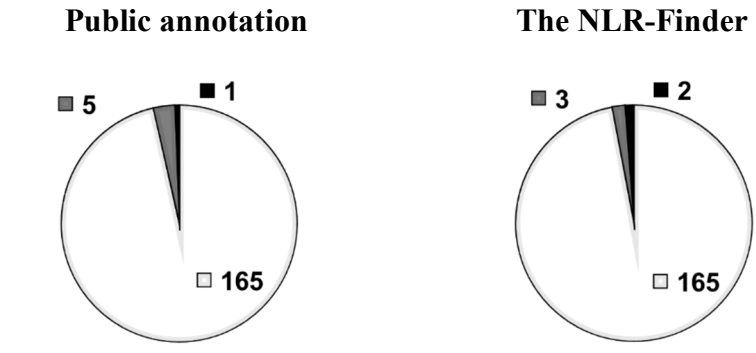
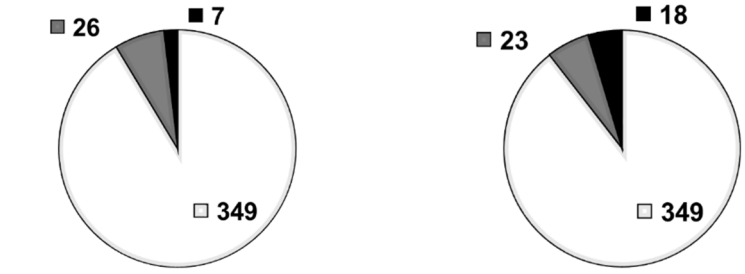


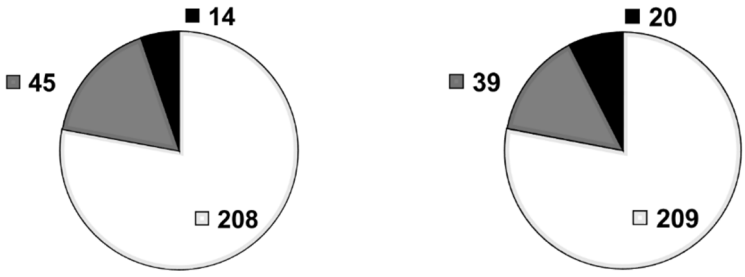
Figure 4. The number of NB-ARC domains identified only by the NLR-Finder is comparable to that found only by public annotation pipelines in three high-quality plant genomes. ”Both NLR-Finder and public data” indicates the number of NB-ARC domains identified by both the NLR-Finder and public annotation pipelines. “NLR-Finder specific” represents domains found only by the NLR-Finder, while “Public data specific” means domains identified only by public annotation pipelines. “No-hit” shows domains not identified by neither the NLR-Finder nor public annotation pipelines.



Arabidopsis thaliana



Brachypodium distachyon



Solanum lycopersicum

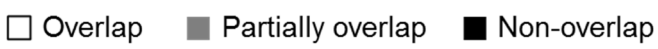


Figure 5. Validation with high-quality plant genomes, *Arabidopsis thaliana, Brachypodium distachyon,* and *Solanum lycopersicum*. ”Overlap” indicates the number of NLR genes annotated by both the NLR-Finder and public annotation pipelines. “Partially overlap” represents the number of NLR genes, which are partially overlapped among the genes annotated by the NLR-Finder and public annotation pipelines. “Non-overlap” shows non-overlapped NLR genes. The pie graph represents the validation results of *A. thaliana, B. distachyon,* and *S. lycopersicum*.

DISCUSSION

By using the NLR-Finder on seventeen plant genomes including three high-quality plant species for validation, annotation results demonstrated that the NLR-Finder provides an effective tool to identify NLR genes in diverse plant genomes of monocot and dicot class. Identification of all candidate domains was performed using HMMERv.3, the domain search program. This is the most important part in the re-annotation process since the pipeline is based on the domains of the NLR gene family for gene annotation. The domain search step was performed to identify NB-ARC domains in genomes and defining potential NLR gene regions from the position of domains. Therefore, all domains identified in the domain search step were not filtered, and potential regions to annotate as NLR gene were not missed.

For an accurate comparison of performance of public annotation tools, it is needed to establish the entire domain number quantity and position within the genome. All the NB-ARC domains found by HMMERv.3 were assumed as whole domains in the genome since it includes all the possible NB-ARC domains. The performance of the NLR-Finder was compared with that of public annotation tools on the basis of the total NB-ARC domains. As a result, there were some NB-ARC domains that were not identified by both the NLR-Finder and public annotation tools. Most of the “no-hit” domains were false-positive domains which cannot be identified even in the protein mapping step in the NLR-Finder.

In the gene models of the NLR-Finder and public annotation tools, the number of NB-ARC domains tends to be overestimated since the number of the domains was counted in the final gene models on the basis of the total NB-ARC domains searched in the domain search step, and the total domains were identified in genomic DNA, not CDS sequence. This method for counting the NB-ARC domains is for performance comparison between the NLR-Finder and public annotation tools based on one criterion. However, the domains in the final gene model were not split like the number of the domains in Table 3.

In a validation analysis performed with three high-quality genomes (Figure 5), non-overlapped genes were found. A non-overlapped NLR gene in public annotation of *A. thaliana* was not identified by the NLR-Finder even in the step of protein mapping. It seems not to be annotated since the splicing sites of the NLR gene are not conserved in the genome. For two non-overlapped NLR genes of the NLR-Finder in *A. thaliana,* NLR homologs were identified in other species. Non-overlapped NLR genes in public annotation of *B. distachyon* and *S. lycopersicum*, there were diverse reasons for not detecting by the NLR-Finder. Some NLR genes were not found in the step of protein mapping. However, a BLAST search using the coding sequence (CDS) revealed that the NLR genes were in the genomes. The public software used in the protein mapping step seems to miss the NLR genes. Other NLR genes were filtered out due to a frame-shift mutation, and the others were not annotated since the splicing sites were not conserved in the genomes. For about 78 percent of non-overlapped NLR genes in the NLR-Finder

of *B. distachyon,* NLR homologs were identified in other species. In *S. lycopersicum,* 95 percent of non-overlapped NLR genes in the NLR-Finder have NLR homologs in other species. For partially overlapped NLR genes, detailed in-depth analysis is needed.

RGAugury (Li *et al*., 2016), NLR-parser (Steuernagel *et al*., 2015) have previously been reported as tools to annotate NLR genes. However, the RGAugury still misses NLR genes which were not annotated in a genome, since the tool uses protein sequence files from a whole genome annotation or manually annotated sequence data as a data input. The NLR-parser can identify NLR genes even though the NLR genes are not in an annotation of a genome, since the input is a protein sequence translated into all six reading frames. The tool annotates NLR genes using Motif Alignment and Search Tool (MAST) (Bailey *et al*., 2009), whereas the NLR-Finder identifies NLR genes using a lot of evidence proteins and transcripts after defining NLR candidate gene regions based on NB-ARC domains. The NLR-Finder would be a useful tool to annotate the NLR gene family and improve the annotation quality in plant genomes.

REFERENCES

Allen, J.E., and Salzberg, S.L. (2005). JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics **21,** 3596-3603.

Andolfo, G., Jupe, F., Witek, K., Etherington, G.J., Ercolano, M.R., and Jones, J.D.G. (2014). Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. BMC Plant Biology **14,** 120.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796-815.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research **37,** 202-208.

Birney, E., and Durbin, R. (2000). Using GeneWise in the *Drosophila* annotation experiment. Genome Research **10,** 547-548.

Brenner, S.E. (1999). Errors in genome annotation. Trends in Genetics **15,** 132-133.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research **18,** 188-196.

Dangl, J.L., Horvath, D.M., and Staskawicz, B.J. (2013). Pivoting the plant immune system from dissection to deployment. Science **341,** 746-751.

Devos, D., and Valencia, A. (2001). Intrinsic errors in genome annotation. Trends in Genetics **17,** 429-431.

DeYoung, B.J., and Innes, R.W. (2006). Plant NBS-LRR proteins in pathogen sensing and host defense. Nature Immunology **7,** 1243-1249.

Dodds, P.N., and Rathjen, J.P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. Nature Reviews Genetics **11,** 539-548.

Eitas, T.K., and Dangl, J.L. (2010). NB-LRR proteins: pairs, pieces, perception, partners, and pathways. Current Opinion in Plant Biology **13,** 472-477.

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S., and Weinstock, G.M. (2007). Creating a honey bee consensus gene set. Genome Biology **8,** R13.

Fei, Q., Zhang, Y., Xia, R., and Meyers, B.C. (2016). Small RNAs add zing to the Zig-Zag-Zig

model of plant defenses. Molecular Plant-microbe Interactions **29,** 165-169.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Research **39,** 29-37.

Ghosh, S., and Chan, C.K. (2016). Analysis of RNA-Seq data using TopHat and Cufflinks. Methods in Molecular Biology **1374,** 339-361.

Gilks, W.R., Audit, B., De Angelis, D., Tsoka, S., and Ouzounis, C.A. (2002). Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics **18,** 1641-1649.

Glowacki, S., Macioszek, V.K., and Kononowicz, A.K. (2011). R proteins as fundamentals of plant innate immunity. Cellular Molecular Biology Letters **16,** 1-24.

Goff, S.A., *et al.* (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science **296,** 92-100.

Goodswen, S.J., Kennedy, P.J., and Ellis, J.T. (2012). Evaluating high-throughput *ab initio* gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. PloS One **7,** e50609.

Gotoh, O., Morita, M., and Nelson, D.R. (2014). Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. BMC Bioinformatics **15,** 189.

Guo, S.G., *et al.* (2013). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nature Genetics **45,** 51-82.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biology **9,** R7.

International Brachypodium Initiative. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463,** 763-768.

Jaillon, O., *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449,** 463-467.

Jones, C.E., Brown, A.L., and Baumann, U. (2007). Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics **8,** 170.

Jupe, F., *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. The Plant Journal **76,** 530-544.

Kang, Y.J., *et al.* (2015). Draft genome sequence of adzuki bean, *Vigna angularis*. Scientific Reports **5,** 8069.

Kang, Y.J., *et al.* (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species. Nature Communications **5**.

Kielbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Research **21,** 487-493.

Kim, S., *et al.* (2015). Integrative structural annotation of *de novo* RNA-Seq provides an accurate reference gene set of the enormous genome of the onion (*Allium cepa* L.). DNA Research **22,** 19-27.

Kim, S., *et al.* (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nature Genetics **46,** 270-278.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology **10,** R25.

Larkin, M.A., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics **23,** 2947-2948.

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S., and You, F.M. (2016). RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. BMC Genomics **17,** 852.

Liu, S., *et al.* (2014). The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nature Communications **5,** 3930.

Lukasik, E., and Takken, F.L. (2009). STANDing strong, resistance proteins instigators of plant defence. Current Opinion in Plant Biology **12,** 427-436.

McHale, L., Tan, X., Koehl, P., and Michelmore, R.W. (2006). Plant NBS-LRR proteins: adaptable guards. Genome Biology **7,** 212.

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.

Ming, R., *et al.* (2015). The pineapple genome and the evolution of CAM photosynthesis. Nature Genetics **47,** 1435-1442.

Potato Genome Sequencing Consortium. (2011). Genome sequence and analysis of the tuber crop potato. Nature **475,** 189-195.

Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into

fleshy fruit evolution. Nature **485,** 635-641.

Schatz, M.C., Witkowski, J., and McCombie, W.R. (2012). Current challenges in *de novo* plant genome sequencing and assembly. Genome Biology **13,** 243.

Schmutz, J., *et al.* (2010). Genome sequence of the palaeopolyploid soybean. Nature **463,** 178-183.

Schnable, P.S., *et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326,** 1112-1115.

Seo, E., Kim, S., Yeom, S.I., and Choi, D. (2016). Genome-wide comparative analyses reveal the dynamic evolution of nucleotide-binding leucine-rich repeat gene family among Solanaceae plants. Frontiers in Plant Science **7,** 1205.

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6,** 31.

Steuernagel, B., Jupe, F., Witek, K., Jones, J.D., and Wulff, B.B. (2015). NLR-parser: rapid annotation of plant NLR complements. Bioinformatics **31,** 1665-1667.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution **30,** 2725-2729.

van den Berg, B.H.J., McCarthy, F.M., Lamont, S.J., and Burgess, S.C. (2010). Re-annotation is an essential step in systems biology modeling of functional genomics data. PloS One **5,** e10642.

van Ooijen, G., Mayr, G., Kasiem, M.M., Albrecht, M., Cornelissen, B.J., and Takken, F.L. (2008). Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. Journal of Experimental Botany **59,** 1383-1397.

Wang, K., *et al.* (2012). The draft genome of a diploid cotton *Gossypium raimondii*. Nature Genetics **44,** 1098-1103.

Xu, Q., *et al.* (2013). The draft genome of sweet orange (*Citrus sinensis*). Nature Genetics **45,** 59-92.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics **13,** 329-342.

Zhang, T., *et al.* (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nature Biotechnology **33,** 531-537.

초 록

유전자 어노테이션이란 유전체에서 유전자의 구조를 찾고 생물학적 기능을 정의하는 것을 의미한다. 이는 유전체를 활용한 거의 모든 추후 분석에서 사용되기 때문에 매우 중요하다. 하지만 알고리즘의 한계와 현존하는 오류 때문에, 현재의 어노테이션은 불완전하며 전체 유전자를 대변하지 못하고 있다. NLR 유전자군은 병 저항성에 관여하며, 식물에서 어노테이션이 잘 되어있지 않은 대표적인 유전자군이다. 이 유전자군은 반복서열을 포함하고 있으며 유전체 내에서 인접해있어 동정이 어렵다. NLR-Finder 는 NLR 유전자군의 어노테이션을 위하여 개발된 생물정보 프로그램이다. 이 프로그램의 가장 큰 특징은 유전체에서 유전자의 후보 지역을 먼저 설정한 후 어노테이션을 수행한다는 것이다. 어노테이션에는 단백질과 전사체 데이터를 사용하였으며, 이 단계에서 찾지 못한 유전자는 *ab initio* gene prediction 이라는 방법을 통해 추가적으로 어노테이션하였다. 식물 17 종의 유전체 서열에서 NLR 유전자의 어노테이션을 새로 수행하였고, 애기장대, 야생잔디, 토마토 등 3 종은 식물에서 어노테이션이 상대적으로 가장 잘 되어있다고 여겨지는 종이기 때문에 프로그램의 타당성을 검증하는데 사용하였다. 이 3 종을 제외한 14 종에서 평균적으로 187 개의 유전자를 더 찾았으며, 타당성 검증 또한 성공적으로 수행되었다. 이를 통해 본 연구는 NLR-Finder 가 NLR 유전자군을 동정하기 위한 쉽고 효율적인 생물정보 프로그램이라는 것을 증명하였다.