RefPlantNLR: a comprehensive collection of experimentally validated plant NLRs

Jiorgos Kourelis and Sophien Kamoun*

The Sainsbury Laboratory, University of East Anglia, Norwich Research Park, NR4 7UH, Norwich, UK

*Correspondence: sophien.kamoun@tsl.ac.uk

ABSTRACT

Reference datasets are critical in computational biology. They help define canonical biological features and are essential for benchmarking studies. Here, we describe a comprehensive reference dataset of experimentally validated plant NLR immune receptors. RefPlantNLR consists of 415 NLRs from 31 genera belonging to 11 orders of flowering plants. We used RefPlantNLR to determine the canonical features of functionally validated plant NLRs. This reference dataset should prove useful for benchmarking NLR annotation tools, guiding comparative analyses of NLRs across the wide spectrum of plant diversity and identifying under-studied taxa. We hope that the RefPlantNLR resource will contribute to moving the field beyond a uniform view of NLR structure and function.

Short title: RefPlantNLR

Keywords: NLR // NB-LRR // NBS-LRR // R gene // Disease resistance gene // Plant immunity

INTRODUCTION

Reference datasets are critical in computational biology (Weber *et al.*, 2019; Schaafsma and Vihinen, 2018). They help define canonical biological features and are essential to benchmarking studies. Reference datasets are particularly important for defining the sequence and domain features of gene and protein families. Despite this, curated collections of experimentally validated sequences are still lacking for several widely studied gene and protein families. One example is the nucleotide-binding leucine-rich repeat (NLR) family of plant proteins. NLRs constitute the predominant class of disease resistance (*R*) genes in plants (Jones *et al.*, 2016; Kourelis and van der Hoorn, 2018; de Araújo *et al.*, 2020). They function as intracellular receptors that detect pathogens and activate an immune response that generally leads to disease resistance. NLRs are thought to be engaged in a coevolutionary tug of war with pathogens and pests. As such, they tend to be among the most polymorphic genes in plant genomes, both in terms of sequence diversity and copy-number variation (Tamborski and Krasileva, 2020). Ever since their first discovery in the 1990s, hundreds of NLRs have been characterized and implicated in pathogen and self-induced immune responses (Kourelis and van der Hoorn, 2018). NLRs are among the most widely studied and economically valuable plant proteins given their importance in breeding crops with disease resistance (Dangl *et al.*, 2013).

NLRs occur widely across all kingdoms of life where they generally function in non-self perception and innate immunity (Jones *et al.*, 2016; Uehling *et al.*, 2017). In the broadest

biochemical definition, NLRs share a similar multidomain architecture consisting of a nucleotide-binding and oligomerization domain (NOD) and a super-structure forming repeat (SSFR) domain (Dyrka *et al.*, 2020). The NOD is either an NB-ARC (nucleotide-binding adaptor shared by APAF-1, certain *R* gene products and CED-4) or NACHT (neuronal apoptosis inhibitory protein, MHC class II transcription activator, HET-E incompatibility locus protein from *Podospora anserina*, and telomerase-associated protein 1), whereas the SSFR domain can be formed by ankyrin (ANK) repeats, tetratricopeptide repeats (TPRs), armadillo (ARM) repeats, WD repeats or leucine-rich repeats (LRRs) (Dyrka *et al.*, 2020; Andolfo *et al.*, 2019). Plant NLRs exclusively carry an NB-ARC domain with the C-terminal SSFR consisting typically of LRRs. The NB-ARC domain has been used to determine the evolutionary relationships between plant NLRs given that it is the only domain that produces reasonably good global alignments across all members of the family. In flowering plants (angiosperms), NLRs form three main monophyletic groups with distinct N-terminal domain fusions: the TIR-NLR subclade containing an N-terminal Toll/interleukin-1 receptor (TIR) domain, the CC-NLR-subclade containing an N-terminal Rx-type coiled-coil (CC) domain and the CC_R-NLR subclade containing an N-terminal RPW8-type CC (CC_R) domain (Tamborski and Krasileva, 2020). Additionally, Lee *et al*. (2020) have recently proposed that the G10-subclade of NLRs is a monophyletic group containing a distinct type of CC (CC_{G10}; CC_{G10}-NLR). NLRs also occur in non-flowering plants where they carry additional types of N-terminal domains such as kinases and α/β hydrolases (Andolfo *et al.*, 2019).

Plant NLRs likely evolved from multifunctional receptors to specialized receptor pairs and networks (Adachi, Contreras, *et al.*, 2019; Adachi, Derevnina, *et al.*, 2019). NLRs which combine pathogen detection and immune signalling activities into a single protein are referred to as "functional singletons", whereas NLRs which have specialized in pathogen recognition or immune signalling are referred to as "sensor" or "helper" NLRs, respectively. About one quarter of NLR genes occur as "genetic singletons" in plant genomes, whereas the others form genetic clusters often near telomeres (Jacob *et al.*, 2013). This genomic clustering likely aids the evolutionary diversification of this gene family and subsequent emergence of pairs and networks (Tamborski and Krasileva, 2020; Adachi, Derevnina, *et al.*, 2019). The emerging picture is that NLRs form genetic and functional receptor networks of varying complexity (Wu *et al.*, 2018; Adachi, Derevnina, *et al.*, 2019).

The mechanism of pathogen detection by NLRs can be either direct or indirect (Kourelis and van der Hoorn, 2018). Direct recognition involves the NLR protein binding a pathogen-derived molecule or serving as a substrate for the enzymatic activity of a pathogen virulence protein (known as effector). Indirect detection is conceptualized by the guard and decoy models where the status of a host component–the guardee or decoy–is monitored by the NLR (van der Biezen and Jones, 1998; van der Hoorn and Kamoun, 2008). Some sensor NLRs known as NLR-IDs contain non-canonical "integrated domains" that can function as decoys to bait pathogen effectors and enable pathogen detection (Cesari *et al.*, 2014; Sarris *et al.*, 2016; Wu *et al.*, 2015). These extraneous domains appear to have evolved by fusion of an effector target domain into an NLR (Cesari *et al.*, 2014; Sarris *et al.*, 2016; Białas *et al.*, 2017). The sequence diversity of integrated domains in NLR-IDs is staggering indicating that novel domain acquisitions have repeatedly occurred throughout the evolution of plant NLRs (Sarris *et al.*, 2016; Kroj *et al.*, 2016).

Given their multidomain nature, sequence diversity and complex evolutionary history, prediction of NLR genes from plant genomes is challenging. Several bioinformatic tools have been developed to extract plant NLRs from sequence datasets. As an input these tools take

either annotated genomic features and transcriptomic data, or alternatively can be run directly on the unannotated genomic sequence. NLR-Parser, RGAugury, RRGPredictor, and DRAGO2 identify transcript and protein sequences that have features of NLRs and are best described as NLR extractors (Steuernagel *et al.*, 2015; Li *et al.*, 2016; Osuna-Cruz *et al.*, 2018; Santana Silva and Micheli, 2020). RGAugury, RRGPredictor, and DRAGO2 also extract other classes of immune-related genes in addition to NLRs. These various tools use pre-defined motifs to classify sequences as NLRs, but they differ in the methods and pipelines. NLR-Annotator, an extension of NLR-Parser–and NLGenomeSweeper, use unannotated genome sequences as input to predict the genomic locations of NLRs (Steuernagel *et al.*, 2020; Toda *et al.*, 2020). This output then requires manual annotation to extract the final gene-models and some of the annotated loci may represent partial or pseudogenized genes.

The goal of this study is to provide a curated reference dataset of experimentally validated plant NLRs. This version of RefPlantNLR (v.20200528_415) consists of 415 NLRs from 31 genera belonging to 11 orders of flowering plants. We used RefPlantNLR to determine the canonical features of functionally validated plant NLRs. RefPlantNLR can also be used to benchmark NLR annotation tools. RefPlantNLR is also useful to guide comparative and phylogenetic analyses of plant NLRs and identify under-studied taxa for future studies.

RESULTS and DISCUSSION

RefPlantNLR pipeline

The current version of RefPlantNLR (v.20200528_415) contains 415 entries from 31 genera of flowering plants (**Figure 1**; **Supplemental dataset 1, 2, 3**). The list was obtained by mining and assessing information gathered from the literature. Briefly, we manually crawled through the literature, extracting plant NLRs that have been experimentally validated to at least some degree. We defined experimental validation broadly as genes reported to be involved in any of the following: 1) disease resistance, 2) disease susceptibility, including effector-triggered immune pathology or trailing necrosis to viruses, 3) hybrid necrosis, 4) autoimmunity, 5) NLR helper function or involvement in downstream immune responses, 6) negative regulation of immunity, and 7) well-described allelic series of NLRs with different pathogen recognition spectra even if not reported to confer disease resistance.

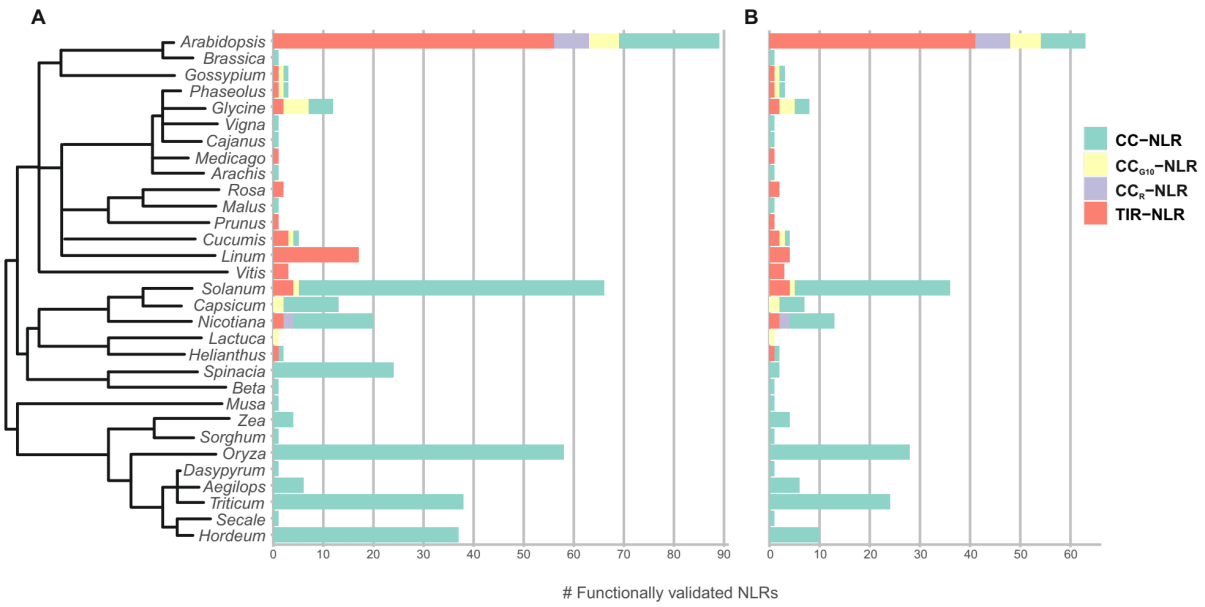

Figure 1: Number of experimentally validated RefPlantNLR sequences per plant genus. **A)** The number of experimentally validated NLRs per plant genus (N = 415), and **B)** the per genus reduced redundancy set at a 90% sequence similarity threshold according to the NB-ARC domain (N = 235) are plotted as a stacked bar graph. The number of experimentally validated NLRs in each genus belonging to the monophyletic TIR-NLR, CC-NLR, CC_R-NLR, or CC_{G10}-NLR subclade members are indicated. There is currently no InterProScan signature for the CC_{G10} N-terminal domain so this sub-class is defined based on phylogeny.

To validate the recovered sequences as genuine NLRs, we annotated them using InterProScan ((Finn *et al.*, 2017); see Material & Methods for the used sequence signatures). We defined NLRs as sequences containing the NB-ARC domain (Pfam signature PF00931) and at least one other annotated domain. This resulted in 414 sequences. We also included the rice protein Pb1, which carries an overall NLR-type domain architecture (N-terminal CC-domain and C-terminal LRRs) but contains an atypical NOD domain instead of the Pfam NB-ARC signature (Hayashi *et al.*, 2010). Altogether these 415 sequences form the current version of RefPlantNLR.

Description of the RefPlantNLR dataset

Table S1 describes the RefPlantNLR dataset, including amino acid, coding sequence (CDS) and locus identifiers, as well as the organism from which the NLR was cloned, the article describing the identification of the NLR, the pathogen type and pathogen to which the NLR provides resistance (when applicable), the matching pathogen effector, additional host components required for pathogen recognition (guardees or decoys) or required for NLR function and the articles describing the identification of the pathogen and host components. We also provide domain annotations of the amino acid and CDS (**Supplemental dataset 4, 5**). From these datasets, we extracted 407 unique NLRs and 347 unique NB-ARC domains (SUPERFAMILY signature SSF52540; **Supplemental dataset 6, 7**). NLRs with identical amino acid sequences were recovered because they have different resistance spectra when genetically linked to different sensor NLR allele (e.g. alleles of Pik), are different in non-coding regions leading to altered regulation (e.g. RPP7 alleles) or have been independently discovered in different plant genotypes (e.g. RRS1-R and SLH1).

The distribution of the RefPlantNLR across plant species mirrors the most heavily studied taxa, i.e. *Arabidopsis*, Solanaceae (*Solanum*, *Capsicum* and *Nicotiana*) and cereals (*Oryza*, *Triticum* and *Hordeum*) (**Figure 1A**). These seven genera comprise 77% (321 out of 415) of the RefPlantNLR sequences. When accounting for redundancy by collapsing similar sequences (>90% NB-ARC amino acid identity per genus), these seven genera would still account for 77% (181 out of 235) sequences (**Figure 1B**). It should be noted that there could be different evolutionary rates of the NB-ARC domain between NLRs, and hence some subfamilies may still be overrepresented in the reduced redundancy set.

In total, 31 plant genera representing 11 taxonomic orders are listed in RefPlantNLR. Interestingly, these species represent only a small fraction of plant diversity with only 11 of 59 major seed plant (spermatophyte) orders described by Smith and Brown represented, and no single entry from non-flowering plants (**Table S2**) (Smith and Brown, 2018). Arabidopsis remains the only species with experimentally validated NLRs from the four major clades (CC-NLR, CC_{G10}-NLR, CC_R-NLR and TIR-NLR) (**Figure 1**).

The average length of RefPlantNLR sequences varies depending on the subclass (**Figure 2A**, **Figure 2C** for the reduced redundancy set). CC-NLRs varied from 696 to 1824 amino acids

(mean = 1089, N = 288), whereas TIR-NLR varied from 380 to 2048 amino acids (mean = 1169, N = 92). NB-ARC domains were more constrained (mean = 265, N = 347, stdev = 20) (**Figure 2B**). Nonetheless, 11 atypically short NB-ARCs (148 to 224 amino acids) and 4 long NB-ARCs (308 to 339 amino acids) were observed at more than two standard deviations of the mean illustrating the overall flexibility of plant NLRs even for this canonical domain (**Figure 2B, Figure 2D** for the reduced redundancy set).

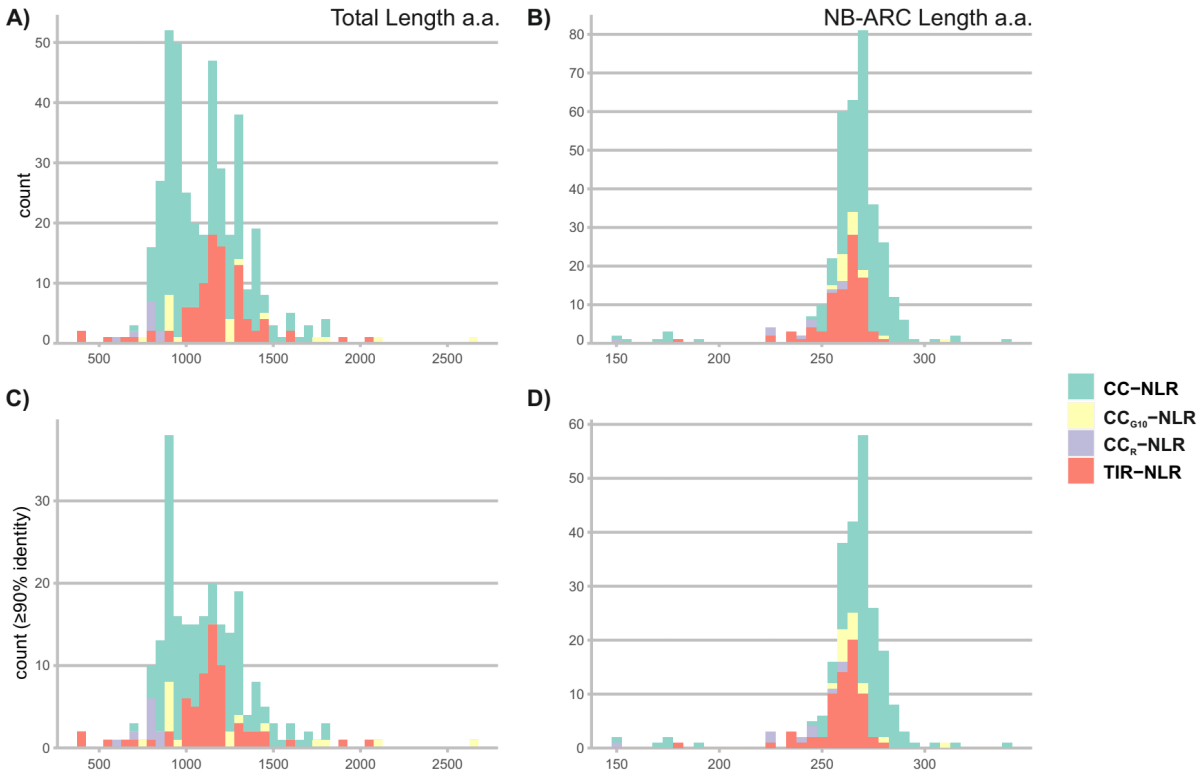

Figure 2: Length distribution RefPlantNLR amino acid sequence and extracted NB-ARC domains. Length distribution of the RefPlantNLR sequences. **A)** Histogram of unique RefPlantNLR amino acid sequence length (binwidth 50aa, N = 407). **B)** Histogram of the unique RefPlantNLR extracted NB-ARC domain (SUPERFAMILY signature SSF52540) amino acid sequence length (binwidth 5aa, N = 347). **C)** Histogram of amino acid sequence length of the reduced redundancy RefPlantNLR set at a 90% amino acid similarity threshold according to the NB-ARC domain length (binwidth 50aa, N = 235). **D)** Histogram of the extracted NB-ARC domain from the reduced redundancy RefPlantNLR set (binwidth 5aa, N = 240). Colour-coding according to NLR subfamily.

We noted that some of the unusually small NLRs lacked a SSFR domain, while some of the small NB-ARC domains appeared to be partial duplications of this domain. In order to look at domain architecture of NLRs more widely and to determine whether these unusual features are common, we mapped the domain architecture of RefPlantNLR proteins (**Figure 3A, Figure 3B** for the reduced redundancy set). Even though CC-NLR and TIR-NLR domain combinations were the most frequent (60% and 19%, respectively), we observed additional domain combinations. In the RefPlantNLR dataset, a subset of NLRs lack the N-terminal domain but still group with the major NLR clades based on the NB-ARC phylogeny. Some TIR-NLRs lack a SSFR domain. Non-canonical integrated domains are found in all NLR subfamilies, and occur at the N-terminus, in between the N-terminal domain and the NB-ARC domain, at the C-terminus or both ends. Of these non-canonical domains, the N-terminal late-blight resistance protein R1 domain (also known as the Solanaceae domain; Pfam signature

PF12061) only occurs in association with the NB-ARC domain, and has an ancient origin likely in the most recent common ancestor of the Asterids and Amaranthaceae (Seong *et al.*, 2020). Other non-canonical domains are also more wide-spread, including the monocot-specific integration of a zinc-finger BED domain, which occurs in between the CC and NB-ARC domain (Bailey *et al.*, 2018; Marchal *et al.*, 2018). Finally, whereas some CC-NLRs contain a partial duplication of the NB-ARC domain, TIR-NLRs appear to allow a more flexible domain architecture with multiple combinations of TIR and NB-ARC domains within the same protein (**Figure 3C**).

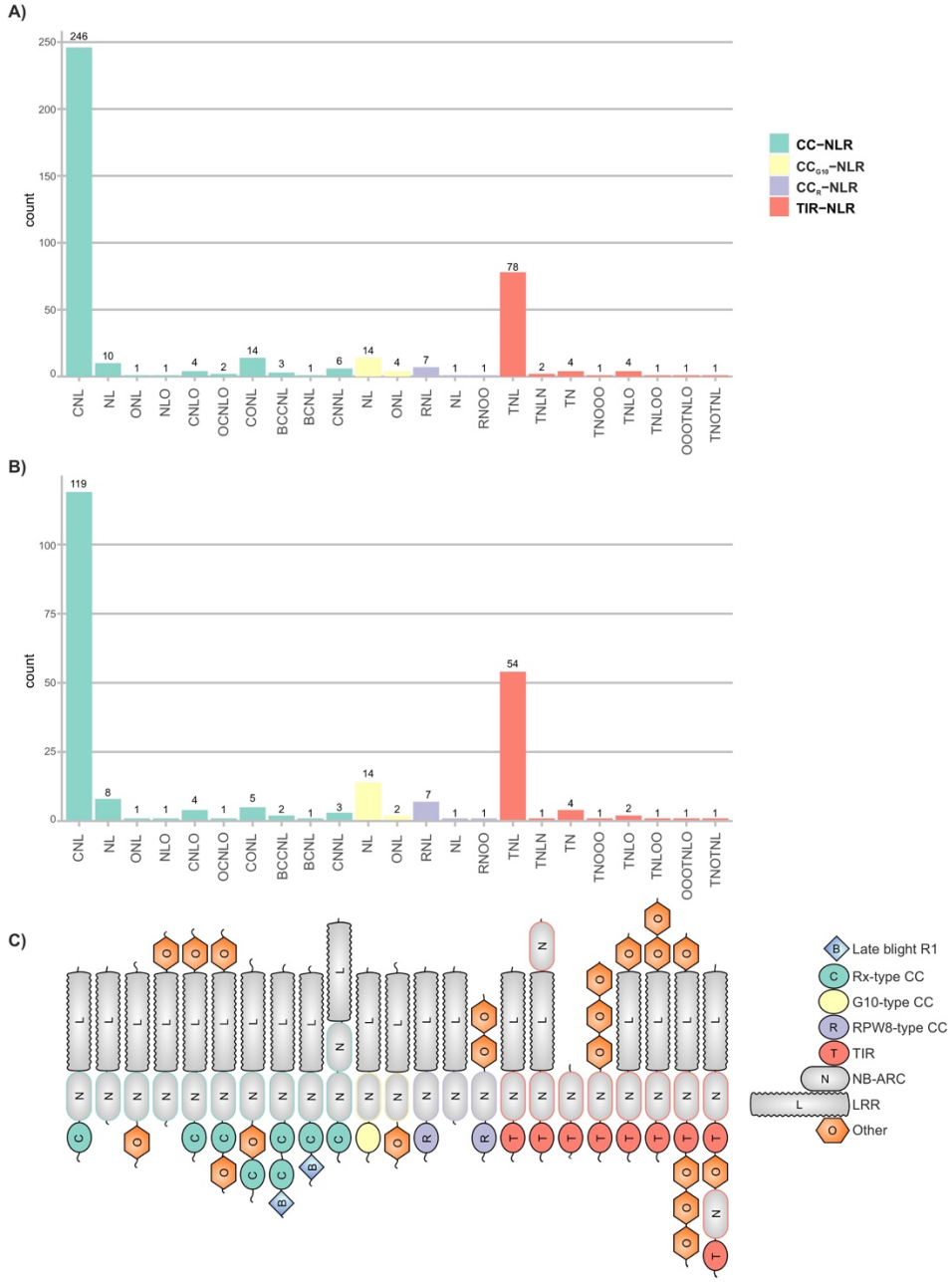

Figure 3: Domain architecture of the RefPlantNLRs. Bar chart of the domain architecture of **A)** the unique RefPlantNLRs (N = 407), or **B)** the per genus reduced redundancy RefPlantNLR set at a 90% amino acid similarity threshold according to the NB-ARC domain (N = 235). **C)** Schematic representation of domain architecture. Used InterPro signatures for each of the domains are highlighted in the material & methods. There is currently no InterProScan signature for the CC_{G10} N-terminal domain.

We explored the phylogenetic diversity of RefPlantNLR proteins using the extracted NB-ARC domains (**Figure 4**; **Supplemental dataset 8, 9)**. As with previously reported NLR phylogenetic analyses, RefPlantNLR sequences generally grouped in well-defined clades, notably CC-NLR, CC_{G10}-NLR, CC_{R}-NLR and TIR-NLR. Within this phylogeny some of the branches, notably Pi54, are long and may represent highly diverged NB-ARC domains. Since Pb1 does not match the Pfam NB-ARC domain it was not included in this phylogenetic analysis.

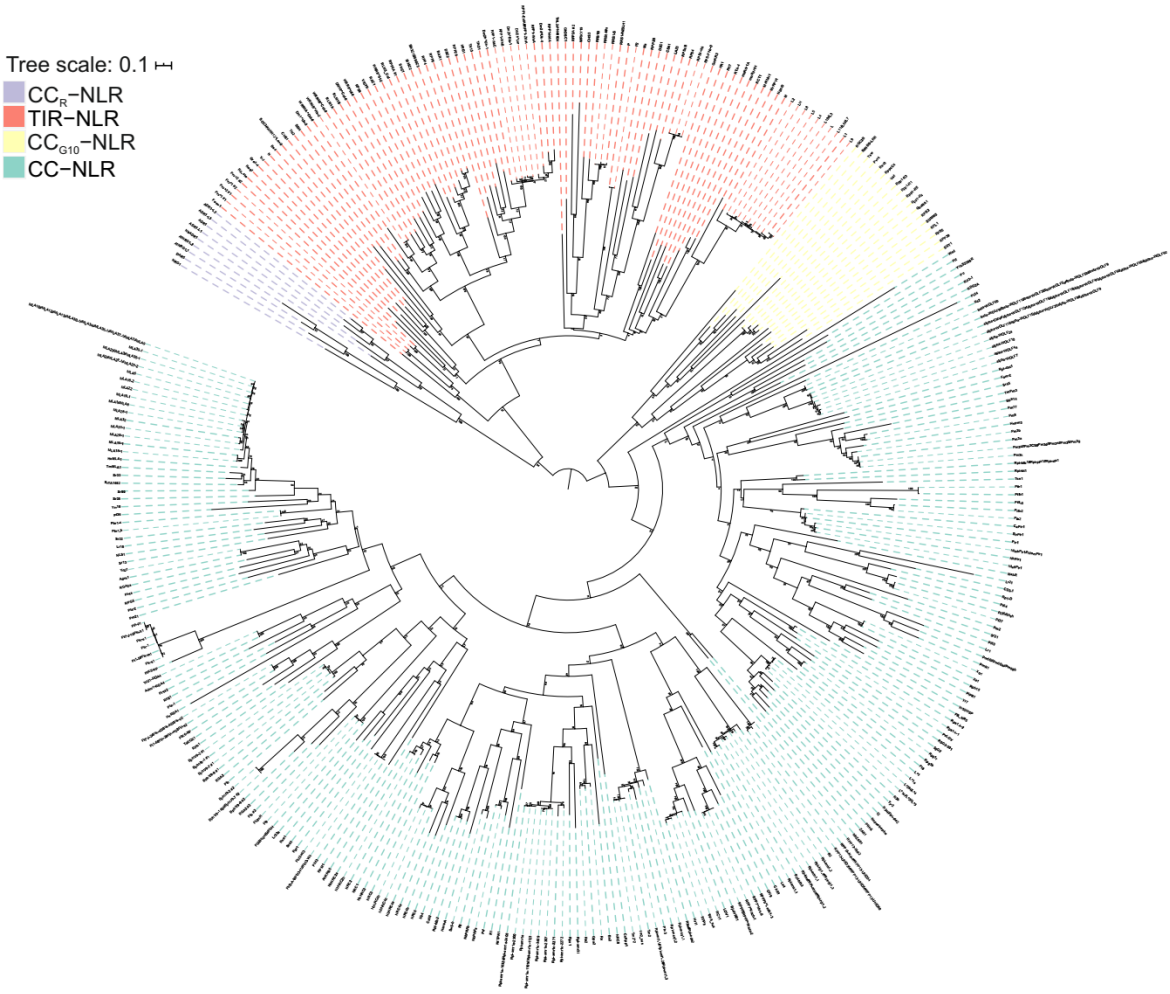

Figure 4: Phylogenetic diversity of RefPlantNLR sequences. The tree, based on the NB-ARC domain (SUPERFAMILY signature SSF52540), was inferred using the Maximum Likelihood method based on the Jones-Taylor-Thornton (JTT) model (Jones *et al.*, 1992). The tree with the highest log likelihood is shown. Branch labels indicates percentage bootstrap values inferred from 1000 rapid bootstrap replicates (Stamatakis *et al.*, 2008). NLRs with identical NB-ARC domains are collapsed, while for those with multiple NB-ARC domains the NB-ARC are numbered according to descending amino acid length. The representative sequences of the 90% sequence identity reduced redundancy set are indicated in bold. The TIR-NLR, CC-NLR, CC_{R}-NLR, and CC_{G10}-NLR subclades are indicated. The tree can be accessed online at https://itol.embl.de/tree/1491551921244583715930141 35.

Applications of the RefPlantNLR dataset

There are many applications of the RefPlantNLR dataset. RefPlantNLR should prove useful for benchmarking NLR extraction and annotation tools. The majority of the reported NLR prediction tools have only been evaluated using the Arabidopsis NLRome, which is not representative of NLR diversity across flowering plants (**Figure 1**). Additionally, it would be

useful to precisely document the limitations and biases of each prediction method to enable users to select the approach that best matches their needs and project objectives. Ideally, NLR extraction tools would correctly identify the N-terminal domain, the boundaries of the NB-ARC, the type of SSFR domain and any non-canonical domains that are present.

Another use of RefPlantNLR would be to provide reference points for newly discovered NLRs and in large-scale phylogenetic analyses of NLRs. Phylogenetic analyses can be readily performed by extracting the SUPERFAMILY signature SSF52540 of the NLRs of interest and aligning the retrieved sequences to the NB-ARC domains of the RefPlantNLR list (**Supplemental dataset 7**). Phylogenetic analyses would help assign NLRs to subclades and provide a basis for generating hypotheses about the function and mode of action of novel NLRs which phylogenetically cluster with experimentally validated NLRs. This type of phylogenetic information can be combined with other features such as genetic clustering and has, for instance, proven valuable in our work on rice and solanaceous NLRs (Białas *et al.*, 2017; Wu *et al.*, 2017) and for defining the CC_{G10}-NLR class (Lee *et al.*, 2020).

RefPlantNLR highlights the under-studied plant species of NLR biology. **Table S2** reveals that 48 out of the 59 major seed plant clades recently defined by Smith and Brown (Smith and Brown, 2018) do not have a single experimentally validated NLR. Certain taxa have subfamily-specific contractions and expansions, and hence may contain unexplored genetic and biochemical diversity of NLR function. For example, while all currently studied CC_R-NLRs act as helper NLRs for TIR-NLRs, it has been reported that the CC_R-NLR subfamily have experienced clade-specific expansions in gymnosperms and rosids, pointing to potential biochemical specialization of this subfamily in these taxa (Van Ghelder *et al.*, 2019). In addition, although NLRs have been reported in non-seed plants and some of these appear to have distinct N-terminal domains (Andolfo *et al.*, 2019), their experimental validation is still lacking.

The RefPlantNLR dataset has inherent limitations due to its focus on experimentally validated NLRs. First, it is biased towards a few highly studied model species and crops as illustrated in **Figure 1**. Additionally, RefPlantNLR entries are somewhat redundant with particular NLR allelic series, such as the monocot MLA and spinach alpha-WOLF, being overrepresented in the dataset (**Figure 1**, **Figure 4**). These issues, notably redundancy, will need to be considered for certain applications such as benchmarking. Our recommendation is to use the reduced redundancy dataset of 235 sequences for benchmarking exercises (**Supplemental dataset 10, 11**).

Conclusion

We hope that the RefPlantNLR resource will contribute to moving the field beyond a uniform view of NLR structure and function. It is now evident that NLRs turned out to be more structurally and functionally diverse than anticipated. Whereas a number of plant NLRs have retained the presumably ancestral three domain architecture of the TIR/CC_R/CC_{G10}/CC fused to the NB-ARC and LRR domains, many NLRs have diversified into specialized proteins with degenerated features and extraneous non-canonical integrated domains (Sarris *et al.*, 2016; Kroj *et al.*, 2016; Adachi, Derevnina, *et al.*, 2019). Therefore, it is time to question holistic concepts such as effector-triggered immunity (ETI) and appreciate the wide structural and functional diversity of NLR-mediated immunity. More specifically, a robust phylogenetic framework of plant NLRs should be fully integrated into the mechanistic study of these exceptionally diverse proteins.

MATERIAL & METHODS

Sequence retrieval

RefPlantNLR was assembled by manually crawling the literature for experimentally validated NLRs according to the criteria described in the results section. NLRs are defined as having a NB-ARC and at least one additional domain. Where possible the amino acid and nucleotide sequences were taken from GenBank. For some NLRs, only the mRNA has been deposited and no genomic locus information was present. When GenBank records were not available, the sequences were extracted from the matching whole-genome sequences projects or from articles and patents describing the identification of these NLRs.

Domain annotation

Protein sequences were annotated with CATH-Gene3D (v4.2.0) (Dawson *et al.*, 2017), SUPERFAMILY (v1.75) (Gough *et al.*, 2001), PRINTS (v42.0) (Attwood *et al.*, 2000), and Pfam (v32.0) (El-Gebali *et al.*, 2019) identifiers using InterProScan (v5.44-79.0) (Finn *et al.*, 2017). A custom R script (**Appendix S1**) was used to convert the InterProScan output to the final GFF3 annotation of protein and CDS using the InterPro descriptions (**Appendix S2**). We routinely use Geneious Prime (v2020.1.2) (https://www.geneious.com) to visualize these annotations on the sequence. The NLR-associated signature motifs/domain IDs are:

- Late blight resistance protein R1: PF12061
- Rx-type CC: PF18052, cd14798
- RPW8-type CC: PF05659, PS51153
- TIR: PF01582, G3DSA:3.40.50.10140, SSF52200, PF13676
- NB-ARC: PF00931
- NB-ARC used for phylogenetic analysis: SSF52540
- LRRs: G3DSA:3.80.10.10, G3DSA:1.25.10.10, PF08263, PF00560, PF07723, PF07725, PF12799, PF13306, PF13516, PF13855, SSF52047, SSF52058
- Other: any other Pfam, SUPERFAMILY, and/or CATH-Gene3D annotation. Additionally, we included the PROSITE Profiles signatures PS51697 (ALOG domain) and PS50808 (zinc-finger BED domain), and the SMART signature SM00614 (zinc-finger BED domain).

Sequence deduplication

The extracted NB-ARC domains (SUPERFAMILY signature SSF52540) were clustered using CD-HIT at 90% sequence identity ((Fu *et al.*, 2012); Usage: cd-hit -i RefPlantNLR_v.20200528_415_SSF52540.fasta -o RefPlantNLR_SSF52540_90 -c 0.90 -n 5 -M 16000 -d 0). A custom R script (**Appendix S1**) was used to assign representative sequences per cluster per genus, i.e. if a single cluster contained sequences from multiple genera we assigned a representative sequence per genus. In case an NLR has multiple NB-ARC domains we grouped this NLR in clusters according to the largest NB-ARC domain. The reduced redundancy sequences are provided in **Supplemental dataset 10, 11**.

Phylogenetics

The NB-ARC domain (SUPERFAMILY signature SSF52540) of all NLRs was extracted using Geneious Prime. A custom R script (**Appendix S1**) was used to deduplicate sequences with

identical NB-ARC domains and assign unique identifiers to sequences containing multiple NB-ARC domains. Sequences were aligned using Clustal Omega (Sievers *et al.*, 2011), and all positions with less than 95% site coverage were removed using QKphylogeny (Moscou, 2020) (**Supplemental dataset 8**). RAxML (v8.2.12) (Stamatakis, 2014) was used (usage: raxmlHPC-PTHREADS-AVX -T 6 -s RefPlantNLR_v.20200528_415_SSF52540_Unique_ClustalOmega_missing5.phy -n RefPlantNLRs_v.20200528_415 -m PROTGAMMAAUTO -f a -# 1000 -x 8153044963028367 -p 644124967711489) to infer the evolutionary history using the Maximum Likelihood method based on the JTT model (Jones *et al.*, 1992). Bootstrap values from 1000 rapid bootstrap replicates as implemented in RAxML are shown (Stamatakis *et al.*, 2008) (**Supplemental dataset 9**). The RefPlantNLR phylogeny was edited using the iTOL suite (v5.5.1; Letunic and Bork, 2019), and can be accessed online through: https://itol.embl.de/shared/JKourelis

ACKNOWLEDGEMENTS

We thank Hiroaki Adachi, Adeline Harant, and Philip Carella for useful comments and feedback, and Aleksandra Białas for the domain architecture illustrations.

FUNDING

This work has been supported by the Gatsby Charitable Foundation, Biotechnology and Biological Sciences Research Council (BBSRC), European Research Council (ERC) and BASF Plant Science. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

COMPETING INTERESTS

The authors receive funding from industry on NLR biology.

AUTHOR CONTRIBUTIONS

Conceptualization: JK, SK; Data curation: JK; Formal analysis: JK; Investigation: JK; Supervision: SK; Funding acquisition: SK; Project administration: SK; Writing initial draft: JK; Editing: JK, SK.

REFERENCES

Adachi, H., Contreras, M.P., Harant, A., et al. (2019) An N-terminal motif in NLR immune receptors is functionally conserved across distantly related plant species. *eLife*, **8**, e49956.

Adachi, H., Derevnina, L. and Kamoun, S. (2019) NLR singletons, pairs, and networks: evolution, assembly, and regulation of the intracellular immunoreceptor circuitry of plants. *Curr. Opin. Plant Biol.*, **50**, 121–131.

Andolfo, G., Di Donato, A., Chiaiese, P., De Natale, A., Pollio, A., Jones, J.D.G., Frusciante, L. and Ercolano, M.R. (2019) Alien domains shaped the modular structure of plant NLR proteins. *Genome Biol. Evol.*, **11**, 3466–3477.

Araújo, A.C. de, Fonseca, F.C.D.A., Cotta, M.G., Alves, G.S.C. and Miller, R.N.G. (2020) Plant NLR receptor proteins and their potential in the development of durable genetic resistance to biotic stresses. *Biotechnol. Res. Innov.*

Attwood, T.K., Croning, M.D.R., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N. and Wright, W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.

Bailey, P.C., Schudoma, C., Jackson, W., Baggs, E., Dagdas, G., Haerty, W., Moscou, M. and Krasileva, K.V. (2018) Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *Genome Biol.*, **19**, 23.

Bialas, A., Zess, E.K., De la Concepcion, J.C., et al. (2017) Lessons in effector and NLR biology of plant-microbe systems. *Mol. Plant. Microbe Interact.*, **31**, 34–45.

Biezen, E.A. van der and Jones, J.D.G. (1998) Plant disease-resistance proteins and the gene-for-gene concept. *Trends Biochem. Sci.*, **23**, 454–456.

Cesari, S., Bernoux, M., Moncuquet, P., Kroj, T. and Dodds, P.N. (2014) A novel conserved mechanism for plant NLR protein pairs: the "integrated decoy" hypothesis. *Plant-Microbe Interact.*, **5**, 606.

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

Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A. and Sillitoe, I. (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.*, **45**, D289–D295.

Dyrka, W., Coustou, V., Daskalov, A., et al. (2020) Identification of NLR-associated amyloid signaling motifs in filamentous bacteria. *bioRxiv*, 2020.01.06.895854.

El-Gebali, S., Mistry, J., Bateman, A., et al. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.

Finn, R.D., Attwood, T.K., Babbitt, P.C., et al. (2017) InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.

Hayashi, N., Inoue, H., Kato, T., et al. (2010) Durable panicle blast-resistance gene *Pb1* encodes an atypical CC-NBS-LRR protein and was generated by acquiring a promoter through local genome duplication. *Plant J.*, **64**, 498–510.

Hoorn, R.A.L. van der and Kamoun, S. (2008) From guard to decoy: a new model for perception of plant pathogen effectors. *Plant Cell Online*, **20**, 2009–2017.

Jacob, F., Vernaldi, S. and Maekawa, T. (2013) Evolution and Conservation of Plant NLR Functions. *Front. Immunol.*, **4**.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, **8**, 275–282.

Jones, J.D.G., Vance, R.E. and Dangl, J.L. (2016) Intracellular innate immune surveillance devices in plants and animals. *Science*, **354**.

Kourelis, J. and Hoorn, R.A.L. van der (2018) Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *Plant Cell*, **30**, 285–299.

Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X. and Morel, J.-B. (2016) Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol.*, **210**, 618–626.

Lee, H.-Y., Mang, H., Choi, E.-H., Seo, Y.-E., Kim, M.-S., Oh, S., Kim, S.-B. and Choi, D. (2020) Genome-wide functional analysis of hot pepper immune receptors reveals an autonomous NLR cluster in seed plants. *bioRxiv*, 2019.12.16.878959.

Letunic, I. and Bork, P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, **47**, W256–W259.

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.

Marchal, C., Zhang, J., Zhang, P., et al. (2018) BED-domain-containing immune receptors confer diverse resistance spectra to yellow rust. *Nat. Plants*, **4**, 662–668.

Moscou, M. (2020) *matthewmoscou/QKphylogeny*, Available at: https://github.com/matthewmoscou/QKphylogeny.

Osuna-Cruz, C.M., Paytuvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Aiese Cigliano, R., Sanseverino, W. and Ercolano, M.R. (2018) PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.*, **46**, D1197–D1201.

Santana Silva, R.J. and Micheli, F. (2020) RRGPredictor, a set-theory-based tool for predicting pathogen-associated molecular pattern receptors (PRRs) and resistance (R) proteins from plants. *Genomics*, **112**, 2666–2676.

Sarris, P.F., Cevik, V., Dagdas, G., Jones, J.D.G. and Krasileva, K.V. (2016) Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol.*, **14**, 8.

Schaafsma, G.C.P. and Vihinen, M. (2018) Representativeness of variation benchmark datasets. *BMC Bioinformatics*, **19**, 461.

Seong, K., Seo, E., Witek, K., Li, M. and Staskawicz, B. (2020) Evolution of NLR resistance genes with noncanonical N-terminal domains in wild tomato species. *New Phytol.*

Sievers, F., Wilm, A., Dineen, D., et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

Smith, S.A. and Brown, J.W. (2018) Constructing a broadly inclusive seed plant phylogeny. *Am. J. Bot.*, **105**, 302–314.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stamatakis, A., Hoover, P. and Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.*, **57**, 758–771.

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

Steuernagel, B., Witek, K., Krattinger, S.G., et al. (2020) The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.*, **183**, 468–482.

Tamborski, J. and Krasileva, K.V. (2020) Evolution of plant NLRs: from natural history to precise modifications. *Annu. Rev. Plant Biol.*, **71**, 355–378.

Toda, N., Rustenholz, C., Baud, A., Le Paslier, M.-C., Amselem, J., Merdinoglu, D. and Faivre-Rampant, P. (2020) NLGenomeSweeper: a tool for genome-wide NBS-LRR resistance gene identification. *Genes*, **11**, 333.

Uehling, J., Deveau, A. and Paoletti, M. (2017) Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems. *PLOS Pathog.*, **13**, e1006578.

Van Ghelder, C., Parent, G.J., Rigault, P., et al. (2019) The large repertoire of conifer NLR resistance genes includes drought responsive and highly diversified RNLs. *Sci. Rep.*, **9**, 11614.

Weber, L.M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P.P., Boulesteix, A.-L., Saeys, Y. and Robinson, M.D. (2019) Essential guidelines for computational method benchmarking. *Genome Biol.*, **20**, 125.

Wu, C.-H., Abd-El-Haliem, A., Bozkurt, T.O., Belhaj, K., Terauchi, R., Vossen, J.H. and Kamoun, S. (2017) NLR network mediates immunity to diverse plant pathogens. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 8113–8118.

Wu, C.-H., Derevnina, L. and Kamoun, S. (2018) Receptor networks underpin plant immunity. *Science*, **360**, 1300–1301.

Wu, C.-H., Krasileva, K.V., Banfield, M.J., Terauchi, R. and Kamoun, S. (2015) The "sensor domains" of plant NLR proteins: more than decoys? *Plant-Microbe Interact.*, **6**, 134.

SUPPLEMENTAL DATA

Table S1: Description of RefPlantNLR.

Table S2: Plant orders represented in RefPlantNLR.

Supplemental dataset 1: Amino acid sequences of RefPlantNLR entries (fasta format). This file contains 415 amino acid sequences.

Supplemental dataset 2: CDS sequences of RefPlantNLR entries (fasta format). This file contains 400 CDS sequences. CDS sequences could not be retrieved for 15 RefPlantNLR entries.

Supplemental dataset 3: Annotated genomic sequences of RefPlantNLR entries (GenBank flat file format). This file contains 329 genomic loci containing the gene models of 344 RefPlantNLR entries and 56 RefPlantNLR mRNA entries lacking genomic information.

Supplemental dataset 4: InterProScan annotation of the RefPlantNLR amino acid sequences (GFF3 format). This file contains the InterProScan annotation of 415 amino acid sequences.

Supplemental dataset 5: InterProScan annotation of the RefPlantNLR CDS sequences (GFF3 format). This file contains the InterProScan annotation of the 400 CDS sequences.

Supplemental dataset 6: Amino acid sequences of the extracted RefPlantNLR NB-ARC domains (fasta format). This file contains 424 NB-ARC domain (SUPERFAMILY signature SSF52540) amino acid sequences belonging to 415 RefPlantNLR entries.

Supplemental dataset 7: Amino acid sequences of the unique RefPlantNLR extracted NB-ARC domains (fasta format). This file contains 347 unique NB-ARC domain (SUPERFAMILY signature SSF52540) amino acid sequences.

Supplemental dataset 8: Clustal Omega alignment of the unique RefPlantNLR extracted NB-ARC domains (PHYLIP format). This file contains the Clustal Omega alignment of 346 unique NB-ARC domains (SUPERFAMILY signature SSF52540) with all positions with less than 95% coverage removed. Pb1 was omitted from this alignment.

Supplemental dataset 9: NB-ARC domain phylogeny of the RefPlantNLR entries using the Maximum likelihood method (Newick format). This file contains the phylogenetic analysis of the NB-ARC domain of the RefPlantNLR entries using the JTT method.

Supplemental dataset 10: Amino acid sequences of the non-redundant RefPlantNLR entries (fasta format). This file contains 235 amino acid sequences representing the non-redundant RefPlantNLR entries at a 90% amino acid identity threshold per genus according to the NB-ARC domain.

Supplemental dataset 11: Amino acid sequences of the NB-ARC domains of the non-redundant RefPlantNLR entries (fasta format). This file contains 241 amino acid sequences representing the extracted NB-ARC domains of the 235 non-redundant RefPlantNLR.

Appendix S1: R script used to generate annotations and figures.

Appendix S2: InterProScan descriptions used for generating annotations.

DATA AVAILABILITY AND UPDATES

Up to date versions of RefPlantNLR can be accessed via Zenodo at http://doi.org/10.5281/zenodo.3936022. This project is part of the OpenPlantNLR community on Zenodo: https://zenodo.org/communities/openplantnlr