Jurassic NLR: conserved and dynamic evolutionary features of the atypically ancient immune receptor ZAR1

Hiroaki Adachi, Toshiyuki Sakai, Jiorgos Kourelis, Abbas Maqbool and Sophien Kamoun

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

ABSTRACT

NLR immune receptors form one of the most diverse protein families in flowering plants (angiosperms). NLRs have massively expanded through birth-and-death evolution and typically exhibit hallmarks of rapid evolution even at the intraspecific level. Here, we reconstructed the evolutionary history of ZAR1, an atypically conserved NLR that traces its origin to early angiosperm lineages ~220 to 150 million years ago (Jurassic period). We used iterative sequence similarity searches coupled with phylogenetic analyses to determine the degree to which ZAR1 orthologs and paralogs occur in plants. We discovered 120 ZAR1 orthologs in 88 species, including the monocot *Colacasia esculenta*, the magnoliid *Cinnamomum micranthum* and the majority of eudicots, notably the early diverging eudicot species *Aquilegia coerulea*. Analyses of the ortholog sequences revealed highly conserved features of ZAR1, including regions for pathogen effector recognition, intramolecular interactions and cell death activation. This also uncovered a new conserved surface on the underside of the activated ZAR1 resistosome wheel. Throughout its evolution, ZAR1 also acquired novel features. Nine ZAR1 orthologs from cassava and cotton carry an integrated thioredoxin-like domain at their C-termini. ZAR1 also duplicated into two paralog families ZAR1-SUB and ZAR1-CIN. ZAR1-SUB, which emerged in the eudicots, is a large class of sequence diverse ZAR1 paralogs that lack several of the conserved motifs of ZAR1. A second family, ZAR1-CIN, comprises an expansion of 11 paralogs unique to a ~500 kb locus in the *C. micranthum* genome and located about 48 Mb from ZAR1. We conclude that ZAR1 stands out among angiosperm NLRs for having an ancient origin and having experienced relatively limited gene duplication and expansion throughout its deep evolutionary history. Nonetheless, ZAR1 did also give rise to non-canonical NLR proteins with integrated domains and degenerated molecular features.

INTRODUCTION

Plants immune receptors, often encoded by disease resistance (*R*) genes, detect invading pathogens and activate innate immune responses that can limit infection (Jones and Dangl, 2006). A major class of immune receptors is formed by intracellular proteins of the nucleotide-binding leucine-rich repeat (NLR) family (Dodds and Rathjen, 2010; Jones et al., 2016; Kourelis and van der Hoorn, 2018). NLRs detect host-translocated pathogen effectors either by directly binding them or indirectly via host proteins known as guardees or decoys. NLRs are arguably the most diverse protein family in flowering plants (angiosperms) with many species having large (>100) and diverse repertoires of NLRs in their genomes (Shao et al., 2016; Baggs et al., 2017). They typically exhibit hallmarks of rapid evolution even at the intraspecific level (Van de Weyer et al., 2019; Lee and Chae, 2020; Prigozhin and Krasileva,

2020). Towards the end of the 20th century, Michelmore and Meyers (1998) proposed that NLRs evolve primarily through the birth-and-death process (Nei and Hughes, 1992). In this model, new NLRs emerge by recurrent cycles of gene duplication and loss—some genes are maintained in the genome acquiring new pathogen detection specificities, whereas others are deleted or become non-functional through the accumulation of deleterious mutations. Such dynamic patterns of evolution enable the NLR immune system to keep up with fast-evolving effector repertoires of pathogenic microbes. However, as already noted over 20 years ago by Michelmore and Meyers (1998), a subset of NLR proteins are slow evolving and have remained fairly conserved throughout evolutionary time (Wu et al., 2017; Stam et al., 2019). These "high-fidelity" NLRs (per Lee and Chae, 2020) offer unique opportunities for comparative analyses, providing a molecular evolution framework to reconstruct key transitions and reveal functionally critical biochemical features (Delaux et al., 2019). Nonetheless, comprehensive evolutionary reconstructions of conserved NLR proteins remain limited despite the availability of a large number of plant genomes across the breadth of plant phylogeny. One of the reasons is that the great majority of NLRs lack clear-cut orthologs across divergent plant taxa. Here, we address this gap in knowledge by investigating ZAR1 (HOPZ-ACTIVATED RESISTANCE1), an atypically ancient NLR, and asking fundamental questions about the conservation and diversification of this immune receptor throughout its deep evolutionary history.

NLRs occur across all kingdoms of life and generally function in non-self perception and innate immunity (Jones et al., 2016; Uehling et al., 2017). In the broadest biochemical definition, plant NLRs share a multidomain architecture typically consisting of a NB-ARC (nucleotide-binding domain shared with APAF-1, various R-proteins and CED-4) followed by a leucine- rich repeat (LRR) domain. Angiosperm NLRs form several major monophyletic groups with distinct N-terminal domain fusions (Shao et al., 2016; Kourelis and Kamoun, 2020). These include the subclades TIR-NLR with the Toll/interleukin-1 receptor (TIR) domain, CC-NLR with the Rx-type coiled-coil (CC) domain, CC_R-NLR with the RPW8-type CC (CC_R) domain (Tamborski and Krasileva, 2020) and the more recently defined CC_{G10}-NLR with a distinct type of CC (CC_{G10}) (Lee et al., 2020). Up to 10% of NLRs carry unconventional "integrated" domains in addition to the canonical tripartite domain architecture. Integrated domains are thought to generally function as decoys to bait pathogen effectors and enable pathogen detection (Cesari et al., 2014; Sarris et al., 2016; Wu et al., 2015; Kourelis and van der Hoorn, 2018). They include dozens of different modules indicating that novel domain acquisitions have repeatedly taken place throughout the evolution of plant NLRs (Sarris et al., 2016; Kroj et al., 2016). To date, over 400 NLRs from 31 genera in 11 orders of flowering plants have been experimentally validated as reported in the RefPlantNLR reference dataset (Kourelis and Kamoun, 2020). Several of these NLRs are coded by *R* genes that function against economically important pathogens and contribute to sustainable agriculture (Dangl et al., 2013).

In recent years, the research community has gained a better understanding of the structure/function relationships of plant NLRs and the immune receptor circuitry they form (Wu et al., 2018; Adachi et al., 2019a; Burdett et al., 2019; Jubic et al., 2019; Bayless and Nishimura, 2020; Feehan et al., 2020; Mermigka et al., 2020; Wang and Chai, 2020; Xiong et al., 2020; Zhou and Zhang, 2020). Some NLRs, such as ZAR1, form a single functional unit that carries both pathogen sensing and immune signalling activities in a single protein (termed 'singleton NLR' per Adachi et al., 2019a). Other NLRs function together in pairs or more

complex networks, where connected NLRs have functionally specialized into sensor NLRs dedicated to pathogen detection or helper NLRs that are required for sensor NLRs to initiate immune signalling (Feehan et al., 2020). Paired and networked NLRs are thought to have evolved from multifunctional ancestral receptors through asymmetrical evolution (Adachi et al., 2019a, 2019b). As a result of their direct coevolution with pathogens, NLR sensors tend to diversify faster than helpers and can be dramatically expanded in some plant taxa (Wu et al., 2017; Stam et al., 2019). For instance, sensor NLRs often exhibit non-canonical biochemical features, such as degenerated functional motifs and unconventional domain integrations (Adachi et al., 2019b; Seong et al., 2020).

The elucidation of plant NLR structures by cryo-electron microscopy has significantly advanced our understanding of the biochemical events associated with the activation of these immune receptors (Wang et al., 2019a; 2019b; Martin et al., 2020). Both the CC-NLR ZAR1 and the TIR-NLR Roq1 oligomerize upon activation into a wheel-like multimeric complex known as the resistosome. In the case of ZAR1, recognition of bacterial effectors occurs through its partner receptor-like cytoplasmic kinases (RLCKs), which tend to vary depending on the pathogen effector and host plant (Lewis et al., 2013; Wang et al., 2015; Seto et al., 2017; Schultink et al. 2019; Laflamme et al., 2020). Activation of ZAR1 induces conformational changes in the nucleotide binding domain resulting in ADP release, dATP/ATP binding and pentamerization of the ZAR1–RLCK complex into the resistosome. The ZAR1 resistosome exposes a funnel-shaped structure formed by the N-terminal α1 helices, which translocates into the plasma membrane and is thought to perturb membrane integrity to trigger cell death response (Wang et al., 2019b). The ZAR1 N-terminal α1 helix matches the MADA consensus sequence motif that is functionally conserved in ~20% of CC-NLRs including NLRs from dicot and monocot plant species (Adachi et al., 2019b). This suggests that the biochemical 'death switch' mechanism of the ZAR1 resistosome may apply to a significant fraction of CC-NLRs. Interestingly, unlike singleton and helper CC-NLRs, sensor CC-NLRs often carry degenerated MADA α1 helix motifs and/or N-terminal domain integrations, which would preclude their capacity to trigger cell death according to the ZAR1 model (Adachi et al., 2019b; Seong et al., 2020).

Comparative sequence analyses based on a robust evolutionary framework can yield insights into molecular mechanisms and help generate experimentally testable hypotheses. ZAR1 was previously reported to be conserved across multiple dicot plant species but whether it occurs in other angiosperms hasn't been systematically studied (Baudin et al. 2017; Schultink et al. 2019; Harant et al. 2020). Here, we used a phylogenomic approach to investigate the molecular evolution of ZAR1 across flowering plants (angiosperms). We discovered 120 ZAR1 orthologs in 88 species, including monocot, magnoliid and eudicot species indicating that ZAR1 is an atypically conserved NLR that traces its origin to early angiosperm lineages ~220 to 150 million years ago (Jurassic period). We took advantage of this large collection of orthologs to identify highly conserved features of ZAR1, revealing regions for effector recognition, intramolecular interactions and cell death activation, along with a new conserved surface on the underside of the activated ZAR1 resistosome wheel. Throughout its evolution, ZAR1 also acquired novel features, including the C-terminal integration of a thioredoxin-like domain and duplication into two paralog families ZAR1-SUB and ZAR1-CIN. Members of the ZAR1-SUB paralog family have highly diversified in eudicots and often lack conserved ZAR1 features. We conclude that ZAR1 has experienced relatively limited gene duplication and

expansion throughout its deep evolutionary history, but still did give rise to non-canonical NLR proteins with integrated domains and degenerated molecular features.

RESULTS

ZAR1 is the most widely conserved CC-NLR across angiosperms

To determine the distribution of ZAR1 across plant species, we applied a computational pipeline based on iterated BLAST searches of plant genome and protein databases (Figure 1A). These comprehensive searches were seeded with previously identified ZAR1 sequences from Arabidopsis, *N. benthamiana*, tomato, sugar beet and cassava (Baudin et al. 2017; Schultink et al. 2019; Harant et al. 2020). We also performed iterated phylogenetic analyses using the NB-ARC domain of the harvested ZAR1-like sequences, and obtained a well-supported clade that includes previously reported ZAR1 from the 5 eudicots (Arabidopsis, cassava, sugar beet, tomato and *N. benthamiana*) as well as new clade members from more distantly related plant species, notably *Colacasia esculenta* (taro, Alismatales), *Cinnamomum micranthum* (Syn. *C. kanehirae*, stout camphor, Magnoliidae) and *Aquilegia coerulea* (columbine, Ranunculales) (Supplementary table 1). In total, we identified 120 ZAR1 from 88 angiosperm species that tightly clustered in the ZAR1 phylogenetic clade (Figure 1B, Supplementary table 1). Among the 120 genes, 108 code for canonical CC-NLR proteins with 52.0 to 97.0% similarity to Arabidopsis ZAR1, whereas another 9 carry the three major domains of CC-NLR proteins but have a C-terminal integrated domain (ZAR1-ID, see below). The remaining 3 genes code for two truncated NLRs and a potentially mis-annotated coding sequence due to a gap in the genome sequence. In summary, we propose that the identified clade consists of ZAR1 orthologs from a diversity of angiosperm species. Our analyses of ZAR1-like sequences also revealed two well-supported sister clades of the ZAR1 ortholog clade (Figure 1B). We named these subclades ZAR1-SUB and ZAR1-CIN and we describe them in more details below.

We have recently proposed that ZAR1 is the most conserved CC-NLR between rosid and asterid plants (Harant et al. 2020). To further evaluate ZAR1 conservation relative to other CC-NLRs across angiosperms, we used a phylogenetic tree of 1475 NLRs from the monocot taro, the magnoliid stout camphor and 6 eudicot species (columbine, Arabidopsis, cassava, sugar beet, tomato, *N. benthamiana*) to calculate the phylogenetic (patristic) distance between each of the 49 Arabidopsis CC-NLRs and their closest neighbor from each of the other plant species. We found that ZAR1 stands out for having the shortest phylogenetic distance to its orthologs relative to other CC-NLRs in this diverse angiosperm species set (Figure 1—figure supplement 1). A similar analysis where we plotted the phylogenetic distance between each of the 159 *N. benthamiana* CC-NLRs to their closest neighbor from the other species also revealed ZAR1 as displaying the shortest patristic distance across all examined species (Figure 1—figure supplement 2). These analyses revealed that ZAR1 is possibly the most widely conserved CC-NLR in flowering plants (angiosperms).

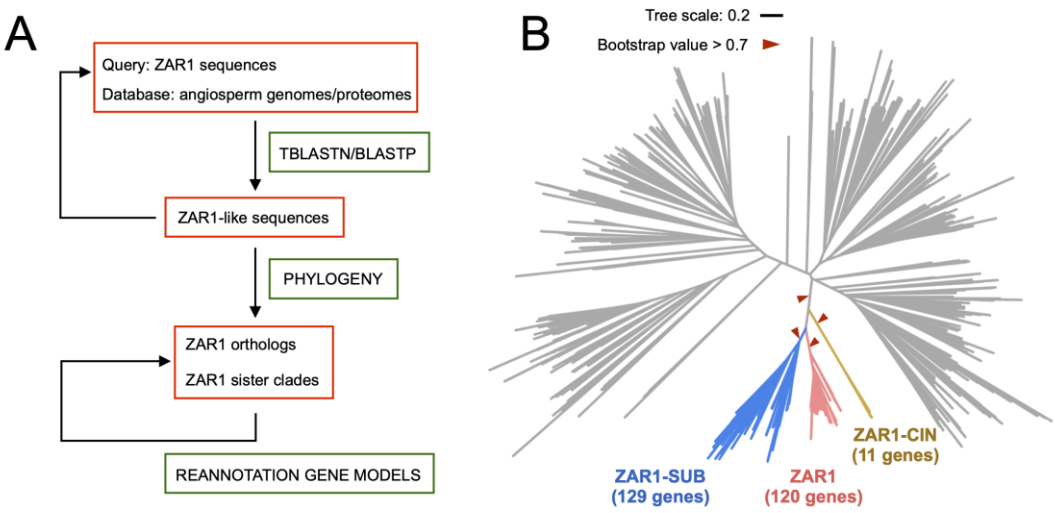


Figure 1. Comparative sequence analyses identify and classify ZAR1 sequences from angiosperms. (**A**) Workflow for computational analyses in searching ZAR1 orthologs. We performed TBLASTN/BLASTP searches and subsequent phylogenetic analyses to identify ZAR1 ortholog genes from angiosperm genome/proteome datasets. (**B**) ZAR1 forms a clade with two closely related sister subclades. The phylogenetic tree was generated in MEGA7 by the neighbour-joining method using NB-ARC domain sequences of ZAR1-like proteins identified from the prior BLAST searches and 1019 NLRs identified from 6 representative plant species, taro, stout camphor, columbine, tomato, sugar beet and Arabidopsis. Each branch is marked with different colours based on the ZAR1 and the sister subclades. Red arrow heads indicate bootstrap support > 0.7 and is shown for the relevant nodes. The scale bar indicates the evolutionary distance in amino acid substitution per site.

Phylogenetic distribution of ZAR1 in angiosperms

Although ZAR1 is distributed across a wide range of angiosperms, we noted particular patterns in its phylogenetic distribution. Supplementary table 1 describes the gene identifiers and other features of ZAR1 orthologs sorted based on the phylogenetic clades reported by Smith and Brown (2018). 68 of the 88 plant species have a single-copy of ZAR1 whereas 20 species have two or more copies. ZAR1 is primarily a eudicot gene but we identified three ZAR1 orthologs outside the eudicots, two in the monocot taro and another one in the magnoliid stout camphor. We failed to detect ZAR1 orthologs in 39 species among the 127 species we examined (Supplementary table 1). Except for taro, ZAR1 is missing in monocot species (17 examined), including in the well-studied *Hordeum vulgare* (barley), *Oryza sativa* (rice), *Triticum aestivum* (wheat) and *Zea mays* (maize). ZAR1 is also missing in all examined species of the eudicot Fabales, Cucurbitales, Apiales and Asterales. However, we found a ZAR1 ortholog in the early diverging eudicot columbine and ZAR1 is widespread in other eudicots, including in 63 rosid, 4 Caryophyllales and 18 asterid species.

ZAR1 is an ancient Jurassic gene that predates the split between monocots, magnoliids and eudicots

The overall conservation of the 120 ZAR1 orthologs enabled us to perform phylogenetic analyses using the full-length protein sequence and not just the NB-ARC domain as generally done with NLRs (Figure 2, Figure 2—figure supplement 1). These analyses yielded a robust

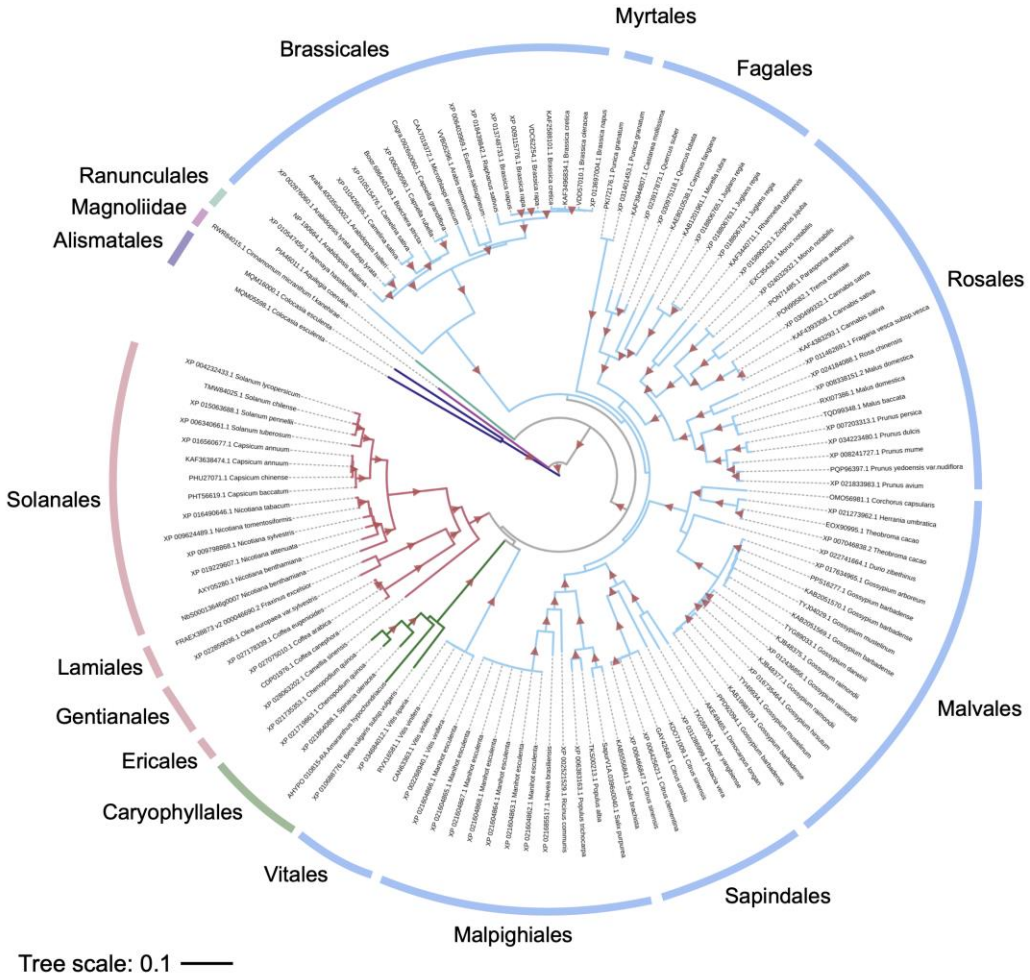


Figure 2. ZAR1 gene is distributed across angiosperms. The phylogenetic tree was generated in MEGA7 by the neighbour-joining method using full length amino acid sequences of 120 ZAR1 orthologs identified in Figure 1. Each branch is marked with different colours based on the plant taxonomy. Red triangles indicate bootstrap support > 0.7. The scale bar indicates the evolutionary distance in amino acid substitution per site.

ZAR1 phylogenetic tree with well-supported branches that generally mirrored established phylogenetic relationships between the examined plant species (Smith and Brown, 2018; Chaw et al., 2019). For example, the ZAR1 tree matched a previously published species tree of angiosperms based on 211 single-copy core ortholog genes (Chaw et al., 2019). We conclude that the origin of the ZAR1 gene predates the split between monocots, magnoliids and eudicots and its evolution traced species divergence ever since. We postulate that ZAR1 probably emerged in the Jurassic era ~220 to 150 million years ago (Mya) based on the species divergence time estimate of Chaw et al. (2019) and consistent with the latest fossil evidence for the emergence of flowering plants (Fu et al., 2018).

ZAR1 is a genetic singleton in a locus that exhibits gene co-linearity across eudicot species

NLR genes are often clustered in loci that are thought to accelerate sequence diversification and evolution (Michelmore and Meyers, 1998; Lee and Chae, 2020). We examined the genetic context of ZAR1 genes using available genome assemblies of taro, stout camphor, columbine, Arabidopsis, cassava, sugar beet, tomato and *N. benthamiana*. The ZAR1 locus is generally

devoid of other NLR genes as the closest NLR is found in the Arabidopsis genome 183 kb away from ZAR1 (Figure 2—figure supplement 2—supplementary table 1). We conclude that ZAR1 has probably remained a genetic singleton NLR gene throughout its evolutionary history in angiosperms.

Next, we examined the ZAR1 locus for gene co-linearity across the examined species. We noted a limited degree of gene co-linearity between Arabidopsis vs. cassava, cassava vs. tomato, and tomato vs. *N. benthamiana* (Figure 2—figure supplement 2). Flanking conserved genes include the ATPase and protein kinase genes that are present at the ZAR1 locus in both rosid and asterid eudicots. In contrast, we didn't observe conserved gene blocks at the ZAR1 locus of taro, stout camphor and columbine, indicating that this locus is divergent in these species. Overall, although limited, the observed gene co-linearity in eudicots is consistent with the conclusion that ZAR1 is a genetic singleton with an ancient origin.

ZAR1 orthologs carry sequence motifs known to be required for Arabidopsis ZAR1 resistosome function

The overall sequence conservation and deep evolutionary origin of ZAR1 orthologs combined with the detailed knowledge of ZAR1 structure and function provide a unique opportunity to explore the evolutionary dynamics of this ancient immune receptor in a manner that cannot be applied to more rapidly evolving NLRs. We used MEME (Multiple EM for Motif Elicitation) (Bailey and Elkan, 1994) to search for conserved sequence patterns among the 117 ZAR1 orthologs (ZAR1 and ZAR1-ID) that encode full-length CC-NLR proteins. This analysis revealed several conserved sequence motifs that span across the ZAR1 orthologs (range of protein lengths: 753-1132 amino acids) (Figure 3A, Figure 3—supplementary table 1). In Figure 3A, we described the major five sequence motifs or interfaces known to be required for Arabidopsis ZAR1 function that are conserved across ZAR1 orthologs.

Effector recognition by ZAR1 occurs indirectly via binding to RLCKs through the LRR domain. Key residues in the Arabidopsis ZAR1-RLCK interfaces are highly conserved among ZAR1 orthologs and were identified by MEME as conserved sequence patterns (Figure 3A). Valine (V) 544, histidine (H) 597, tryptophan (W) 825 and phenylalanine (F) 839 in the Arabidopsis ZAR1 LRR domain were validated by mutagenesis as important residues for RLCK binding whereas isoleucine (I) 600 was not essential (Wang et al. 2019a; Hu et al. 2020). In the 117 ZAR1 orthologs, V544, H597, W825 and F839 are conserved in 97-100% of the proteins compared to only 63% for I600.

After effector recognition, Arabidopsis ZAR1 undergoes conformational changes from inactive to active state. This is mediated by ADP release from the NB-ARC domain and subsequent ATP binding, which triggers further structural remodelling of ZAR1 into the pentameric resistosome (Wang et al. 2019b). NB-ARC sequences that coordinate binding and hydrolysis of dATP, namely P-loop and MHD motifs, are highly conserved across ZAR1 orthologs (Figure 3A). Histidine (H) 488 and lysine (K) 195, located in the ADP/ATP binding pocket (Wang et al. 2019a; Wang et al. 2019b), are invariant in all 117 orthologs. In addition, three NB-ARC residues, W150, S152 and V154, known to form the NBD-NBD oligomerisation

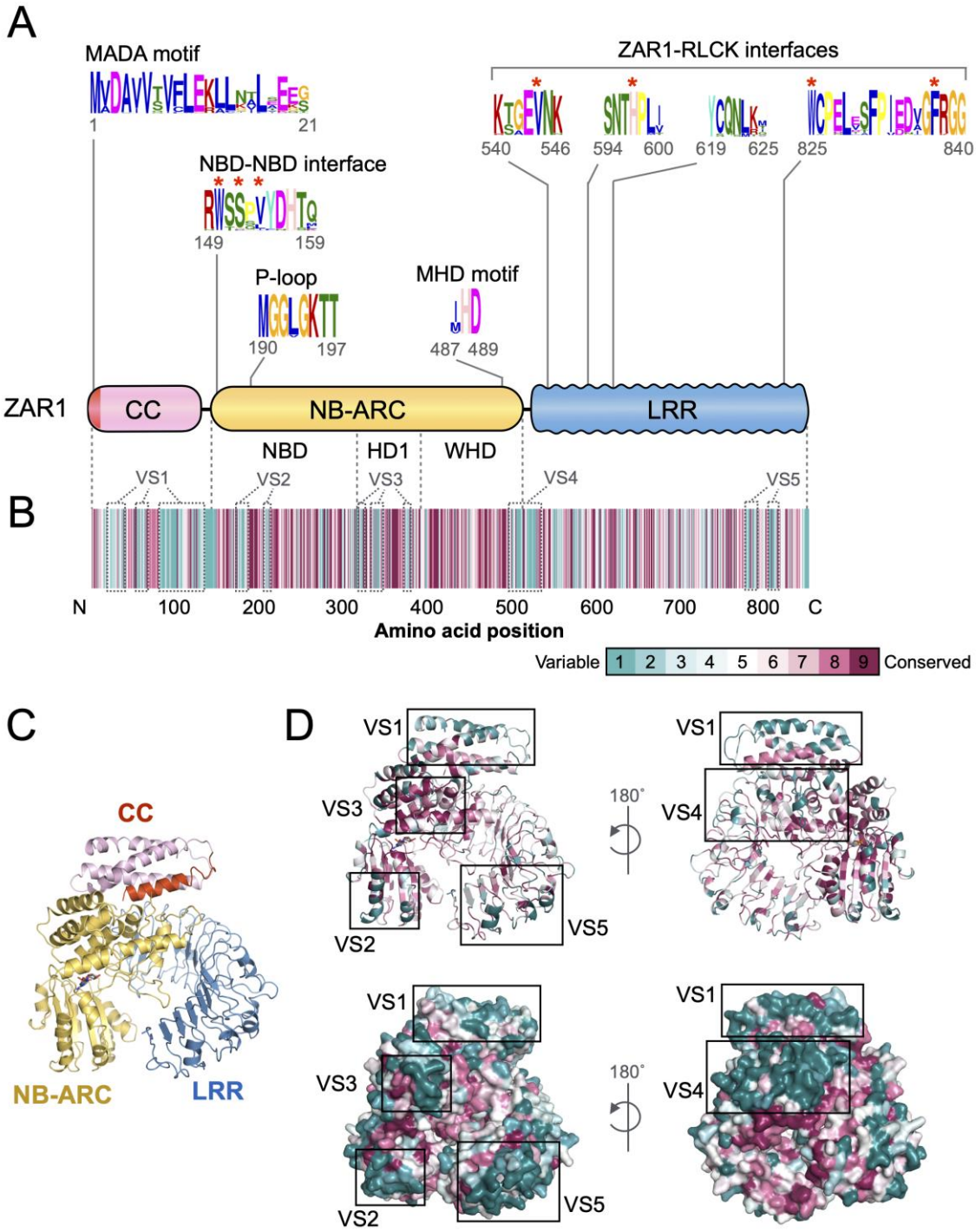


Figure 3. ZAR1 orthologs carry conserved sequence patterns required for Arabidopsis ZAR1 resistosome function. (**A**) Schematic representation of the Arabidopsis ZAR1 protein highlighting the position of conserved sequence patterns across ZAR1 orthologs. Consensus sequence patterns were identified by MEME using 117 ZAR1 ortholog sequences. Raw MEME motifs are listed in Figure 3—Supplementary table 1. Red asterisks indicate residues functionally validated in Arabidopsis ZAR1 for NBD-NBD and ZAR1-RLCK interfaces. (**B**) Conservation and variation of each amino acid among ZAR1 orthologs across angiosperms. Amino acid alignment of 117 ZAR1 orthologs was used for conservation score calculation via the ConSurf server (https://consurf.tau.ac.il). The conservation scores are mapped onto each amino acid position in Arabidopsis ZAR1 (NP_190664.1). (**C, D**) Distribution of the ConSurf conservation score on the Arabidopsis ZAR1 structure. The inactive ZAR1 monomer is illustrated in cartoon representation with different colours based on each canonical domain (C) and the conservation score (D). Major five variable surfaces (VS1 to VS5) on the inactive ZAR1 monomer structure are described in grey dot or black boxes in panel B or D, respectively.

interface for resistosome formation (Wang et al. 2019b; Hu et al. 2020), are present in 82-97% of the ZAR1 orthologs and were also part of a MEME motif (Figure 3A).

The N-terminal CC domain of Arabidopsis ZAR1 mediates cell death signalling thorough the N-terminal α1 helix/MADA motif, that becomes exposed in activated ZAR1 resistosome to form a funnel like structure that perturbs the plasma membrane (Baudin et al., 2017; 2019; Wang et al. 2019b; Adachi et al., 2019b). We detected an N-terminal MEME motif that matches the α1 helix/MADA motif (Figure 3A). We also used the HMMER software (Eddy, 1998) to query the ZAR1 orthologs with a previously reported MADA motif-Hidden Markov Model (HMM) (Adachi et al., 2019b). This HMMER search detected a MADA-like sequence at the N-terminus of all 117 ZAR1 orthologs (Supplementary table 1).

Taken together, based on the conserved motifs depicted in Figure 3A, we propose that angiosperm ZAR1 orthologs share the main functional features of Arabidopsis ZAR1: 1) effector recognition via RLCK binding, 2) remodelling of intramolecular interactions via ADP/ATP switch, 3) oligomerisation via the NBD-NBD interface and 4) α1 helix/MADA motif-mediated activation of hypersensitive cell death.

A novel conserved surface on the underside of the ZAR1 resistosome

To identify additional conserved and variable features in ZAR1 orthologs, we used ConSurf (Ashkenazy et al., 2016) to calculate a conservation score for each amino acid and generate a diversity barcode for ZAR1 orthologs (Figure 3B). The overall pattern is that the 117 ZAR1 orthologs are fairly conserved. Nonetheless, the CC domain (except for the N-terminal MADA motif and a few conserved stretches), the junction between the NB-ARC and LRR domains and the very C-terminus were distinctly more variable than the rest of the protein (Figure 3B).

We also used the cryo-EM structures of Arabidopsis ZAR1 to determine how the ConSurf score map onto the 3D structures (Figure 3C, D and Figure 4). First, we found five major variable surfaces (VS1 to VS5) on the inactive ZAR1 monomer structure (Figure 3C, D), as depicted in the ZAR1 diversity barcode (Figure 3B). VS1 comprises α2/α4 helices and a loop between α3 and α4 helices of the CC domain. VS2 and VS3 corresponds to α1/α2 helices of NBD and a loop between α2 and α3 helices of HD1, respectively. VS4 comprises a loop between WHD and LRR and first three helices of the LRR domain. VS5 is mainly derived from the last three helices of the LRR domain and the loops between these helices (Figure 3B, D).

We also noted significant sequence variation at the glutamate rings inside the Arabidopsis ZAR1 resistosome (Figure 4—figure supplement 1). Mutations of glutamic acid (E) 11 and E18 impaired Arabidopsis ZAR1-mediated cell death without interfering with oligomerization and plasma membrane association (Wang et al. 2019b). The E130/E134 ring was previously discussed as potentially having Ca^{2+} transporter activity because of structural similarity to rings in the structures of the mitochondrial calcium uniporter from *Caenorhabditis elegans* and the calcium release-activated calcium channel ORAI from *Drosophila melanogaster* (Burdett et al., 2019). Whereas E11 is conserved in 94% of ZAR1 orthologs, only 3-18% retain E18, E130 and E134 in the same positions as Arabidopsis ZAR1.

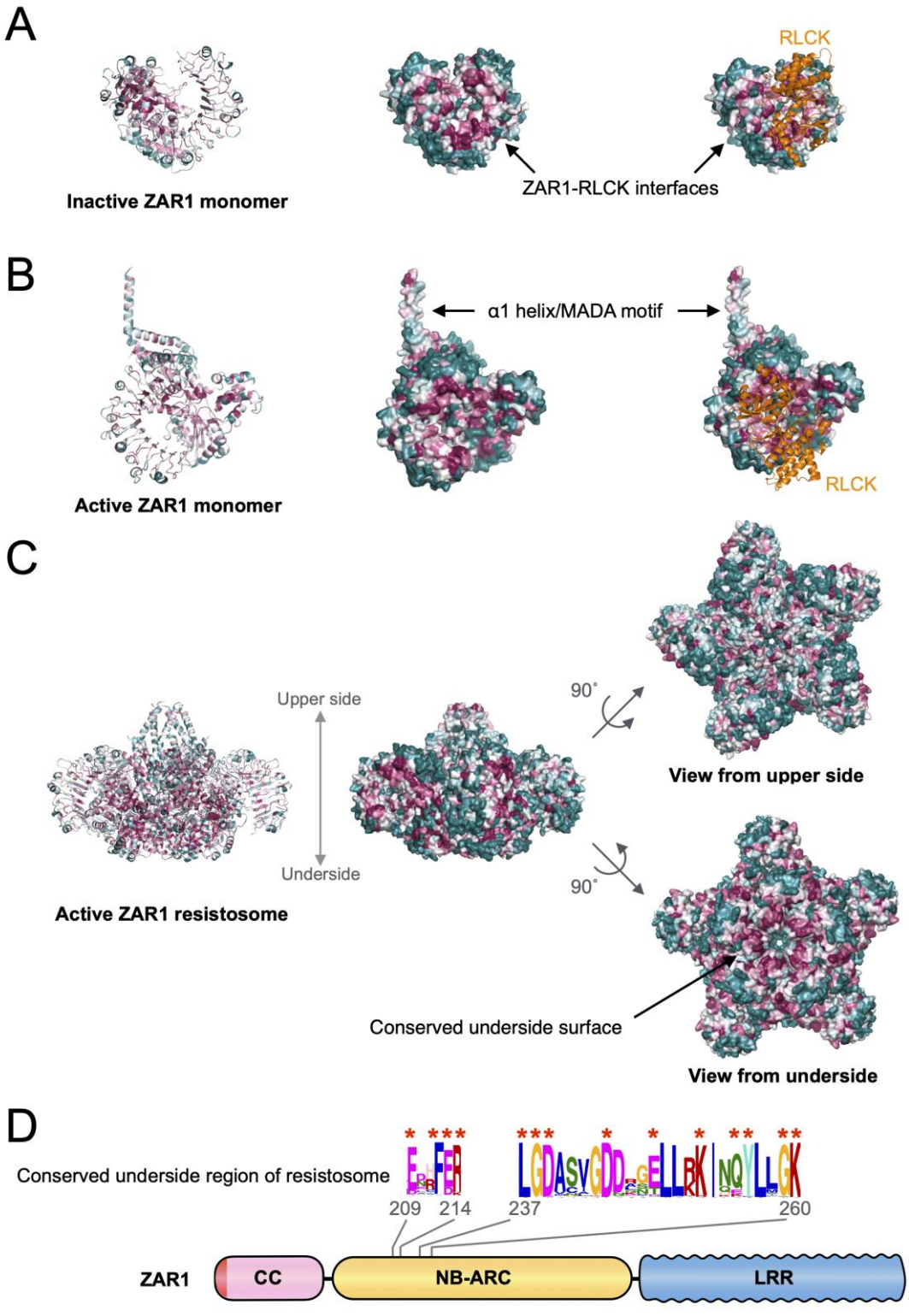


Figure 4. ZAR1 orthologs across angiosperms display multiple conserved surfaces on the resistosome structure. Distribution of the ConSurf conservation score was visualized on the inactive monomer (**A**), active monomer (**B**) and resistosome (**C**) structures of Arabidopsis ZAR1. Each structure and cartoon representation are illustrated with different colours based on the conservation score shown in Figure 3. (**D**) Schematic representation of the conserved underside surface region among ZAR1 orthologs. The conserved underside regions are described with consensus sequence patterns identified by MEME. Red asterisks indicate residues exposed on resistosome surfaces. The raw MEME motif is listed in Figure 3—Supplementary table 1.

Next, we examined highly conserved surfaces on inactive and active ZAR1 structures (Figure 4A, B). Consistent with the MEME analyses, we confirmed that highly conserved surfaces match to the RLCK binding interfaces (Figure 4A, B). We also confirmed that the N-terminal α1 helix/MADA motif is conserved on the resistosome surfaces, although the first four N-terminal amino acids are missing from the N terminus of the active ZAR1 cryo-EM structures (Figure 4B).

Remarkably, these analyses revealed a highly conserved ring that is exposed on the underside surface of the ZAR1 resistosome opposite to the funnel-shaped structure (Figure 4C). This conserved underside surface is mainly formed by α2 helix-loop-ß2 sheet and α3 helix-loop-α4 helix regions in the NBD (Figure 4—figure supplement 2 and 3). Within the conserved patch, arginine (R) 214, K252 and K260, are positively charged residues that are exposed on the underside surface and are conserved in 94%, 96% and 99% of ZAR1 orthologs, respectively (Figure 4—figure supplement 4). The three residues form a positive electrostatic potential ring on the underside surface of the ZAR1 resistosomes (Figure 4—figure supplement 4). The conserved underside ring is composed of a 14 amino acid motif as revealed by MEME (Figure 4D). We propose that this underside sequence pattern has been maintained throughout the more than 150 million years of ZAR1 evolution and is likely to be functionally important.

Integration of a PLP3a thioredoxin-like domain at the C-termini of cassava and cotton ZAR1

As noted earlier, 9 ZAR1 orthologs carry an integrated domain (ID) at their C-termini (Supplementary table 1). These ZAR1-ID include 2 predicted proteins (XP_021604862.1 and XP_021604864.1) from *Manihot esculenta* (cassava) and 7 predicted proteins (KAB1998109.1, PPD92094.1, KAB2051569.1, TYG89033.1, TYI49934.1, TYJ04029.1, KJB48375.1) from the cotton plant species *Gossypium barbadense*, *Gossypium darwinii*, *Gossypium mustelinum* and *Gossypium raimondii* (Supplementary table 1). The integrations follow an otherwise intact LRR domain and vary in length from 108 to 266 amino acids (Figure 5A). We confirmed that the ZAR1-ID gene models of cassava XP_021604862.1 and XP_021604864.1 are correct based on RNA-seq exon coverage in the NCBI database (database ID: LOC110609538). However, cassava ZAR1-ID XP_021604862.1 and XP_021604864.1 are isoforms encoded by transcripts from a single locus on chromosome LG2 (RefSeq sequence NC_035162.1) of the cassava RefSeq assembly (GCF_001659605.1) which also produces transcripts encoding isoforms lacking the C-terminal ID (XP_021604863.1, XP_021604865.1, XP_021604866.1, XP_021604867.1 and XP_021604868.1). Thus, cassava ZAR1-ID are probably splicing variants from a unique cassava ZAR1 gene locus (Figure 5—figure supplement 1).

To determine the phylogenetic relationship between ZAR1-ID and canonical ZAR1, we mapped the domain architectures of ZAR1 orthologs on the phylogenetic tree shown in Figure 2 (Figure 5—figure supplement 2). Cassava and cotton ZAR1-ID occur in different branches of the ZAR1 rosid clade indicating that they may have evolved as independent integrations although alternative evolutionary scenarios such as a common origin followed by subsequent deletion of the ID or lineage sorting remain possible (Figure 5—figure supplement 2).

We annotated all the C-terminal extensions as thioredoxin-like using InterProScan (Trx, IPR036249; IPR013766; cd02989). The integrated Trx domain sequences share sequence similarity to each other (Figure 5B). They are also similar to Arabidopsis AT3G50960

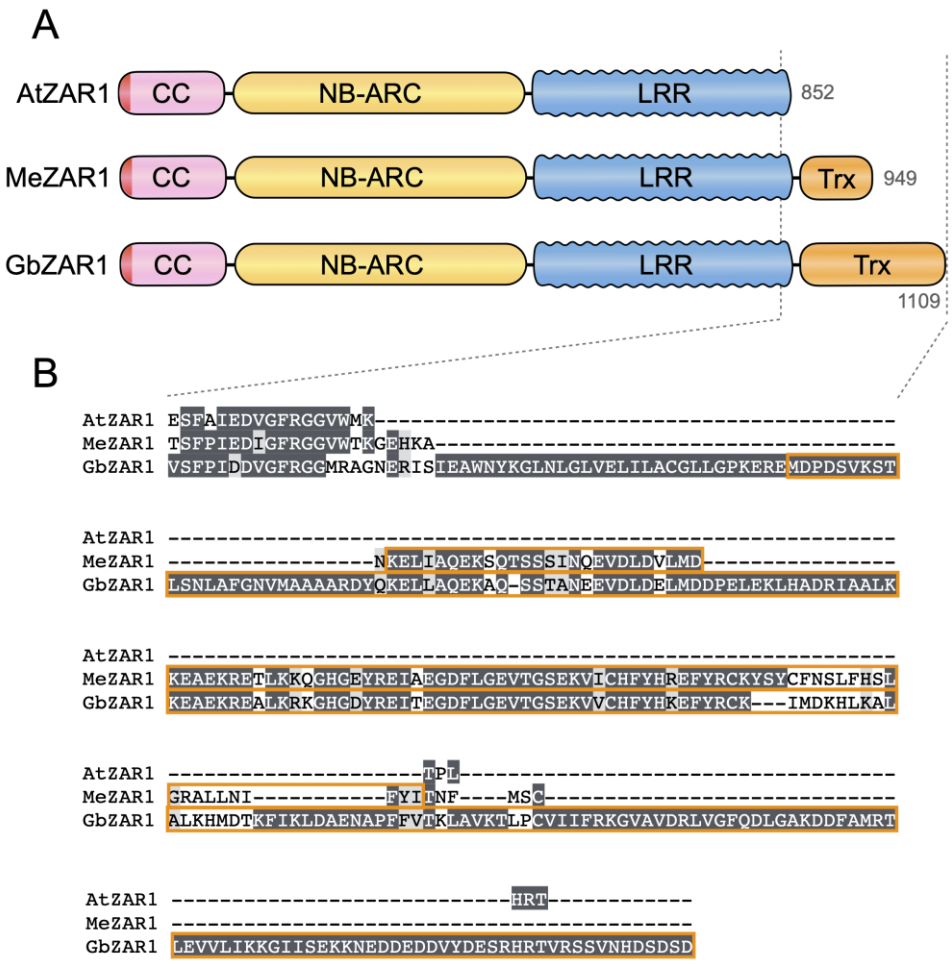


Figure 5. Cassava and cotton ZAR1-ID carry an additional Trx domain at the C terminus. (**A**) Schematic representation of NLR domain architecture with C-terminal Trx domain. (**B**) Description of Trx domain sequences on amino acid sequence alignment. Cassava XP_021604862.1 (MeZAR1) and cotton KAB1998109.1 (GbZAR1) were used for MAFFT version 7 alignment as representative ZAR1-ID. Arabidopsis ZAR1 (AtZAR1) was used as a control of ZAR1 without ID.

(phosphoducin-like PLP3a; 34.8-90% similarity to integrated Trx domains), which is located immediately downstream of ZAR1 in a tail-to-tail configuration in the Arabidopsis genome (Figure 2—figure supplement 2; Figure 5—figure supplement 3). We also noted additional genetic linkage between ZAR1 and Trx genes in other rosid species, namely field mustard, orange, cacao, grapevine and apple, and in the asterid species coffee (Figure 2—figure supplement 2—supplementary table 2). We conclude that ZAR1 is often genetically linked to a PLP3a-like Trx domain gene and that the integrated domain in ZAR1-ID has probably originated from a genetically linked sequence.

The ZAR1-SUB clade emerged early in eudicot evolution from a single ZAR1 duplication event

Phylogenetic analyses revealed ZAR1-SUB as a sister clade of the ZAR1 ortholog clade (Figure 1B, Figure 6). ZAR1-SUB clade comprises 129 genes from a total of 55 plant species (Supplementary table 2). 21 of the 55 plant species carry a single-copy of ZAR1-SUB whereas

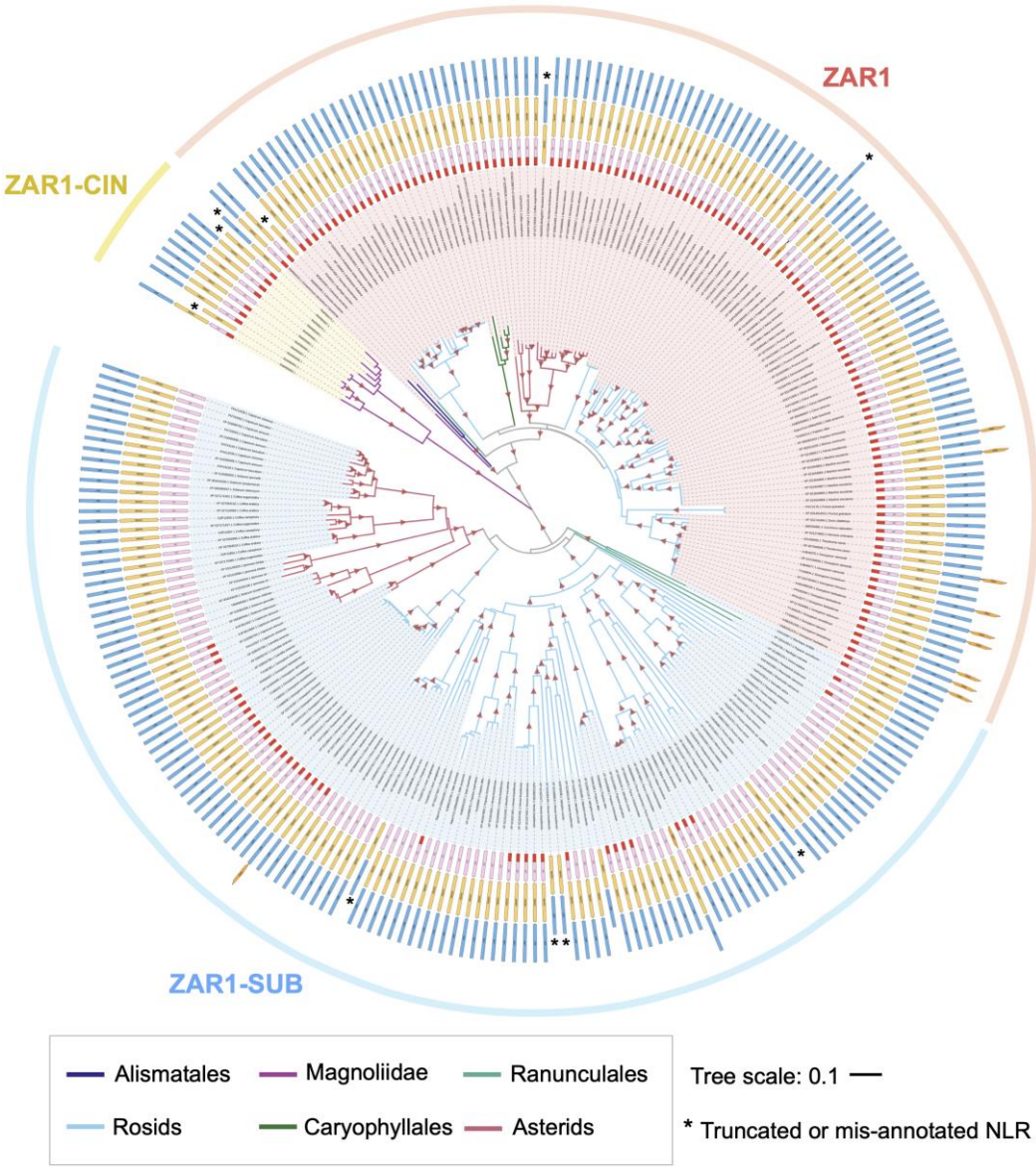


Figure 6. ZAR1-SUB has emerged early in eudicots and diverged at MADA motif sequence. The phylogenetic tree was generated in MEGA7 by the neighbour-joining method using full length amino acid sequences of 120 ZAR1, 129 ZAR1-SUB and 11 ZAR1-CIN identified in Figure 1. Each branch is marked with different colours based on the plant taxonomy. Red triangles indicate bootstrap support > 0.7. The scale bar indicates the evolutionary distance in amino acid substitution per site. NLR domain architectures are illustrated outside of the leaf labels: MADA is red, CC is pink, NB-ARC is yellow, LRR is blue and other domain is orange. Black asterisks on domain schemes describe truncated NLRs or potentially mis-annotated NLR.

34 species have two or more copies. Of the 129 genes, 122 code for canonical CC-NLR proteins (692-1038 amino acid length) with shared sequence similarities ranging from 36.5 to 99.9% (Figure 6).

Unlike ZAR1, ZAR1-SUB NLRs are restricted to eudicots (Figure 6—figure supplement 1, Supplementary table 2). Three out of 129 genes are from the early diverging eudicot clade Ranunculales species, namely columbine, *Macleaya cordata* (plume poppy) and *Papaver somniferum* (opium poppy) (Figure 6—figure supplement 1). The remaining ZAR1-SUB are

spread across rosid and asterid species (Figure 6—figure supplement 1). We found that 11 species have ZAR1-SUB genes but lack a ZAR1 ortholog (Supplementary table 3). These 11 species include two of the early diverging eudicots plume poppy and opium poppy, and the Brassicales *Carica papaya* (papaya). Interestingly, papaya is the only Brassicales species carrying a ZAR1-SUB gene, whereas the 16 other Brassicales species have ZAR1 but lack ZAR1-SUB genes (Figure 6—figure supplement 1, Supplementary table 3). In total, we didn't detect ZAR1-SUB genes in 44 species that have ZAR1 orthologs, and these 44 species include the monocot taro, the magnoliid stout camphor and 42 eudicots, such as Arabidopsis, sugar beet and *N. benthamiana* (Supplementary table 3).

In summary, given the taxonomic distribution of the ZAR1-SUB clade genes, we propose that ZAR1-SUB has emerged from a single duplication event of ZAR1 prior to the split between Ranunculales and other eudicot lineages about ~120-130 Mya based on the species divergence time estimate of Chaw et al. (2019).

ZAR1-SUB paralogs have significantly diverged from ZAR1

We investigated the sequence patterns of ZAR1-SUB proteins and compared them to the sequence features of canonical ZAR1 proteins that we identified earlier (Figures 3, 4). MEME analyses revealed several conserved sequence motifs (Figure 7—supplementary table 1). Especially, the MEME motifs in the ZAR1-SUB NB-ARC domain were similar to ZAR1 ortholog motifs (Figure 7—supplementary table 2). These include P-loop and MHD motifs, which are broadly conserved in NB-ARC of 97% and 100% of the ZAR1-SUB NLRs, respectively (Figure 7A). MEME also revealed sequence motifs in the ZAR1-SUB LRR domain that partially overlaps in position with the conserved ZAR1-RLCK interfaces (Figure 7A, Figure 7—figure supplement 1). However, the ZAR1-SUB MEME motifs in the LRR domain were variable at the ZAR1-RLCK interface positions compared to ZAR1, and the motif sequences were markedly different between ZAR1-SUB and ZAR1 proteins (Figures 3A, 7A, Figure 7—supplementary table 2).

Remarkably, unlike ZAR1 orthologs, MEME did not predict conserved sequence pattern from a region corresponding to the MADA motif, indicating that these sequences have diverged across ZAR1-SUB proteins (Figure 7A). We confirmed the low frequency of MADA motifs in ZAR1-SUB proteins using HMMER searches with only ~30% (38 out of 129) of the tested proteins having a MADA-like sequence (Supplementary table 2, Figure 6). Moreover, conserved sequence patterns were not predicted for the NBD-NBD interface and the conserved underside surface of the ZAR1 resistosome (Figure 7A, Figure 7—figure supplement 1). This indicates that the NB-ARC domain of ZAR1-SUB proteins is highly diversified in contrast to the relatively conserved equivalent region of ZAR1 proteins.

We generated a diversity barcode for ZAR1-SUB proteins using the ConSurf as we did earlier with ZAR1 orthologs (Figure 7B). This revealed that there are several conserved sequence blocks in each of the CC, NB-ARC and LRR domains, such as the regions corresponding to P-loop, MHD motif and the equivalent of the ZAR1-RLCK interfaces. Nonetheless, ZAR1-SUB proteins are overall more diverse than ZAR1 orthologs especially in the CC domain, including the N-terminal MADA motif, and the NBD/HD1 regions of the NB-ARC domain where the NBD-NBD interface is located.

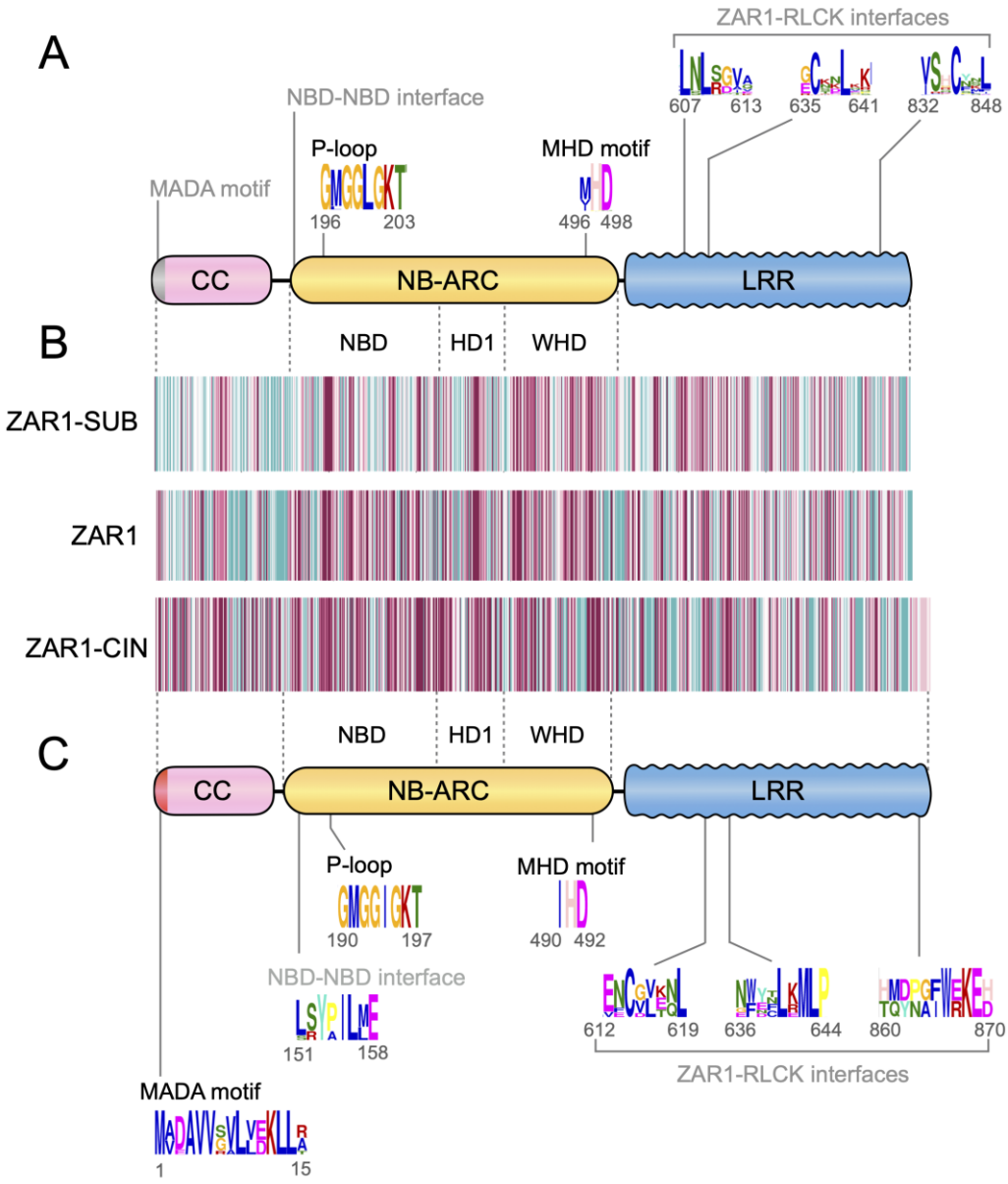


Figure 7. Conserved sequence distributions in ZAR1-SUB and ZAR1-CIN. (**A**) Schematic representation of the ZAR1-SUB protein highlighting the position of the representative conserved sequence patterns across ZAR1-SUB. Representative consensus sequence patterns identified by MEME are described on the scheme. Raw MEME motifs are listed in Figure 7–supplementary tables 1 and 2. (**B**) Conservation and variation of each amino acid among ZAR1-SUB and ZAR1-CIN. Amino acid alignment of 129 ZAR1-SUB or 8 ZAR1-CIN was used for conservation score calculation via the ConSurf server (https://consurf.tau.ac.il). The conservation scores are mapped onto each amino acid position in queries XP_004243429.1 (ZAR1-SUB) and RWR85656.1 (ZAR1-CIN), respectively. (**C**) Schematic representation of the ZAR1-CIN protein highlighting the position of the representative conserved sequence patterns across 8 ZAR1-CIN. Raw MEME motifs are listed in Figure 7–supplementary tables 3 and 4.

Next, we mapped the ConSurf conservation scores onto a homology model of a representative ZAR1-SUB protein (XP_004243429.1 from tomato) built based on the Arabidopsis ZAR1 cryo-EM structures (Figure 8). As highlighted in Figure 8B and C, conserved residues, such as MHD motif region in the WHD, are located inside of the monomer and resistosome structures. Interestingly, although the prior MEME prediction analyses revealed

conserved motifs in positions matching the ZAR1-RLCK interfaces in the LRR domain, the ZAR1-SUB structure homology models displayed variable surfaces in this region (Figures 7A, 8A). This indicates that the variable residues within these sequence motifs are predicted to be on the outer surfaces of the LRR domain and may reflect interaction with different ligands.

Taken together, these results suggest that unlike ZAR1 orthologs, the ZAR1-SUB paralogs have divergent molecular patterns for regions known to be involved in effector recognition, resistosome formation and activation of hypersensitive cell death.

Eleven tandemly duplicated ZAR1-CIN genes occur in a 500 kb cluster in the *Cinnamomum micranthum* (stout camphor) genome

The ZAR1-CIN clade, identified by phylogenetic analyses as a sister clade to ZAR1 and ZAR1-SUB, consists of 11 genes from the magnoliid species stout camphor (Figure 1B, Figure 6, Supplementary table 4). 8 of the 11 ZAR1-CIN genes code for canonical CC-NLR proteins with 63.8 to 98.9% sequence similarities to each other, whereas the remaining 3 genes code for truncated NLR proteins. Interestingly, all ZAR1-CIN genes occur in a ~500 kb cluster on scaffold QPKB01000005.1 of the stout camphor genome assembly (GenBank assembly accession GCA_003546025.1) (Figure 6—figure supplement 2). This scaffold also contains the stout camphor ZAR1 ortholog (CmZAR1, RWR84015), which is located 48 Mb from the ZAR1-CIN cluster (Figure 6—figure supplement 2). Based on the observed phylogeny and gene clustering, we suggest that the ZAR1-CIN cluster emerged from segmental duplication and expansion of the ancestral ZAR1 gene after stout camphor split from the other examined ZAR1 containing species.

Tandemly duplicated ZAR1-CIN display variable ligand binding interfaces on the LRR domain

We performed MEME and ConSurf analyses of the 8 intact ZAR1-CIN proteins as described above for ZAR1 and ZAR1-SUB. The ConSurf barcode revealed that although ZAR1-CIN proteins are overall conserved, their WHD region and LRR domain include some clearly variable blocks (Figure 7B). MEME analyses of ZAR1-CIN sequences revealed that like ZAR1 orthologs, the MADA, P-loop and MHD motifs match highly conserved blocks of the ZAR1-CIN ConSurf barcode (Figure 7B, C, Figure 7—supplementary tables 3 and 4). Consistently, 87.5% (7 out of 8) of the ZAR1-CIN proteins were predicted to have a MADA-type N-terminal sequence based on MADA-HMM analyses (Supplementary table 4, Figure 6).

MEME picked up additional sequence motifs in ZAR1-CIN proteins that overlap in position with the NBD-NBD and ZAR1-RLCK interfaces (Figure 7C, Figure 7—figure supplement 2). However, the sequence consensus at the NBD-NBD and ZAR1-RLCK interfaces indicated these motifs are more variable among ZAR1-CIN proteins relative to ZAR1 orthologs, and the motif sequences were markedly different from the matching region in ZAR1 (Figures 3A, 7C).

We also mapped the ConSurf conservation scores onto a homology model of a representative ZAR1-CIN protein (RWR85656.1) built based on the Arabidopsis ZAR1 cryo-EM structures (Figure 8). This model revealed several conserved surfaces, such as on the α1 helix in the CC domain, the WHD of the NB-ARC domain and underside surface of the resistosome (Figure

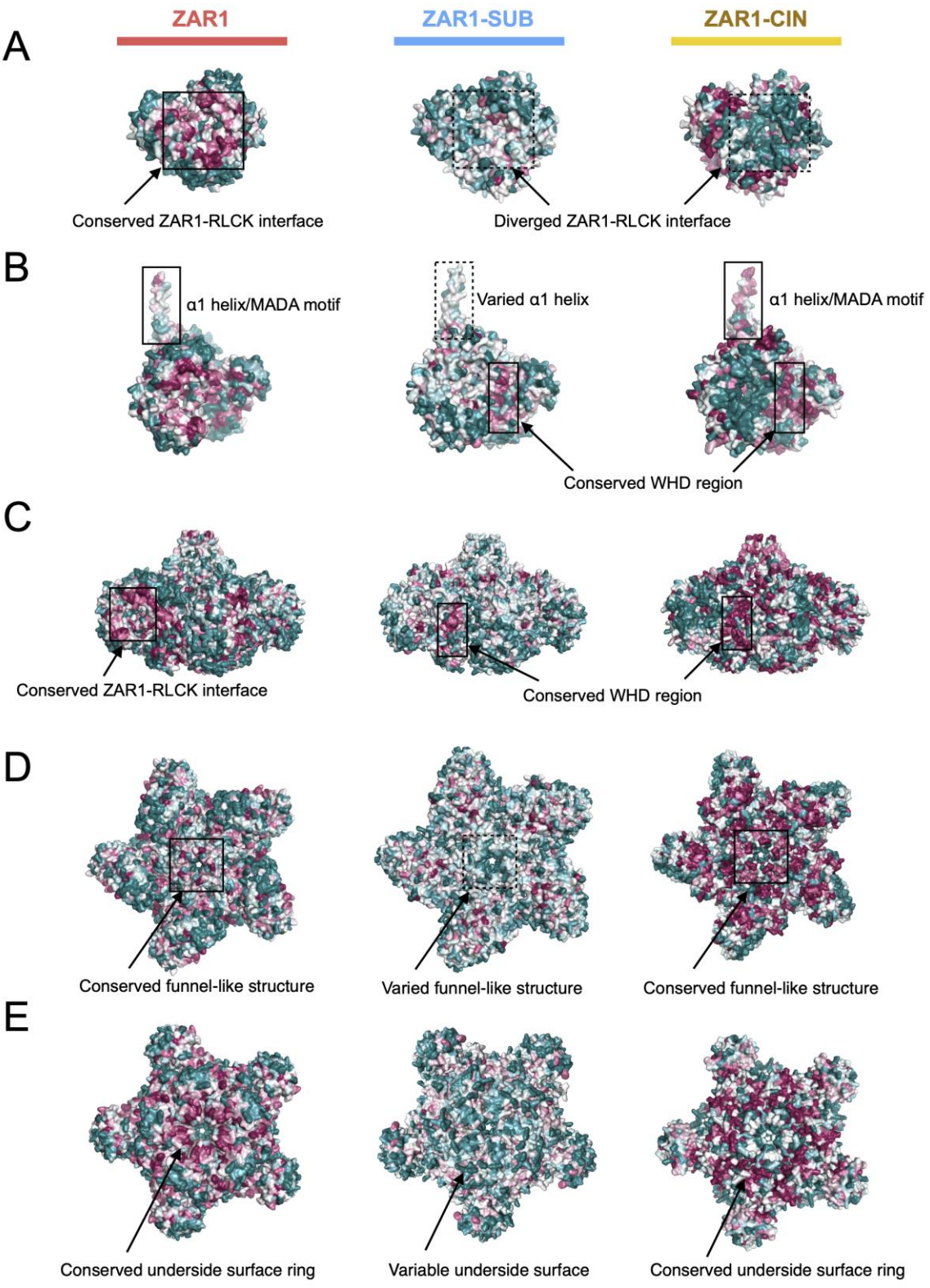


Figure 8. ZAR1 and the sister subclade NLRs display different conserved surfaces on the resistosome structure. Distribution of the ConSurf conservation score was visualized on the inactive monomer (**A**), active monomer (**B**) and resistosome structures (**C-E**) of Arabidopsis ZAR1 or the structure homology models of ZAR1-SUB (XP_004243429.1) and ZAR1-CIN (RWR85656.1). Each structure and cartoon representation are illustrated with different colours based on the conservation score shown in Figures 3 and 7. Resistosome structures are shown from different angles, from side (C), from upper side (D) and from underside (E).

8B, C, E). In contrast, the ZAR1-CIN structure homology models displayed highly varied surfaces especially in the LRR region matching the RLCK binding interfaces of ZAR1 (Figure 8A). This sequence diversification on the LRR surface suggests that the ZAR1-CIN paralogs may have different host partner proteins and/or effector recognition specificities compared to ZAR1.

DISCUSSION

This study originated from phylogenomic analyses we initiated during the COVID-19 lockdown of March 2020. We performed iterated comparative sequence similarity searches of plant genomes using the CC-NLR immune receptor ZAR1 as a query, and subsequent phylogenetic evaluation of the recovered ZAR1-like sequences. This revealed that ZAR1 is an ancient gene with 120 orthologs recovered from 88 species including monocot, magnoliid and eudicot plants. ZAR1 is an atypically conserved NLR in these species with the gene phylogeny tracing species phylogeny, and consistent with the view that ZAR1 originated early in angiosperms during the Jurassic geologic period ~220 to 150 Mya (Figure 9). The ortholog series enabled us to determine that resistosome sequences that are known to be functionally important and have remained highly conserved throughout the long evolutionary history of ZAR1. This also revealed a new conserved sequence ring on the underside of the resistosome, which has remained constrained in ZAR1 ortholog proteins (Figure 4). The only unexpected feature among ZAR1 orthologs is the acquisition of a C-terminal thioredoxin-like domain in cassava and cotton species (Figures 5, 9). Our phylogenetic analyses also indicated that ZAR1 duplicated twice throughout its evolution (Figure 9). In the eudicots, ZAR1 spawned a large paralog family, ZAR1-SUB, which greatly diversified and often lost the typical sequence features of ZAR1. A second paralog, ZAR1-CIN, is restricted to a tandemly repeated 11-gene cluster in stout camphor. Overall, our findings map patterns of functional conservation, expansion and diversification onto the evolutionary history of ZAR1 and its paralogs (Figure 9). Phylogenomics analyses, such as this work, provide a unique evolutionary perspective on the function of a plant NLR immune receptor and generate experimentally testable hypotheses that can be challenged in the future.

ZAR1 most likely emerged prior to the split between monocots, Magnoliids and eudicots, which corresponds to ~220 to 150 Mya based on the dating analyses of Chaw et al. (2019). The origin of the angiosperms remains hotly debated with uncertainties surrounding some of the fossil record coupled with molecular clock analyses that would benefit from additional genome sequences of undersampled taxa. However, recently Fu et al. (2018) provided credence to an earlier emergence of angiosperms with the discovery of the fossil flower *Nanjinganthus dendrostyla*, which places the emergence of flowering plants at the Early Jurassic. It is tempting to speculate that ZAR1 emerged among these early flowering plants during the period when dinosaurs dominated planet earth.

NLRs are notorious for their rapid and dynamic evolutionary patterns. In contrast, ZAR1 is an atypically core NLR gene conserved in a wide range of angiosperm species (Figures 3 and 4). Nevertheless, Arabidopsis ZAR1 can recognize diverse bacterial pathogen effectors, including five different effector families distributed among nearly half of a collection of ~500 *Pseudomonas syringae* strains (Laflamme et al., 2020) and an effector AvrAC from

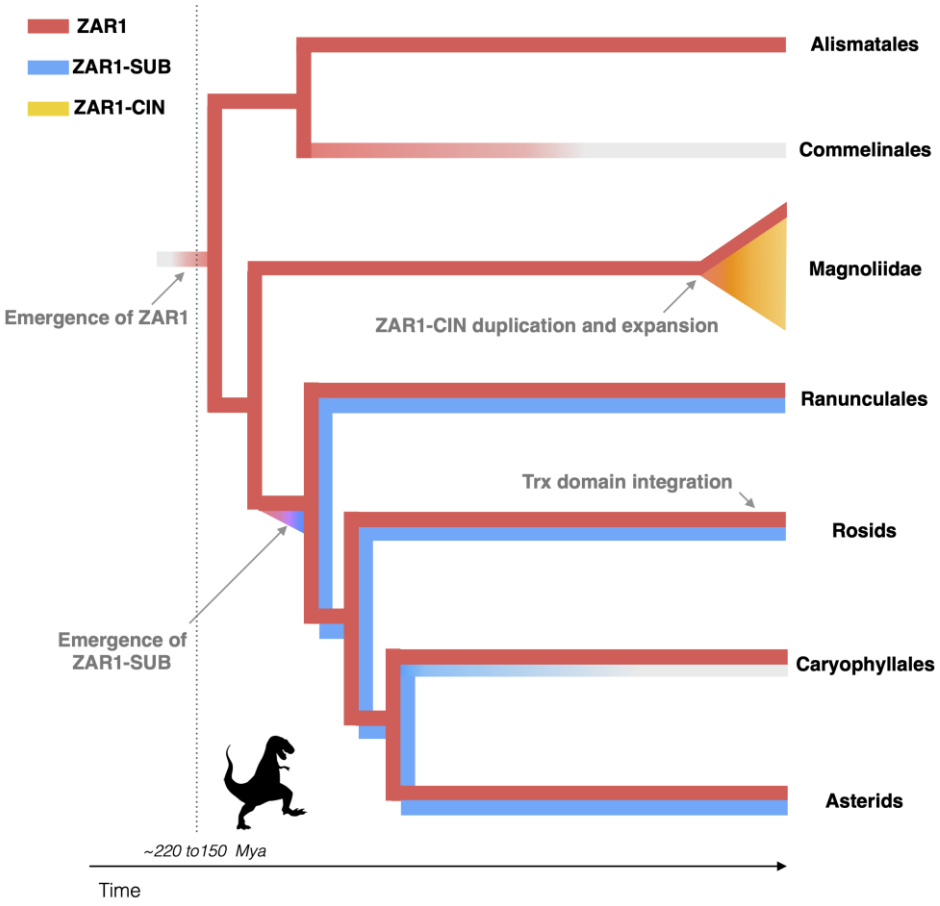


Figure 9. Evolution of ZAR1 and the paralogs in angiosperms. We propose that the ancestral ZAR1 gene has emerged ~220 to 150 million years ago (Mya) before monocot and eudicot lineages split. ZAR1 gene is widely conserved CC-NLR in angiosperms, but it is likely that ZAR1 has lost in a monocot lineage, Commelinales. A sister clade paralog ZAR1-SUB has emerged early in the eudicot lineages and may have lost in Caryophyllales. Another sister clade paralog ZAR1-CIN has duplicated from ZAR1 gene and expanded in the Magnoliidae *C. micranthum*. Trx domain integration to C terminus of ZAR1 has independently occurred in few rosid lineages.

Xanthomonas campestris (Wang et al., 2015). How did ZAR1 remain conserved throughout its evolutionary history while managing to detect a diversity of effectors? The answer to the riddle lies in the fact that ZAR1 effector recognition occurs via its partner RLCKs. HopZ-ETI-deficient 1 (ZED1) and ZED1-related kinases (ZRKs) of the RLCK XII-2 subfamily rest in complex with inactive ZAR1 proteins and bait effectors by binding them directly or by recruiting other effector-binding RLCKs, such as the family VII PBS1-like protein 2 (PBL2) (Lewis et al., 2013; Wang et al., 2015). These ZAR1-associated RLCKs are highly diversified in Arabidopsis, with 8 of the 13 RLCK XII-2 members occurring in the expanded ZRK gene cluster (Lewis et al., 2013). In this ZRK cluster, RKS1/ZRK1 is required for recognition of *X. campestris* effector AvrAC (Wang et al., 2015) and ZRK3 and ZRK5/ZED1 are required for recognition of *P. syringae* effectors HopF2a and HopZ1a, respectively (Lewis et al., 2013; Seto et al., 2017). Therefore, as in the model discussed by Schultink et al. (2019), RLCKs have evolved as pathogen 'sensors' whereas ZAR1 acts as a conserved signal executor to activate immune response. Future phylogenomic analyses of the RLCK subfamilies coupled with functional analyses with ZAR1 across angiosperms will help test and sharpen this model.

Our MEME and ConSurf analyses are consistent with the model of ZAR1/RLCK evolution described above. ZAR1 is not just exceptionally conserved across angiosperms but it has also preserved sequence patterns that are key to resistosome-mediated immunity (Figures 3 and 4). In particular, within the LRR domain, ZAR1 orthologs display highly conserved surfaces for RLCK binding (Figure 4). We conclude that ZAR1 has been guarding host kinases throughout its evolution ever since the Jurassic period. These findings strikingly contrast with observations recently made by Prigozhin and Krasileva (2020) on highly variable Arabidopsis NLRs (hvNLRs), which tend to have diverse LRR sequences. For instance, the CC-NLR RPP13 displays variable LRR surfaces across 62 Arabidopsis accessions, presumably because these regions are effector recognition interfaces that are caught in arms race coevolution with the oomycete pathogen *Hyaloperonospora arabidopsidis* (Prigozhin and Krasileva, 2020). The emerging view is that the mode of pathogen detection (direct vs indirect recognition) drives an NLR evolutionary trajectory by accelerating sequence diversification at the effector binding site or by maintaining the binding interface with the partner guardee/decoy proteins (Prigozhin and Krasileva, 2020).

ZAR1 orthologs display a patchy distribution across angiosperms (Figure 9, Supplementary table 1). Given the low number of non-eudicot species with ZAR1 it is challenging to develop a conclusive evolutionary model. Nonetheless, the most parsimonious explanation is that ZAR1 was lost in the monocot Commelinales lineage (Figure 9, Supplementary table 1). ZAR1 is also missing in some eudicot lineages, notably Fabales, Cucurbitales, Apiales and Asterales (Supplementary table 1). Cucurbitaceae (Cucurbitales) species are known to have reduced repertoires of NLR genes possibly due to low levels of gene duplications and frequent deletions (Lin et al., 2013). ZAR1 may have been lost in this and other plant lineages as part of an overall shrinkage of their NLRomes or as a consequence of selection against autoimmune phenotypes triggered by NLR mis-regulation (Karasov et al., 2017; Adachi et al., 2019a). In the future, it would be interesting to investigate the repertoires of RLCK subfamilies VII and XII in species that lack ZAR1 orthologs.

We unexpectedly discovered that some ZAR1 orthologs from cassava and cotton species carry a C-terminal thioredoxin-like domain (ZAR1-ID in Figure 5). What is the function of these integrated domains? The occurrence of unconventional domains in NLRs is relatively frequent and ranges from 5 to 10% of all NLRs. In several cases, integrated domains have emerged from pathogen effector targets and became decoys that mediate detection of the effectors (Kourelis and van der Hoorn, 2018). Whether or not the integrated Trx domain of ZAR1-ID functions to bait effectors will need to be investigated. Since ZAR1-ID proteins still carry intact kinase binding interfaces (Supplementary table 1—source data 2), they may have evolved dual or multiple recognition specificities via RLCKs and the Trx domain. In addition, all ZAR1-ID proteins have an intact N-terminal MADA motif (Figure 5—figure supplement 2), suggesting that they probably can execute the hypersensitive cell death through their N-terminal CC domains even though they carry a C-terminal domain extension (Adachi et al., 2019b). However, we noted multiple splice variants of the ZAR1-ID gene of cassava, some of which lack the Trx integration (Figure 5—figure supplement 1). It is possible that both ZAR1 and ZAR1-ID isoforms are produced, potentially functioning together as a pair of sensor and helper NLRs.

Our sequence analyses of ZAR1-ID indicate that the integrated Trx domain originates from the PLP3 phosphoducin gene, which is immediately downstream of ZAR1 in the Arabidopsis genome and adjacent to ZAR1 in several other eudicot species (Figure 5—figure supplement 3). Whether or not PLP3 plays a role in ZAR1 function and the degree to which close genetic linkage facilitated domain fusion between these two genes are provocative questions for future studies.

ZAR1 spawned two classes of paralogs through two independent duplication events. The ZAR1-SUB paralog clade emerged early in the eudicot lineage—most likely tens of millions of years after the emergence of ZAR1—and has diversified into at least 129 genes in 55 species (Figure 9). ZAR1-SUB proteins are distinctly more diverse in sequence than ZAR1 orthologs and generally lack key sequence features of ZAR1, like the MADA motif and the NBD-NBD oligomerisation interface (Figures 6, 7) (Adachi et al., 2019b; Wang et al. 2019b; Hu et al. 2020). This pattern is consistent with 'use-it-or-lose-it' evolutionary model, in which NLRs that specialize for pathogen detection lose some of the molecular features of their multifunctional ancestors (Adachi et al., 2019b). Therefore, we predict that many ZAR1-SUB proteins evolved into specialized sensor NLRs that require NLR helper mates for executing the hypersensitive response. It is possible that ZAR1-SUB helper mate is ZAR1 itself, and that these NLRs evolved into a phylogenetically linked network of sensors and helpers similar to the NRC network of asterid plants (Wu et al., 2017). However, 11 species have a ZAR1-SUB gene but lack a canonical ZAR1 (Supplementary table 3), indicating that these ZAR1-SUB NLRs may have evolved to depend on other classes of NLR helpers.

How would ZAR1-SUB sense pathogens? Given that the LRR domains of most ZAR1-SUB proteins markedly diverged from the RLCK binding interfaces of ZAR1, it is unlikely that ZAR1-SUB proteins bind RLCKs in a ZAR1-type manner (Figure 8). This leads us to draw the hypothesis that ZAR1-SUB proteins have diversified to recognize other ligands than RLCKs. In the future, functional investigations of ZAR1-SUB proteins could provide insights into how multifunctional NLRs, such as ZAR1, evolve into functionally specialized NLRs.

The ZAR1-CIN clade consists of 11 clustered paralogs that are unique to the magnoliid species stout camphor as revealed from the genome sequence of the Taiwanese small-flowered camphor tree (also known as *Cinnamomum kanehirae*, Chinese name niu zhang 牛樟) (Chaw et al., 2019). This cluster probably expanded from ZAR1, which is ~48 Mbp on the same genome sequence scaffold (Figure 6—figure supplement 2). The relatively rapid expansion pattern of ZAR1-CIN into a tandemly duplicated gene cluster is more in line with the classical model of NLR evolution compared to ZAR1 maintenance as a genetic singleton over tens of millions of years (Michelmore and Meyers, 1998). ZAR1-CIN proteins may have neofunctionalized after duplication, acquiring new recognition specificities as a consequence of coevolution with host partner proteins and/or pathogen effectors. Consistent with this view, ZAR1-CIN proteins display distinct surfaces at the ZAR1-RLCK binding interfaces and may bind to other ligands than RLCKs as we hypothesized above for ZAR1-SUB (Figure 8). ZAR1-CIN could be viewed as intraspecific highly variable NLRs (hvNLR) per the nomenclature of Prigozhin and Krasileva (2020).

Unlike ZAR1-SUB, ZAR1-CIN have retained the N-terminal MADA sequence (Figures 7 and 8). We propose that ZAR1-CIN are able to execute the hypersensitive cell death on their own

similar to ZAR1. However, ZAR1-CIN display divergent sequence patterns at NBD-NBD oligomerisation interfaces compared to ZAR1 (Figure 7C, Figure 7—figure supplement 2). Therefore, ZAR1-CIN may form resistosome-type complexes that are independent of ZAR1. One intriguing hypothesis is that ZAR1-CIN may associate with each other to form heterocomplexes of varying complexity and functionality operating as an NLR receptor network. In any case, the clear-cut evolutionary trajectory from ZAR1 to the ZAR1-CIN paralog cluster provides a robust evolutionary framework to study functional transitions and diversifications in this CC-NLR lineage.

In summary, our phylogenomics analyses raise a number of intriguing questions about ZAR1 evolution. The primary hypothesis we draw is that ZAR1 is an ancient CC-NLR that has been guarding RLCKs ever since the Jurassic Period. Throughout over 150 million years, ZAR1 has maintained its molecular features for sensing pathogens and activating hypersensitive cell death, but it also has retained an intriguingly conserved NB-ARC ring surface on the underside of the ZAR1 resistosome (Figure 4). We propose that this underside surface may play an important function in the resistosome, similar to other highly conserved regions such as α1 helix/MADA motif, NBD-NBD oligomerisation and RLCK binding interfaces. The equivalent region of the NB-ARC underside ring is apparently not exposed onto the underside surface of the TIR-NLR Roq1 resistosome structure (Martin et al., 2020). Therefore, ZAR1 conserved underside surface may be a specific feature of CC-NLR resistosomes. Further comparative analyses, combining molecular evolution and structural biology, of plant resistosomes and between resistosomes and the apoptosomes and inflammasome of animal NLR systems (Wang and Chai, 2020) will yield novel experimentally testable hypotheses for NLR research.

Materials and Methods

ZAR1 sequence retrieval

We performed BLAST (Altschul et al., 1990) using previously identified ZAR1 sequences as queries (Baudin et al. 2017; Schultink et al. 2019; Harant et al. 2020) to search ZAR1 like sequences in NCBI nr or nr/nt database (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and Phytozome12.1 (https://phytozome.jgi.doe.gov/pz/portal.html#!search?show=BLAST). In the BLAST search, we used cut-offs, percent identity ≥ 40% and query coverage ≥ 95%. The BLAST pipeline was circulated by using the obtained sequences as new queries to search ZAR1 like genes over the angiosperm species. We also performed the BLAST pipeline against a plant NLR dataset annotated by NLR-parser (Steuernagel et al., 2015) from 38 plant reference genome databases (Supplementary table 5).

Phylogenetic analyses

For the phylogenetic analysis, we aligned NLR amino acid sequences using MAFFT v.7 (Katoh and Standley, 2013) and manually deleted the gaps in the alignments in MEGA7 (Kumar et al., 2016). Full-length or NB-ARC domain sequences of the aligned NLR datasets were used for generating phylogenetic trees. The neighbour-joining tree was made using MEGA7 with JTT model and bootstrap values based on 100 iterations. All datasets used for phylogenetic analyses are in source data files.

Patristic distance analyses

To calculate the phylogenetic (patristic) distance, we used Python script based on DendroPy (Sukumaran and Mark, 2010). We calculated patristic distances from each CC-NLR to the other CC-NLRs on the phylogenetic tree (Figure 1—source data 3) and extracted the distance between CC-NLRs of Arabidopsis or *N. benthamiana* to the closest NLR from the other plant species. The script used for the patristic distance calculation is available from GitHub (https://github.com/slt666666/ Phylogenetic_distance_plot2).

Gene co-linearity analyses

To investigate genetic co-linearity at ZAR1 loci, we extracted the 3 genes upstream and downstream of ZAR1 using GFF files derived from reference genome databases (Supplementary table 5). To identify conserved gene blocks, we used gene annotation from NCBI Protein database and confirmed protein domain information based on InterProScan (Jones et al, 2014).

Sequence conservation analyses

Full-length NLR sequences of the each subfamily ZAR1, ZAR1-SUB or ZAR1-CIN were subjected to motif searches using the MEME (Multiple EM for Motif Elicitation) (Bailey and Elkan, 1994) with parameters 'zero or one occurrence per sequence, top twenty motifs', to detect consensus motifs conserved in ≥90% of the input sequences. The output data are summarized in Figure 3–supplementary table 1, Figure 7–supplementary table 1 and Figure 7–supplementary table 3.

To predict the MADA motif from ZAR1, ZAR1-SUB and ZAR1-CIN datasets, we used the MADA-HMM previously developed (Adachi et al., 2019b), with the hmmsearch program (hmmsearch –max -o <outputfile> <hmmfile> <seqdb>) implemented in HMMER v2.3.2 (Eddy, 1998). We termed sequences over the HMMER cut-off score of 10.0 as the MADA motif and sequences having the score 0-to-10.0 as the MADA-like motif.

To analyze sequence conservation and variation in ZAR1, ZAR1-SUB and ZAR1-CIN proteins, aligned full-length NLR sequences in MAFFT v.7 were used for ConSurf (Ashkenazy et al., 2016). Arabidopsis ZAR1 (NP_190664.1), a tomato ZAR1-SUB (XP_004243429.1) or a Stout camphor ZAR1-CIN (RWR85656.1) was used as a query for each analysis of ZAR1, ZAR1-SUB or ZAR1-CIN, respectively. The output datasets are in Figure 3—source data 1, Figure 7—source data 1 and Figure 7—source data 2.

Protein structure analyses

We used the cryo-EM structure of activated ZAR1 (Wang et al., 2019b) as template to generate a homology model of ZAR1-SUB and ZAR1-CIN. The amino acid sequence of a tomato ZAR1-SUB (XP_004243429.1) and a Stout camphor ZAR1-CIN (RWR85656.1) were submitted to Protein Homology Recognition Engine V2.0 (Phyre2) for modelling (Kelley et a., 2015). The coordinates of ZAR1 structure (6J5T) were retrieved from the Protein Data Bank and assigned

as modelling template by using Phyre2 Expert Mode. The resulting model of ZAR1-SUB and ZAR1-CIN, and the ZAR1 structures (6J5T) were illustrated with the ConSurf conservation scores in PyMol.

ACKNOWLEDGEMENTS

We are thankful to several colleagues for discussions and ideas. We thank Sebastian Schornack (Sainsbury Laboratory, University of Cambridge, Cambridge, UK) for valuable comments on this paper. This work was funded by the Gatsby Charitable Foundation, Biotechnology and Biological Sciences Research Council (BBSRC, UK), and European Research Council (ERC BLASTOFF projects). We thank the Prime Minister of the United Kingdom for announcing a stay-at-home order on 23th March 2020.

AUTHOR CONTRIBUTIONS

H.A. and S.K. mainly wrote the paper; H.A., A.M. and S.K. designed the research and supervised the work; H.A., T.S., J.K. and A.M. performed research.

DECLARATION OF INTERESTS

S.K. receives funding from industry on NLR biology.

References

Adachi H, Derevnina L, Kamoun S. (2019a) NLR singletons, pairs, and networks: evolution, assembly, and regulation of the intracellular immunoreceptor circuitry of plants. *Curr Opin Plant Biol* **50**:121-131. doi:10.1016/j.pbi.2019.04.007.

Adachi H, Contreras MP, Harant A, Wu CH, Derevnina L, Sakai T, Duggan C, Moratto E, Bozkurt TO, Maqbool A, Win J, Kamoun S. (2019b) An N-terminal motif in NLR immune receptors is functionally conserved across distantly related plant species. *eLife* **8:** e49956. doi: 10.7554/eLife.49956.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-10. doi: 10.1016/S0022-2836(05)80360-2.

Ashkenazy H, Abadi S, Martz E, et al. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**: W344-W350. doi:10.1093/nar/gkw408.

Baggs E, Dagdas G, Krasileva KV. (2017) NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. *Curr Opin Plant Biol* **38**: 59-67. doi:10.1016/j.pbi.2017.04.012.

Bailey TL, Elkan C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**:28-36.

Baudin M, Hassan JA, Schreiber KJ, Lewis JD. (2017) Analysis of the ZAR1 immune complex reveals determinants for immunity and molecular interactions. *Plant Physiol* **174**: 2038-2053. doi: 10.1104/pp.17.00441.

Baudin M, Schreiber KJ, Martin EC, Petrescu AJ, Lewis JD. (2019) Structure-function analysis of ZAR1 immune receptor reveals key molecular interactions for activity. *Plant J* **101**: 352-370. doi: 10.1111/tpj.14547.

Bayless AM, Nishimura MT. (2020) Enzymatic functions for Toll/Interleukin-1 receptor domain proteins in the plant immune system. *Front Genet* **11**:539. doi:10.3389/fgene.2020.00539.

Bentham AR, Zdrzalek R, De la Concepcion JC, Banfield MJ. (2018) Uncoiling CNLs: structure/Function approaches to understanding CC domain function in plant NLRs. *Plant and Cell Physiology* **59**:2398–2408. doi: https://doi.org/10.1093/pcp/pcy185.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. (2013) GenBank. *Nucleic Acids Res* **41**: D36-42. doi: 10.1093/nar/gks1195.

Boutrot F, Zipfel C. (2017) Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. *Annu Rev Phytopathol* **55**: 257-286. doi:10.1146/annurev-phyto-080614-120106.

Burdett H, Bentham AR, Williams SJ, et al. (2019) The plant "resistosome": structural insights into immune signaling. *Cell Host Microbe* **26**:193-201. doi:10.1016/j.chom.2019.07.020.

Cesari S, Bernoux M, Moncuquet P, Kroj T, Dodds PN. (2014) A novel conserved mechanism for plant NLR protein pairs: the "integrated decoy" hypothesis. *Front Plant Sci* **5**:606. doi:10.3389/fpls.2014.00606.

Chaw SM, Liu YC, Wu YW, et al. (2019) Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat Plants* **5**:63-73. doi:10.1038/s41477-018-0337-0.

Dangl JL, Horvath DM, Staskawicz BJ. (2013) Pivoting the plant immune system from dissection to deployment. *Science* **341**:746-751. doi:10.1126/science.1236011.

Delaux PM, Hetherington AJ, Coudert Y, et al. (2019) Reconstructing trait evolution in plant evo-devo studies. *Curr Biol* **29**:R1110-R1118. doi:10.1016/j.cub.2019.09.044.

Dodds PN, Rathjen JP. (2010) Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat Rev Genet* **11**:539-548. doi:10.1038/nrg2812.

Duxbury Z, Wang S, MacKenzie CI, et al. (2020) Induced proximity of a TIR signaling domain on a plant-mammalian NLR chimera activates defense in plants. *Proc Natl Acad Sci U S A* **117**: 18832-18839. doi:10.1073/pnas.2001185117.

Eddy SR. (1998) Profile hidden markov models. *Bioinformatics* **14**:755-763. DOI: https://doi.org/10.1093/bioinformatics/14.9.755.

Feehan JM, Castel B, Bentham AR, Jones JD. (2020) Plant NLRs get by with a little help from their friends *Curr Opin Plant Biol* **56**:99-108. doi:10.1016/j.pbi.2020.04.006.

Fu Q, Diez JB, Pole M, et al. (2018) An unexpected noncarpellate epigynous flower from the Jurassic of China. *Elife* **7**: e38827. doi:10.7554/eLife.38827.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178-1186. doi: 10.1093/nar/gkr944.

Harant A, Sakai T, Kamoun S, Adachi H. (2020) A vector system for fast-forward in vivo studies of the ZAR1 resistosome in the model plant *Nicotiana benthamiana*. *bioRxiv* doi: https://doi.org/10.1101/2020.05.15.097584.

Hu M, Qi J, Bi G, Zhou JM. 2020. Bacterial effectors induce oligomerization of immune receptor ZAR1 *in vivo*. *Mol Plant* pii: S1674-2052(20)30066-6. doi: 10.1016/j.molp.2020.03.004.

Jones JD, Dangl JL. (2006) The plant immune system. *Nature* **444**: 323-329. doi:10.1038/nature05286.

Jones JD, Vance RE, Dangl JL. (2016) Intracellular innate immune surveillance devices in plants and animals. *Science* **354**: aaf6395. doi:10.1126/science.aaf6395.

Jones P, Binns D, Chang H, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn A, Sangrador-Vegas A, Scheremetjew M, Yong S, Lopez R, Hunter S. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240. doi: 10.1093/bioinformatics/btu031.

Jubic LM, Saile S, Furzer OJ, El Kasmi F, Dangl JL. (2019) Help wanted: helper NLRs and plant immune responses. *Curr Opin Plant Biol* **50**: 82-94. doi:10.1016/j.pbi.2019.03.013.

Karasov TL, Chae E, Herman JJ, Bergelson J. (2017) Mechanisms to mitigate the trade-off between growth and defense. *Plant Cell* **29**: 666-680. doi:10.1105/tpc.16.00931.

Katoh K, Standley DM. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780. doi: 10.1093/molbev/mst010.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**: 845-858. doi: 10.1038/nprot.2015.053.

Kourelis J, van der Hoorn RAL. (2018) Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *The Plant Cell* **30**: 285-299. doi: https://doi.org/10.1105/tpc.17.00579.

Kourelis J, Kamoun S. (2020) RefPlantNLR: a comprehensive collection of experimentally validated plant NLRs. *bioRxiv* doi: https://doi.org/10.1101/2020.07.08.193961.

Kroj T, Chanclud E, Michel-Romiti C, Grand X, Morel JB. (2016) Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol* **210**: 618-626. doi:10.1111/nph.13869.

Kumar S, Stecher G, Tamura K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**: 1870-1874. doi: 10.1093/molbev/msw054.

Laflamme B, Dillon MM, Martel A, Almeida RND, Desveaux D, Guttman DS. (2020) The pan-genome effector-triggered immunity landscape of a host-pathogen interaction. *Science* **367**: 763-768. doi:10.1126/science.aax4079.

Lee H-Y, Mang H, Choi E-H, Seo Y-E, Kim M-S, Oh S, Kim S-B, Choi D. (2020) Genome-wide functional analysis of hot pepper immune receptors reveals an autonomous NLR cluster in seed plants. *bioRxiv* doi: https://doi.org/10.1101/2019.12.16.878959.

Lee RRQ, Chae E. (2020) Variation patterns of NLR clusters in *Arabidopsis thaliana* genomes. *Plant Communications* https://doi.org/10.1016/j.xplc.2020.100089.

Lewis JD, Lee AH, Hassan JA, et al. (2013) The *Arabidopsis* ZED1 pseudokinase is required for ZAR1-mediated immunity induced by the *Pseudomonas syringae* type III effector HopZ1a. *Proc Natl Acad Sci U S A*. **110**: 18722-18727. doi:10.1073/pnas.1315520110.

Liang X, Zhou JM. (2018) Receptor-like cytoplasmic kinases: central players in plant receptor kinase-mediated signaling. *Annu Rev Plant Biol* **69**: 267-299. doi:10.1146/annurev-arplant-042817-040540.

Lin X, Zhang Y, Kuang H, Chen J. (2013) Frequent loss of lineages and deficient duplications accounted for low copy number of disease resistance genes in Cucurbitaceae. *BMC Genomics* **14**: 335. doi:10.1186/1471-2164-14-335.

Martin R, Qi T, Zhang H, Liu F, King M, Toth C, Nogales E, Staskawicz BJ. (2020) Structure of the activated Roq1 resistosome directly recognizing the pathogen effector XopQ. *bioRxiv* doi: https://doi.org/10.1101/2020.08.13.246413.

Mermigka G, Amprazi M, Mentzelopoulou A, Amartolou A, Sarris PF. (2020) Plant and animal innate immunity complexes: fighting different enemies with similar weapons. *Trends Plant Sci* **25**: 80-91. doi:10.1016/j.tplants.2019.09.008.

Michelmore RW, Meyers BC. (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res* **8**: 1113-1130. doi:10.1101/gr.8.11.1113.

Nei M, Hughes AL. (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: Tsuji K, Aizawa M, Sasazuki T, editors. 11th Histocompatibility Workshop and Conference. Oxford Univ. Press; Oxford, UK

Prigozhin DM, Krasileva KV. (2020) Intraspecies diversity reveals a subset of highly variable plant immune receptors and predicts their binding sites. *bioRxiv* doi: https://doi.org/10.1101/2020.07.10.190785.

Sarris PF, Cevik V, Dagdas G, Jones JD, Krasileva KV. (2016) Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol* **14**: 8. doi:10.1186/s12915-016-0228-7.

Seong K, Seo E, Witek K, Li M, Staskawicz B. (2020) Evolution of NLR resistance genes with noncanonical N-terminal domains in wild tomato species. *New Phytol* **227**: 1530-1543. doi:10.1111/nph.16628.

Seto D, Koulena N, Lo T, Menna A, Guttman DS, Desveaux D. (2017) Expanded type III effector recognition by the ZAR1 NLR protein using ZED1-related kinases. *Nat Plants* **3**: 17027. doi:10.1038/nplants.2017.27.

Schultink A, Qi T, Bally J, Staskawicz B. (2019) Using forward genetics in *Nicotiana benthamiana* to uncover the immune signaling pathway mediating recognition of the *Xanthomonas perforans* Effector XopJ4. *New Phytol* **221**: 1001-1009. doi: 10.1111/nph.15411.

Smith SA, Brown JW. (2018) Constructing a broadly inclusive seed plant phylogeny. *Am J Bot* **105**: 302-314. doi:10.1002/ajb2.1019.

Shao ZQ, Xue JY, Wu P, et al. (2016) Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiol* **170**: 2095-2109. doi:10.1104/pp.15.01487.

Stam R, Silva-Arias GA, Tellier A. (2019) Subsets of NLR genes show differential signatures of adaptation during colonization of new habitats. *New Phytol* **224**: 367-379. doi:10.1111/nph.16017.

Steuernagel B, Jupe F, Witek K, Jones JD, Wulff BB. (2015) NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics* **31**: 1665-1667. doi: 10.1093/bioinformatics/btv005.

Sukumaran J, Holder MT. (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**: 1569-1571. doi: 10.1093/bioinformatics/btq228.

Tamborski J, Krasileva KV. (2020) Evolution of plant NLRs: from natural history to precise modifications. *Annu Rev Plant Biol* **71**: 355-378. doi:10.1146/annurev-arplant-081519-035901.

Uehling J, Deveau A, Paoletti M. (2017) Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems. *PLoS Pathog* **13**: e1006578. doi:10.1371/journal.ppat.1006578.

Van de Weyer AL, Monteiro F, Furzer OJ, et al. (2019) A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell* **178**: 1260-1272.e14. doi:10.1016/j.cell.2019.07.038.

Wang G, Roux B, Feng F, et al. (2015) The decoy substrate of a pathogen effector and a pseudokinase specify pathogen-induced modified-self recognition and immunity in plants. *Cell Host Microbe* **18**: 285-295. doi:10.1016/j.chom.2015.08.004.

Wang J, Wang J, Hu M, W Shan, Qi J, Wang G, Han Z, Qi Y, Gao N, Wang H-W, Zhou J-M, Chai J. (2019a) Ligand-triggered allosteric ADP release primes a plant NLR complex. *Science* **364**: eaav5868. doi: 10.1126/science.aav5868.

Wang J, Hu M, Wang J, Qi J, Han Z, Wang G, Qi Y, Wang H-W, Zhou J-M, Chai J. (2019b) Reconstitution and structure of a plant NLR resistosome conferring immunity. *Science* **364**: eaav5870. doi: 10.1126/science.aav5870.

Wang J, Chai J. (2020) Structural Insights into the Plant Immune Receptors PRRs and NLRs. *Plant Physiol* **182**: 1566-1581. doi:10.1104/pp.19.01252.

Wu CH, Krasileva KV, Banfield MJ, Terauchi R, Kamoun S. (2015) The "sensor domains" of plant NLR proteins: more than decoys? *Front Plant Sci* **6**: 134. doi:10.3389/fpls.2015.00134.

Wu CH, Abd-El-Haliem A, Bozkurt TO, et al. (2017) NLR network mediates immunity to diverse plant pathogens. *Proc Natl Acad Sci U S A* **114**: 8113-8118. doi:10.1073/pnas.1702041114.

Wu CH, Derevnina L, Kamoun S. (2018) Receptor networks underpin plant immunity. *Science* **360**: 1300-1301. doi:10.1126/science.aat2623.

Xiong Y, Han Z, Chai J. (2020) Resistosome and inflammasome: platforms mediating innate immunity. *Curr Opin Plant Biol* **56**: 47-55. doi:10.1016/j.pbi.2020.03.010.

Zhou JM, Zhang Y. (2020) Plant immunity: danger perception and signaling. *Cell* **181**: 978-989. doi:10.1016/j.cell.2020.04.028.

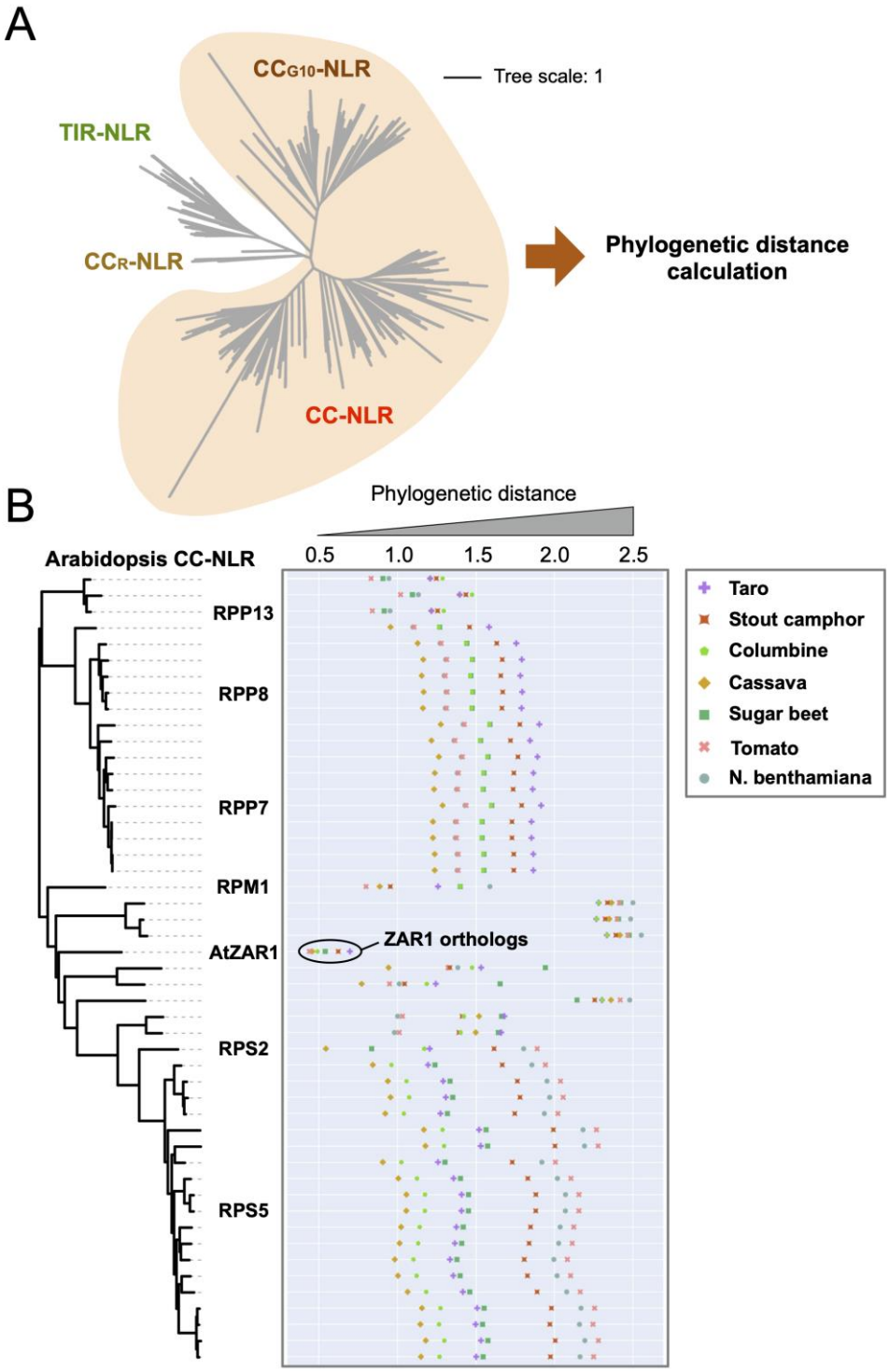


Figure 1—figure supplement 1. Arabidopsis ZAR1 is the most conserved CC-NLR across angiosperms. (**A**) Phylogenetic tree of NLR proteins from 8 plant species. The phylogenetic tree was generated in MEGA7 by the neighbour-joining method using NB-ARC domain sequences of 1475 NLRs identified from taro, stout camphor, columbine, Arabidopsis, cassava, sugar beet, tomato and *N. benthamiana*. The scale bars indicate the evolutionary distance in amino acid substitution per site. We used CC-NLR and CC_{G10}-NLR superclades for calculating phylogenetic distances. (**B**) The phylogenetic (patristic) distance of two CC-NLR nodes between Arabidopsis and other plant species were calculated from the NB-ARC phylogenetic tree in A. The closest patristic distances are plotted with different colours based on plant species. Representative Arabidopsis NLRs are highlighted. The closest patristic distances of two CC-NLR nodes between *N. benthamiana* and other plant species can be found in Figure 1—figure supplement 2.

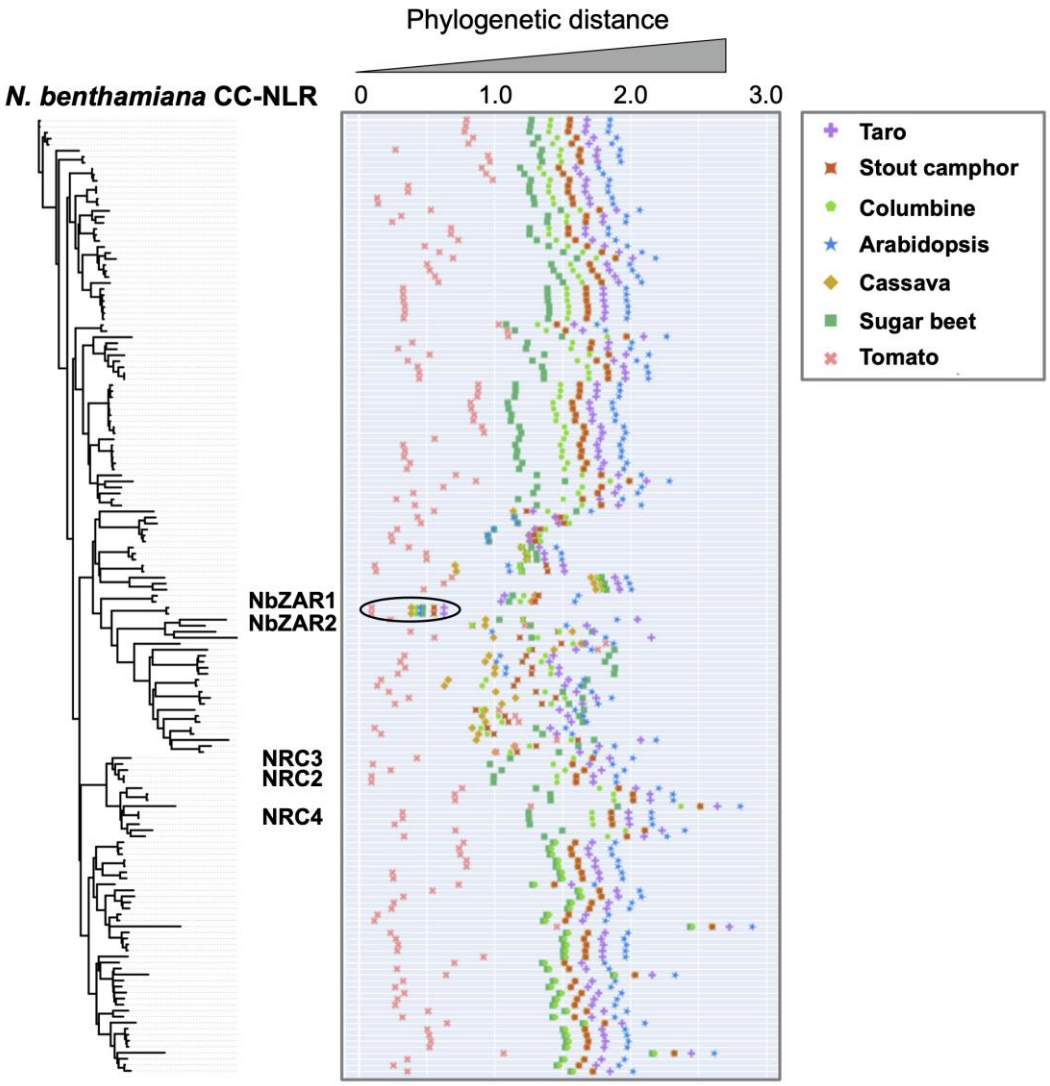


Figure 1—figure supplement 2. NbZAR1 is highly conserved across angiosperms. The phylogenetic (patristic) distance of two CC-NLR nodes between *N. benthamiana* and the closest NLR from the other plant species were calculated from the NB-ARC phylogenetic tree in Figure 1—figure supplement 1. The closest patristic distances are plotted with different colours based on plant species.

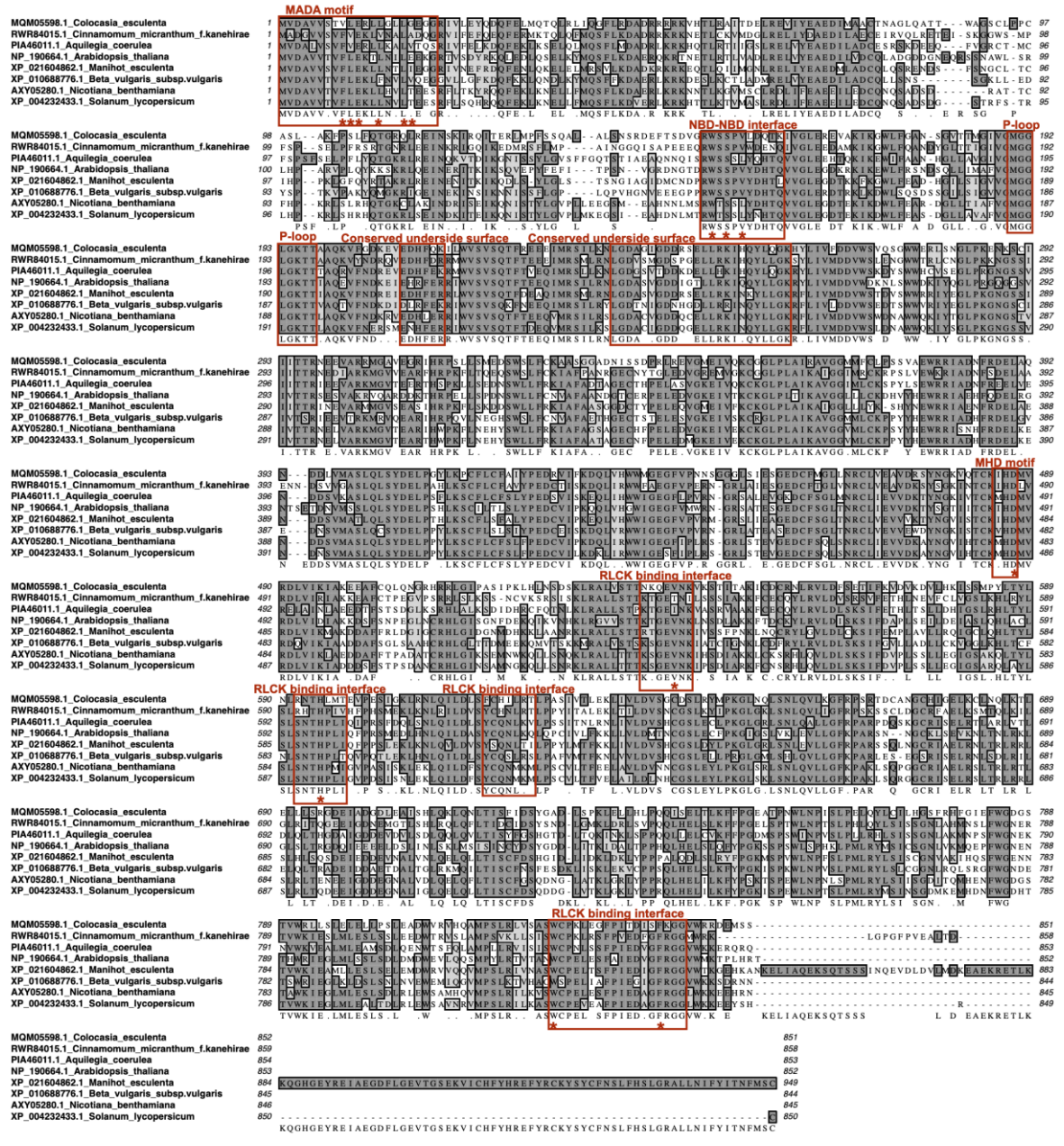


Figure 2—figure supplement 1. Sequence alignment of full-length ZAR1 ortholog proteins across angiosperms. Amino acid sequences of ZAR1 orthologs were aligned by MAFFT version 7 program. Conserved motif sequences highlighted in this study are marked with red boxes. Red asterisks indicate substitution sites for introducing gain or loss of ZAR1 protein function.

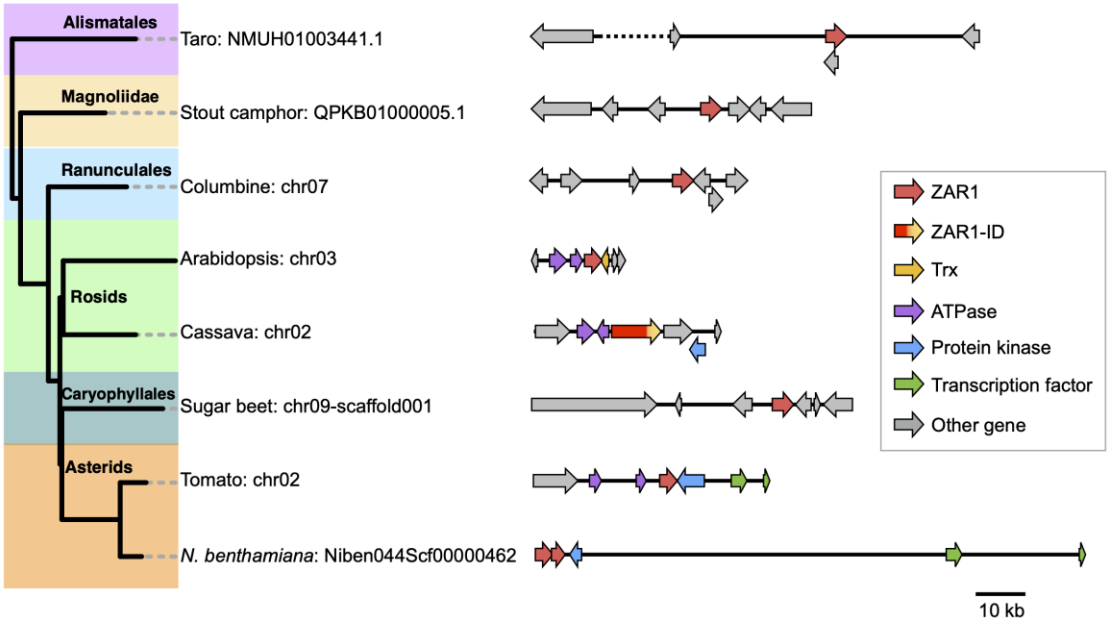


Figure 2—figure supplement 2. Schematic representation of the intragenomic relationship at ZAR1 loci across angiosperm genomes. We selected representative 8 plant species genome assemblies based on the phylogenetic tree in Figure 2 and used them for the synteny-based analysis of the ZAR1 loci. We highlight genes showing intragenomic linkages with different colours based on the gene annotations. Genes genetically linked to ZAR1 in eudicots are listed in Figure 2—figure supplement 2—Supplementary table 2.

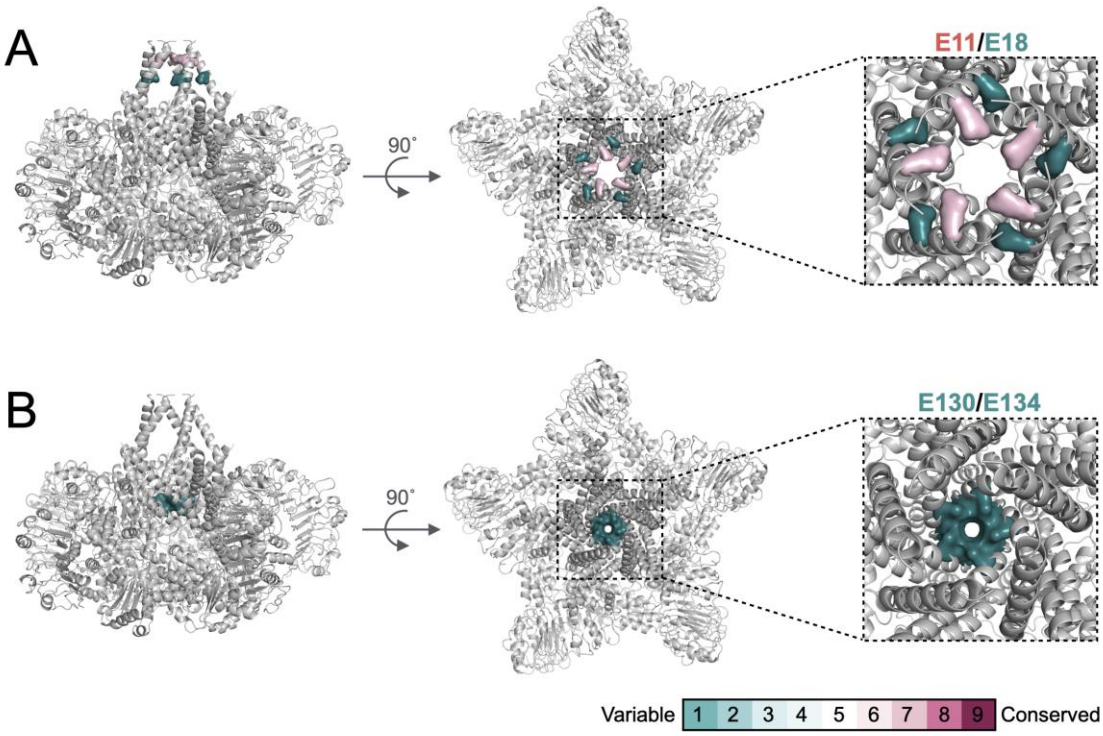


Figure 4—figure supplement 1. E18, E130 and E134 on glutamate rings inside of the Arabidopsis ZAR1 resistosome are variable across the orthologs. The ConSurf conservation scores at E11 and E18 (**A**) or at E130 and E134 (**B**) are illustrated in cartoon representation of the Arabidopsis ZAR1 resistosome structure.

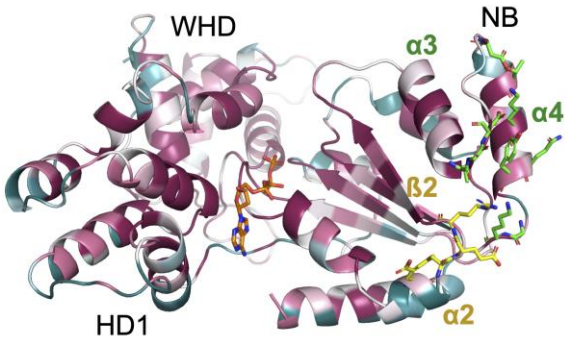
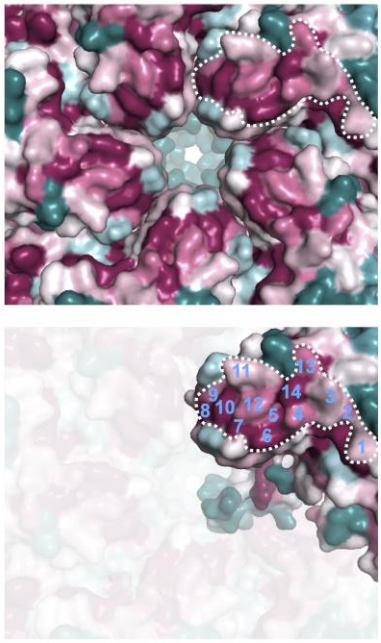


Figure 4—figure supplement 2. α2 helix-loop-ß2 sheet and α3 helix-loop-α4 helix in NBD locate on the underside surface of the activated ZAR1. NB-ARC domain of activated ZAR1 monomer in the resistosome is described as schematic representation. Residues that mainly locate on the conserved underside surface are shown as stick representation with different colour codes: yellow and green indicate α2 helix-loop-ß2 sheet and α3 helix-loop-α4 helix, respectively. A dATP molecule is present in nucleotide binding pocket and shown in stick representation with orange colour code.



Number	Amino acid	ConSurf score	Location in NBD
1	Glu-209	7	α2
2	Phe-212	8	loop
3	Glu-213	7	loop
4	Arg-214	8	loop
5	Leu-237	9	α3
6	Gly-238	9	loop
7	Asp-239	9	loop
8	Asp-244	9	α4
9	Thr-248	8	α4
10	Lys-252	9	α4
11	Gln-255	7	α4
12	Tyr-256	8	α4
13	Gly-259	8	loop
14	Lys-260	9	loop

Figure 4—figure supplement 3. Amino acid residues on the conserved underside surface of the ZAR1 resistosome. Conserved underside surface of the ZAR1 resistosome with five (top left) or single (bottom left) ZAR1 molecule(s). Both images are zoomed in from the underside view in Figure 4C. Conserved amino acids are labelled as numbers and listed in the right table.

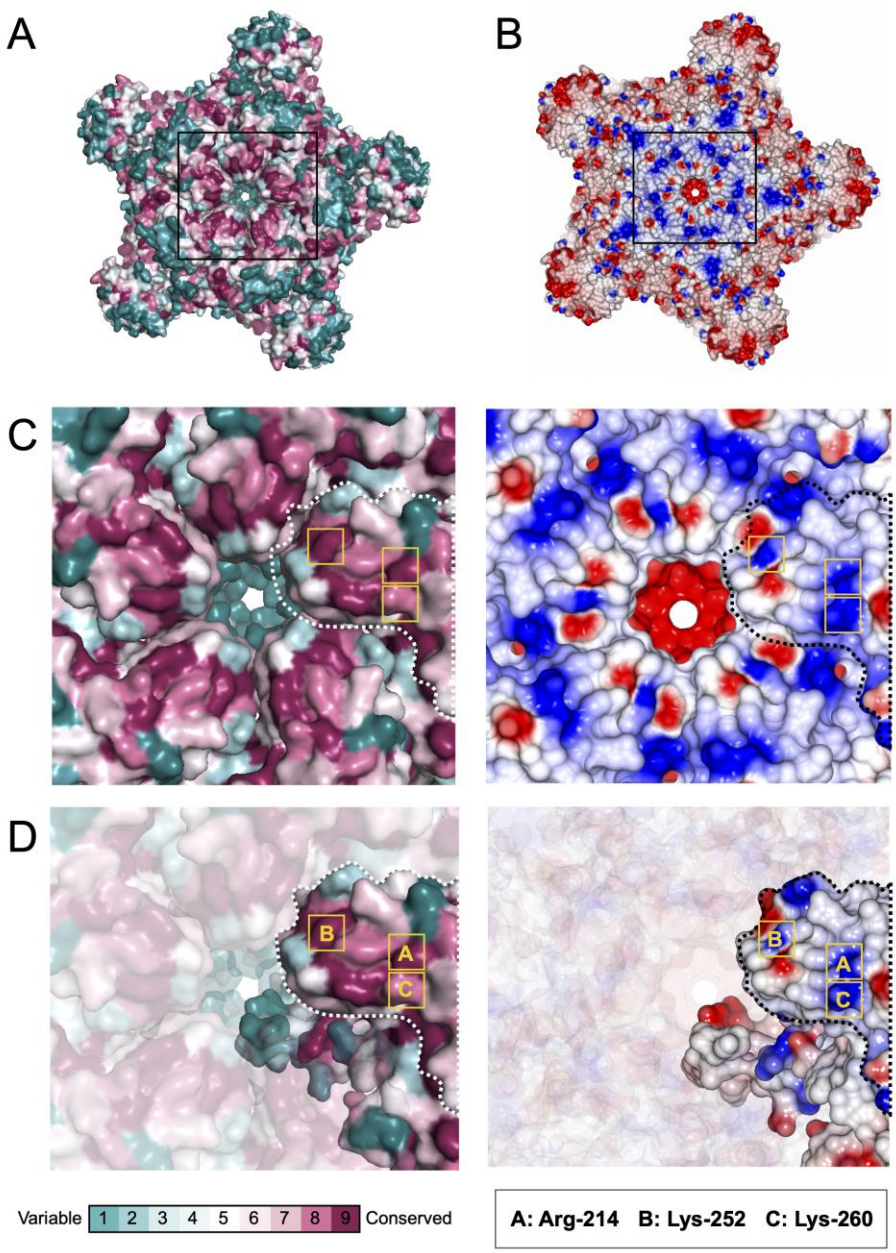


Figure 4—figure supplement 4. Three conserved residues form a positive electrostatic potential ring on underside surface of the ZAR1 resistosome. (**A, B**) Underside view of the ZAR1 resistosome with the ConSurf conservation score (**A**) or electrostatic potential (**B**). The colour gradient from red to blue represents negative to positive electrostatic potentials. Black boxes indicate regions zoomed in panels C and D. (**C, D**) Conserved underside surface of the ZAR1 resistosome with five (**C**) or single (**D**) ZAR1 molecule(s). The underside surface is illustrated with the ConSurf conservation score (Left) and electrostatic potential (right). Regions marked by white or black dot lines are exposed to the underside surface from single ZAR1 protein. Yellow boxes indicate positive charged residues that are conserved on the underside surface across ZAR1 orthologs.

Figure 5—figure supplement 1. Cassava ZAR1 and ZAR1-ID are transcribed from a single locus on the genome. The gene locus of cassava ZAR1 (XP_021604863.1, XP_021604865.1, XP_021604866.1, XP_021604867.1 and XP_021604868.1) and ZAR1-ID (XP_021604862.1 and XP_021604864.1) is shown with RNA-seq exon coverage and is extracted from NCBI database (database ID: LOC110609538).

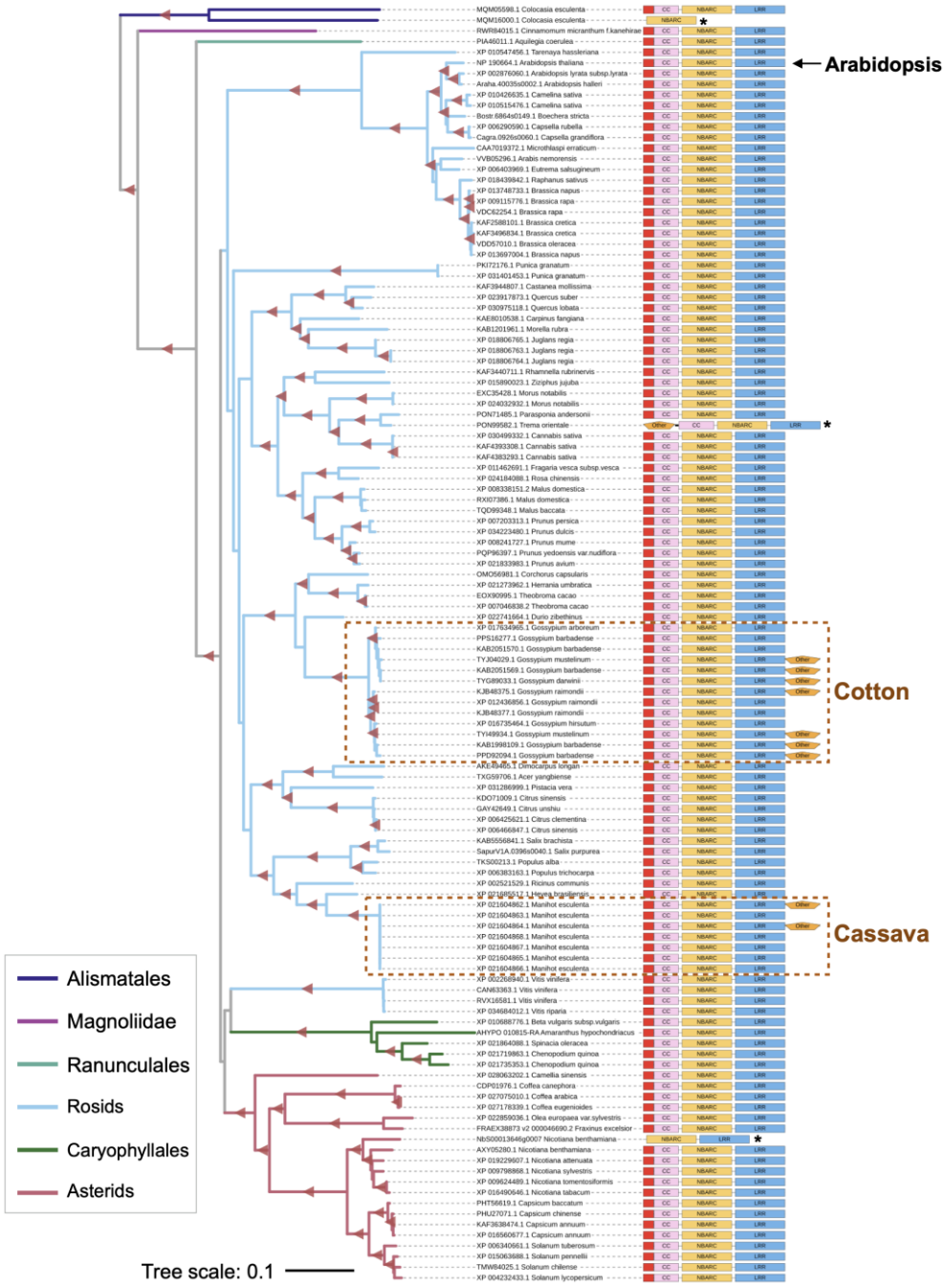


Figure 5—figure supplement 2. Trx domain integration occurred in two independent rosid ZAR1 subclades. The phylogenetic tree shown in Figure 2 was used to describe NLR domain architectures. Domain schemes are aligned to right side of the leaf labels: MADA is red, CC is pink, NB-ARC is yellow, LRR is blue and other domain is orange. Black asterisks on domain schemes describe truncated NLRs or potentially mis-annotated NLR. Each branch is marked with different colours based on the plant taxonomy. Red triangles indicate bootstrap support > 0.7. The scale bar indicates the evolutionary distance in amino acid substitution per site.

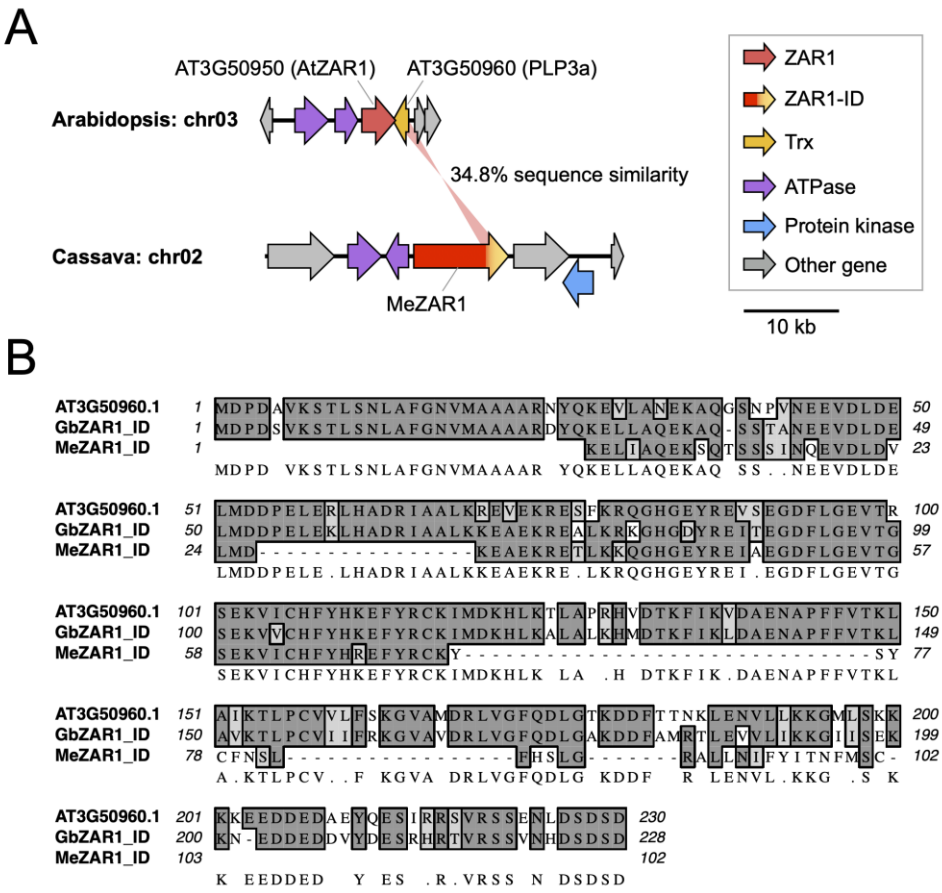


Figure 5—figure supplement 3. Integrated Trx domains show high sequence similarity to ZAR1-linked PLP3a gene in Arabidopsis. (**A**) Schematic representation of the intragenomic relationship at ZAR1 loci between Arabidopsis and cassava. We highlight sequence similarity of integrated Trx domain in Cassava ZAR1 (MeZAR1) to PLP3a gene genetically linked to Arabidopsis ZAR1 (AtZAR1). Details are explained in Figure 2—figure supplement 2. (**B**) Amino acid sequences of Arabidopsis PLP3a gene (AT3G50960) and integrated domains of an MeZAR1 (XP_021604862.1) and a cotton ZAR1 (GbZAR1; KAB1998109.1).

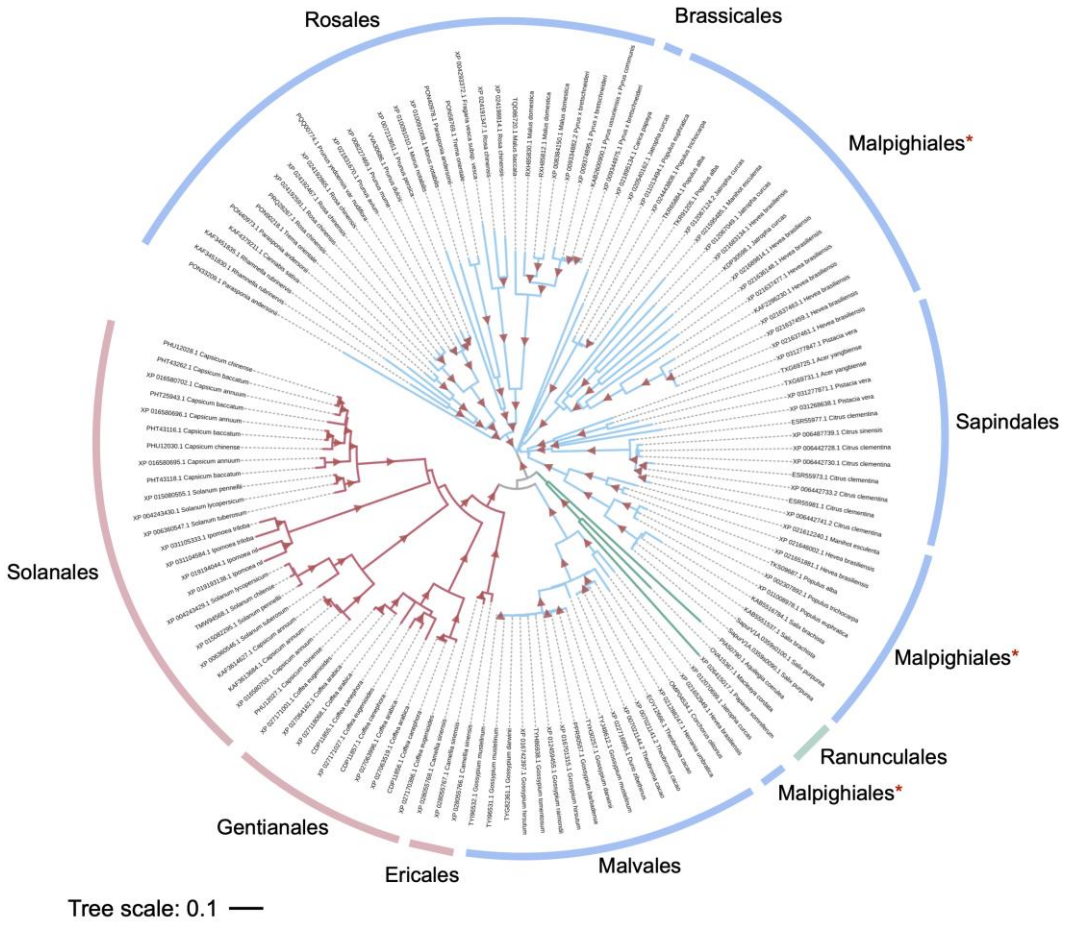


Figure 6—figure supplement 1. ZAR1-SUB gene is distributed across eudicots. The phylogenetic tree was generated in MEGA7 by the neighbour-joining method using full length amino acid sequences of 129 ZAR1-SUB orthologs identified in Figure 1. Each branch is marked with different colours based on the plant taxonomy. Red triangles indicate bootstrap support > 0.7. The scale bar indicates the evolutionary distance in amino acid substitution per site. Red asterisks on plant order term describe that NLRs from Malpighiales are distributed in three independent clades.

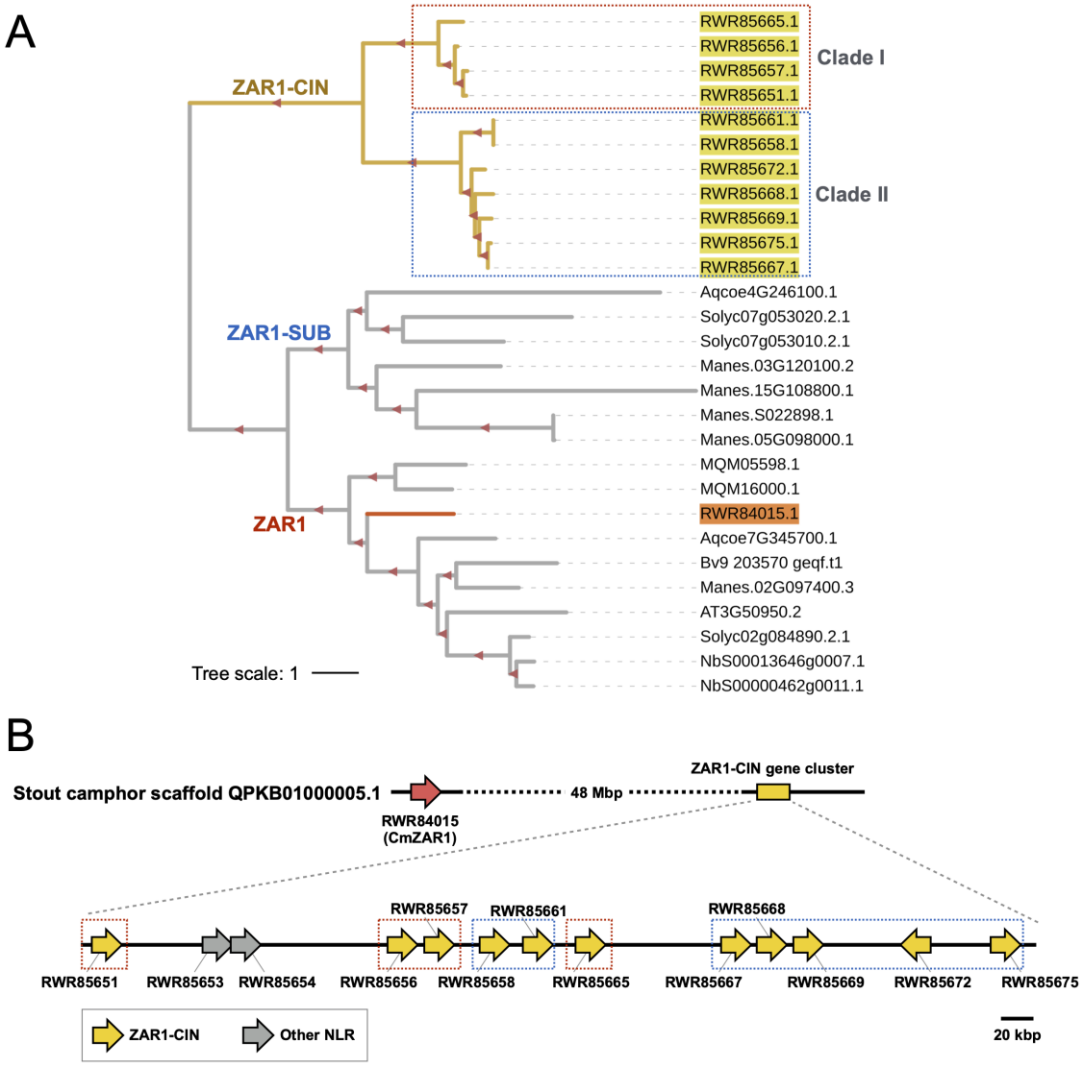


Figure 6—figure supplement 2. ZAR1-CIN gene cluster occurs in the *Cinnamomum micranthum* genome. (**A**) The subclades including ZAR1, ZAR1-SUB and ZAR1-CIN were zoomed in from the phylogenetic tree constructed in Figure 1—figure supplement 1. Red triangles indicate bootstrap support > 0.7. The scale bar indicates the evolutionary distance in amino acid substitution per site. Well supported subclades (I and II) in ZAR1-CIN are described with red or blue dot box. The gene IDs: taro (MQM-), stout camphor (RWR-), columbine (Aqcoe-), Arabidopsis (AT-), cassava (Manes-), sugar beet (Bv-), tomato (Solyc-) and *N. benthamiana* (NbS-). (**B**) Schematic representation of the ZAR1-CIN gene cluster on a *C. micranthum* (Stout camphor) scaffold. Stout camphor ZAR1 (CmZAR1) and ZAR1-CIN genes are highlighted in orange and yellow, respectively.

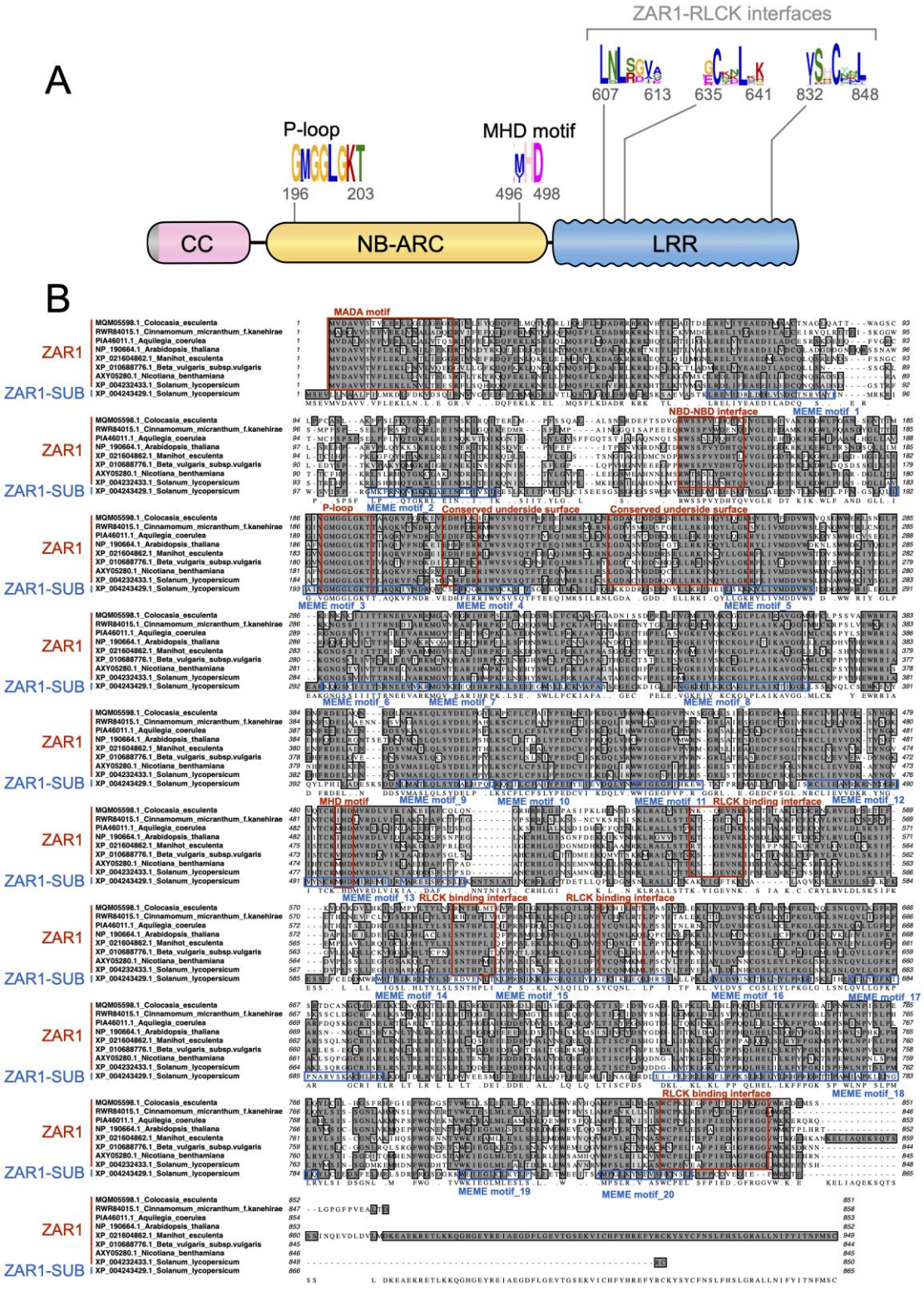


Figure 7—figure supplement 1. Sequence alignment of full-length ZAR1 and ZAR1-SUB proteins. (**A**) Schematic representation of the ZAR1-SUB protein highlighting the position of the representative conserved sequence patterns across ZAR1-SUB. (**B**) Amino acid sequences of ZAR1 orthologs and a representative ZAR1-SUB (XP_004243429.1 from tomato) were aligned by MAFFT version 7 program. ZAR1 motif sequences highlighted in this study are marked with red boxes. Positions of MEME motifs identified from ZAR1-SUB are marked in blue boxes. Raw MEME motifs are listed in Figure 7–supplementary tables 1 and 2.

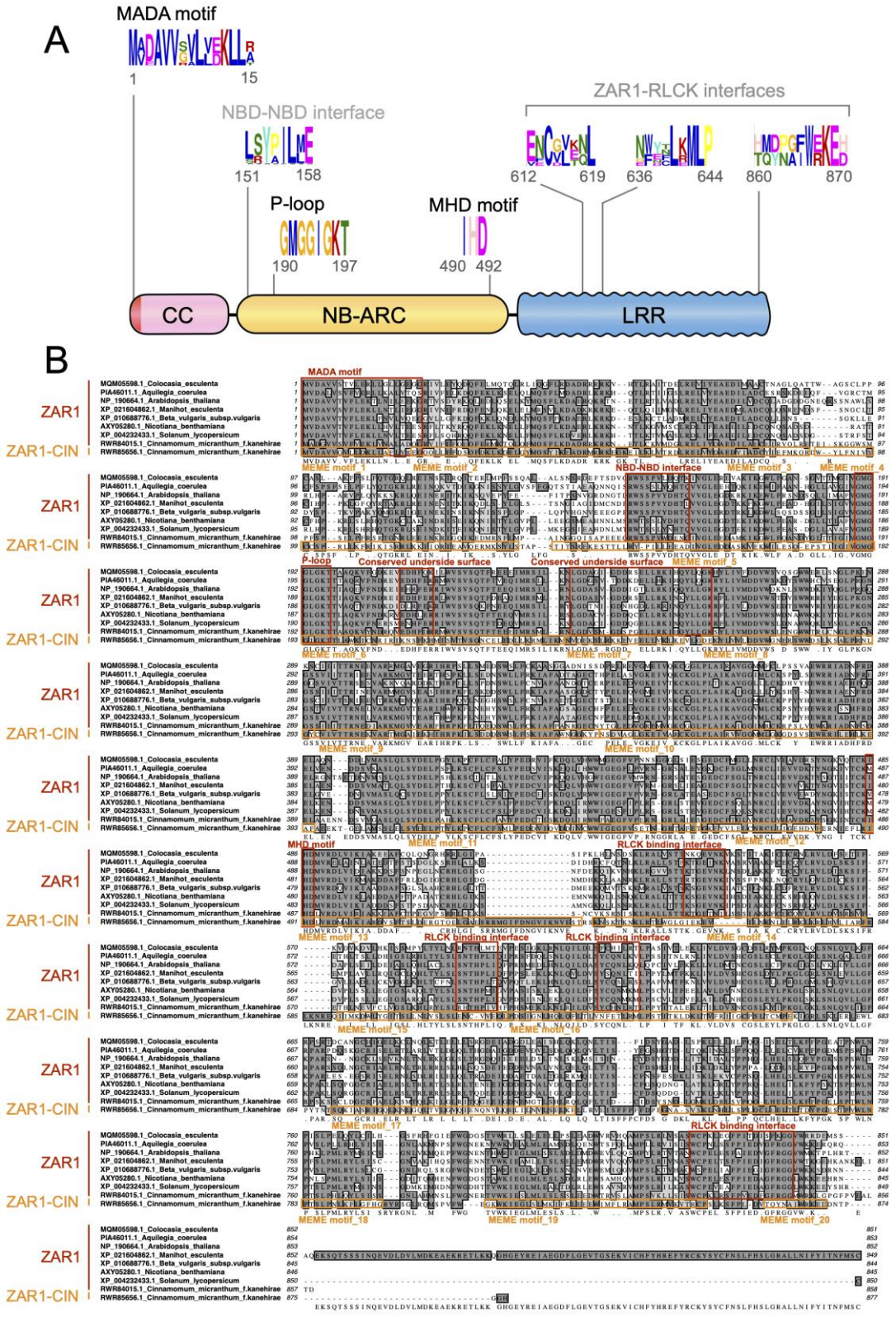


Figure 7—figure supplement 2. Sequence alignment of full-length ZAR1 and ZAR1-CIN proteins. (**A**) Schematic representation of the ZAR1-CIN protein highlighting the position of the represented conserved sequence patterns across ZAR1-SUB. (**B**) Amino acid sequences of ZAR1 orthologs and a representative ZAR1-CIN (RWR85656.1) were aligned by MAFFT version 7 program. ZAR1 motif sequences highlighted in this study are marked with red boxes. Positions of MEME motifs identified from ZAR1-CIN are marked in orange boxes. Raw MEME motifs are listed in Figure 7–supplementary tables 3 and 4.

Supplementary files

Supplementary table 1. List of ZAR1 in angiosperms.

Supplementary table 2. List of ZAR1-SUB.

Supplementary table 3. List of plant species with the number of ZAR1, ZAR1-SUB and ZAR1-CIN genes.

Supplementary table 4. List of ZAR1-CIN.

Supplementary table 5. Reference genome databases used for NLR annotation with NLR-parser.

Figure 2—figure supplement 2—Supplementary table 1. List of the closest NLR genes to ZAR1 locus.

Figure 2—figure supplement 2—Supplementary table 2. List of genes genetically linked to ZAR1 in eudicots.

Figure 3—Supplementary table 1. List of MEME motifs predicted from ZAR1 in angiosperms.

Figure 7—figure supplement 1. List of MEME motifs predicted from ZAR1-SUB.

Figure 7—figure supplement 2. Comparison of MEME motifs between ZAR1-SUB and ZAR1. Black dot boxes indicate corresponding regions between ZAR1-SUB and ZAR1 based on MAFFT v7 alignment in Figure 7—figure supplement 1. Red boxes indicate motifs which are highlighted in Figures 3 and 4.

Figure 7—figure supplement 3. List of MEME motifs predicted from ZAR1-CIN.

Figure 7—figure supplement 4. Comparison of MEME motifs between ZAR1-CIN and ZAR1. Black dot boxes indicate corresponding regions between ZAR1-CIN and ZAR1 based on MAFFT v7 alignment in Figure 7—figure supplement 2. Red boxes indicate motifs which are highlighted in Figures 3 and 4.

Source data files

Figure 1—source data 1. Amino acid sequences of full-length NLRs used for phylogenetic analysis in Figure 1B. This file contains 1268 NLR amino acid sequences with the IDs, taro (MQM-), stout camphor (RWR-), columbine (Aqcoe-), Arabidopsis (AT-), sugar beet (Bv-) and tomato (Solyc-).

Figure 1—source data 2. Amino acid sequences for NLR phylogenetic tree in Figure 1B. This file contains NB-ARC domain sequences used for phylogenetic analysis.

Figure 1—source data 3. NLR phylogenetic tree file in Figure 1B. The phylogenetic tree was saved in newick file format.

Figure 1—figure supplement 1—source data 1. Amino acid sequences of full-length NLRs used for phylogenetic analysis in Figure 1—figure supplement 1. This file contains 1475 NLR amino acid sequences with the IDs, taro (MQM-), stout camphor (RWR-), columbine (Aqcoe-), Arabidopsis (AT-), cassava (Manes-), sugar beet (Bv-), tomato (Solyc-) and *N. benthamiana* (NbS-).

Figure 1—figure supplement 1—source data 2. Amino acid sequences for NLR phylogenetic tree in Figure 1—figure supplement 1. This file contains NB-ARC domain sequences used for phylogenetic analysis.

Figure 1—figure supplement 1—source data 3. NLR phylogenetic tree file in Figure 1—figure supplement 1. The phylogenetic tree was saved in newick file format.

Figure 2—source data 1. NLR phylogenetic tree file in Figure 2. The phylogenetic tree was saved in newick file format.

Figure 3—source data 1. The ConSurf conservation score among ZAR1 proteins. The table contains conservation score on each position of amino acid sequences in Arabidopsis ZAR1.

Figure 6—source data 1. NLR phylogenetic tree file in Figure 6. The phylogenetic tree was saved in newick file format.

Figure 7—source data 1. The ConSurf conservation score among ZAR1-SUB proteins. The table contains conservation score on each position of amino acid sequences in a tomato ZAR1-SUB, XP_004243429.1.

Figure 7—source data 2. The ConSurf conservation score among ZAR1-CIN proteins. The table contains conservation score on each position of amino acid sequences in a cinnamomum ZAR1-CIN, RWR85656.1.

Supplementary table 1—source data 1. Amino acid sequences of 120 ZAR1 in angiosperms. This file contains 120 ZAR1 amino acid sequences identified from computational pipeline in Figure 1A.

Supplementary table 1—source data 2. Amino acid alignment file of 120 ZAR1 in angiosperms. Full-length amino acid sequences of ZAR1 orthologs were aligned by MAFFT version 7.

Supplementary table 1—source data 3. List of plant species carrying ZAR1 as single-copy gene.

Supplementary table 1—source data 4. List of plant species carrying 2 or more ZAR1 genes.

Supplementary table 2—source data 1. Amino acid sequences of 129 ZAR1-SUB. This file contains 129 ZAR1-SUB amino acid sequences identified from computational pipeline in Figure 1A.

Supplementary table 2—source data 2. Amino acid alignment file of 129 ZAR1-SUB. Full-length amino acid sequences of ZAR1-SUB were aligned by MAFFT version 7.

Supplementary table 3—source data 1. Amino acid sequences of 11 ZAR1-CIN. This file contains 11 ZAR1-CIN amino acid sequences identified from computational pipeline in Figure 1A.

Supplementary table 3—source data 2. Amino acid alignment file of 11 ZAR1-CIN. Full-length amino acid sequences of ZAR1-CIN were aligned by MAFFT version 7.