Pathogen-driven coevolution across CBP60 plant immune regulator subfamilies confers resilience on the regulator module

Qi Zheng[1,2,‡], Kristina Majsec[1], and Fumiaki Katagiri[1,*]

[1]Department of Plant and Microbial Biology, Microbial and Plant Genomics Institute, University of Minnesota, St Paul, MN 55108, USA. [2]Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Maize Research Institute, Sichuan Agricultural University, Chengdu, Sichuan 611130, China.

*Corresponding author, katagiri@umn.edu

[‡]Current address: School of Life Sciences, Shandong University, Qingdao, Shandong 266237, China

ABSTRACT

Among eight Arabidopsis CaM-Binding Protein (CBP) 60 family members, AtCBP60g and AtSARD1 are partially functionally redundant, major positive immune regulators while AtCBP60a is a negative immune regulator. Phylogenetic analysis of CBP60 protein sequences of 247 diverse land plant species indicated that the immune regulator CBP60a, CBP60g, and SARD1 subfamilies diversified at divergence of Angiosperms and have been evolving very fast, suggesting strong selection pressure from pathogen effectors. We developed the Protein Evolution Analysis in Euclidean Space (PEAES) approach to investigate effects of this potential selection. We detected significant coevolutionary interactions across the immune regulator subfamilies specific to different Core Eudicot lineages, which are consistent with hypothetical coevolutionary mechanisms that protect the positive immune regulator function from being targeted by pathogen effectors. Thus, fast coevolution across the subfamilies with overlapping or opposing functions appears crucial to maintain resilience of the CBP60 immune regulator module against fast-evolving pathogen effectors.

INTRODUCTION

Pathogen effectors, typically proteinaceous, are delivered from pathogens into plant cells and manipulate plant functions to improve pathogen environments *in planta*, including compromising plant immune signaling [1]. The plant immune signaling network characterized in Arabidopsis has a remarkable level of resilience as different parts of the network have compensating functions that buffer the disabling impacts of mutations or pathogen effectors [2–4]. We are interested in how such network resilience evolved in biological systems. To gain insights into this topic, we have been studying evolution of important components of the immune

signaling network. Members of the CaM-Binding Protein (CBP) 60 family are such network components [5–8].

Arabidopsis has eight CBP60 family members (AtCBP60a-g and AtSARD1) [6,8,9]. The CBP60 protein family was defined by the domain conserved among known CBP60 proteins (Pfam: “calmodulin_bind”, PF07887). Despite the Pfam domain name, the actual CaM-binding domains are outside the conserved domain (Fig 1) [5,6,9]. Thus, we call the conserved domain the CBP60-conserved domain. AtCBP60g and AtSARD1 are major positive regulators of immunity controlling activation of many immune responses, including synthesis of the important immune hormone salicylic acid (SA) [5,6,8]. They are transcriptionally induced during immune responses. Their functions as positive immune regulators are partially redundant since an *atcbp60g atsard1* double mutant has a more severe immune deficiency than either of the single mutants [6,8]. In contrast, AtCBP60a is a negative regulator of SA signaling and immunity [7]. In an *atcbp60a* mutant the basal level of SA is higher [7] and decrease of the post-induction level of AtCBP60g mRNA is slowed [10]. The functions of the other AtCBP60 members (AtCBP60b-f) are unknown as mutants lacking any one of them did not show substantial changes in immunity or other obvious phenotypes [7]. AtCBP60g and AtSARD1 are DNA-binding proteins [8,11]. Since the DNA-binding activity resides within the CBP60-conserved domain, it is likely that all CBP60 proteins are DNA-binding proteins. Phylogenetic analysis of AtCBP60 members with a CBP60 from the moss *Physcomitrella patens* as the outgroup suggested that the immune regulators AtCBP60a, AtCBP60g, and AtSARD1 form one clade and AtCBP60b-f form a separate clade [6].

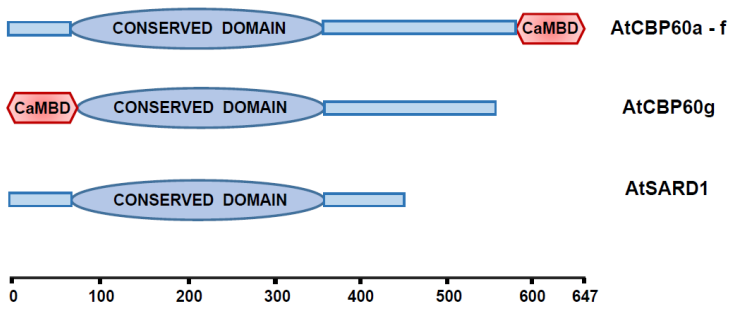


Fig 1. Domain structures of AtCBP60 proteins. The amino acid lengths between the CBP60-conserved domain and the C-CaMBD in AtCBP60a-f vary from 133 to 261: in this figure, the length for AtCBP60b was used.

We applied phylogenetic analysis to the CBP60 protein family members in land plants and found that the immune-related clade neofunctionalized from the highly conserved, prototypical group at or immediately before divergence of Seed Plants. The immune regulator CBP60a, CBP60g, and SARD1 subfamilies diversified within the immune-related clade at or immediately before the emergence of Angiosperms. The immune regulator subfamilies have been evolving very fast, which suggests strong selection imposed by fast-evolving pathogen effectors. We developed an analytical platform, Protein Evolution Analysis in Euclidean Space (PEAES), based on physical-chemical characteristics of amino acids. Using PEAES, we detected significant coevolutionary interactions among the three immune regulator subfamilies in the CBP60-conserved domain in separate Core Eudicot species linages. The observed coevolutionary interactions were consistent with two coevolutionary mechanism hypotheses. First, two positive regulators evolve to be as different as possible, which makes it difficult for a single pathogen effector to target both at once. Second, the negative regulator evolves to stay similar to the positive regulators at the sites where the positive regulators cannot be very different. If a pathogen effector targets both positive and negative regulators, the negative impact of the effector on immunity due to targeting the positive regulators is tempered by its targeting of the negative regulators. Our discoveries suggest that functionally related protein subfamilies

influence the evolution of each other so that high levels of functional resilience can be achieved under strong and rapidly varying selection imposed by fast-evolving pathogen effectors.

RESULTS

The immune-related CBP60 clade neofunctionalized from the prototypical CBP60 group.

We inferred phylogenetic relationships among 1024 CBP60 protein sequences from 247 diverse land plants: one Liverwort, 35 Moss, 9 Lycophyte, 44 Fern, 69 Gymnosperm, and 89 Angiosperm species (Figs 2 and S1; the latter includes sequence names).

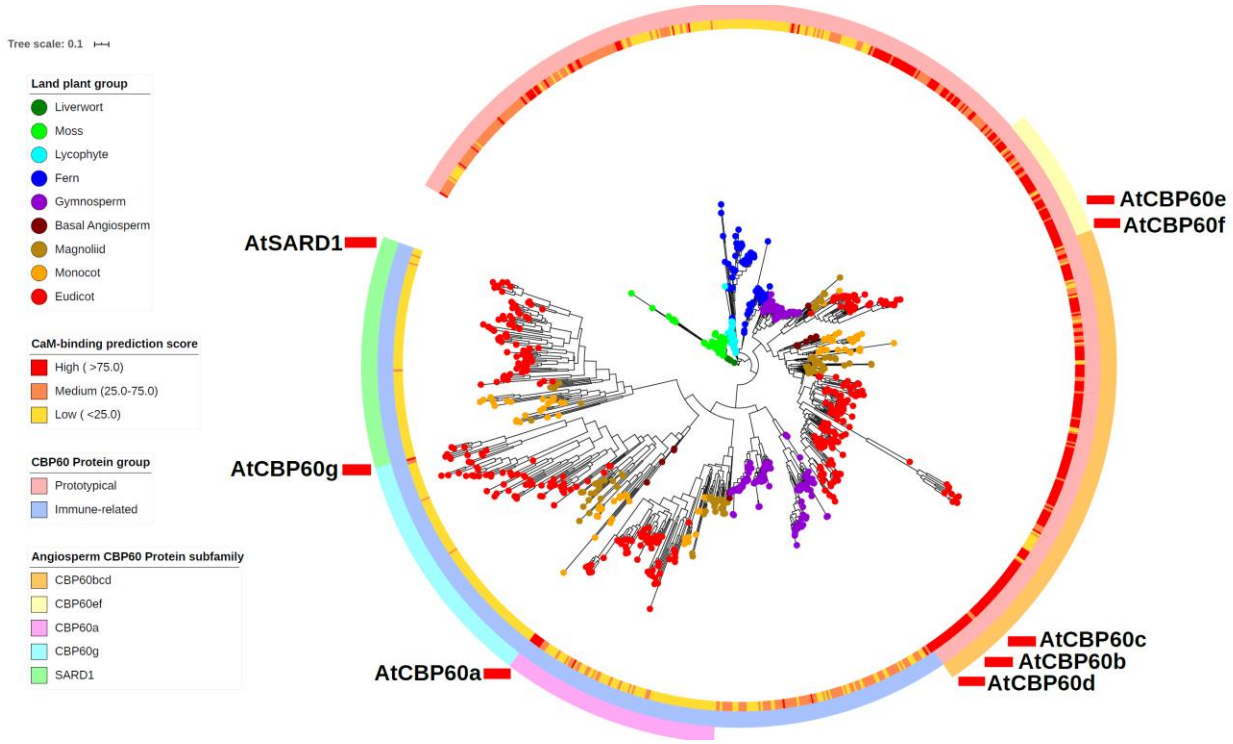


Fig 2. The immune-related clade diversified at or immediately before the Gymnosperm divergence and has been evolving very fast. The phylogenetic tree of 1024 CBP60 protein sequences from 247 land plant species was inferred using RaXML [48] based on ClustalW alignment [43] and visualized using iTOL [45]. Colored dot at the leaf of the tree represents the land plant group (from Liverwort to Angiosperm) the CBP60 member originated from according to the color code for the phylogenetic clade. The inner ring with the yellow (low) to red (high) color range shows the CaM-binding score predicted by CaMELS [13]. The middle ring shows the CBP60 subfamilies according to AtCBP60 names. The outer ring shows the prototypical group (salmon) and the immune-related clade (blue). The AtCBP60 leaf positions are indicated in the outermost layer. The branch length scale of 0.1 (substitution per site) is shown at the top left.

Most plant species had at least one member of the highly conserved, “prototypical” CBP60 group that includes AtCBP60b-f (salmon color arc of the outermost ring in Fig 2). The tree topology indicates that a major clade diversified from the prototypical group at or immediately before the divergence of Gymnosperms (blue arc of the outermost ring in Fig 2). This clade further diversified into the three immune regulator CBP60a, CBP60g, and SARD1 subfamilies (pink, cyan, and green arcs in the second outermost ring in Fig 2) at or immediately before the divergence of Angiosperms. Hence, we call this clade the immune-related clade. It is evident from the branch length that the immune regulator subfamilies have been evolving much faster than the prototypical group. The very fast evolution of the positive immune regulator subfamilies is consistent with the notion that they are subject to selection to evade targeting by fast-evolving pathogen effectors. A fungal effector is known to target AtCBP60g [11]. It is conceivable that multiple effectors target these major positive immune regulator subfamily members at a variety of sites.

None of the six Basal Angiosperm species had SARD1 subfamily members. This could be due to SARD1 deletion in Basal Angiosperms or to SARD1 subfamily diversification from the CBP60g subfamily after Basal Angiosperm divergence. A likelihood-ratio test of the tree topologies for these two possible cases favored the tree structure with SARD1 deletion in Basal Angiosperms ($p < 0.01$; Figs S2A and S2B). We identified multiple SARD1 subfamily members in recently-published genome sequences of a Basal Angiosperm water lily order [12] (Fig S2C). This new information strongly supports our conclusion that diversification of the three immune regulator subfamilies occurred at or immediately before divergence of Angiosperms.

The common ancestor of the immune regulator subfamilies appears to have lost CaM-binding ability

AtCBP60a-f have CaM-binding domains (CaMBD) in their C-terminal regions (C-CaMBD), AtCBP60g has one in its N-terminal region (N-CaMBD), but AtSARD1 lacks CaM-binding ability [5,6,9] (Fig 1). We investigated when the C-CaMBD domain was lost during evolution of the immune regulator subfamilies using CaMELS [13], which was the only algorithm among those we tested that correctly predicted C-CaMBD and N-CaMBD among AtCBP60 members. Generally, C-CaMBD was conserved among CBP60 proteins, except for some specific clades, including the immune regulator subfamilies (Fig 2, innermost ring). C-CaMBD in AtCBP60a and N-CaMBD in AtCBP60g appear to have been reacquired relatively recently. These observations suggest that CaM binding activity may not be essential for the CBP60 immune regulator functions, or alternatively, loss of C-CaMBD may have coincided with evolution of a CaM-binding adaptor protein that works together with the immune regulators. The latter could explain reacquisition of CaMBD in some of the immune regulators.

There were two forms of C-CaMBD loss (Fig 3). One form is deletion of the C-terminal region corresponding to C-CaMBD, such as in the SARD1 subfamily members (Fig 3A). The other form is mutations in C-CaMBD with weakly to moderately alignable sequences remaining in the the C-terminal region. [5]. A secondary structural property of amphipathic helix may be important for CaMBDs [14]. When the CaM-binding prediction scores at the C-CaMBD region were compared to the α-helix length and the hydrophobic moment (as an indication of potential amphipathic α-helix), a long α-helix appeared important while its amphipathicity did not (Fig 3B). Although we cannot exclude a possible high false negative rate in CaMELS predictions,

mutations that led to loss of CaMBD might have disrupted the CaMBDs through disruption of alpha-helices.

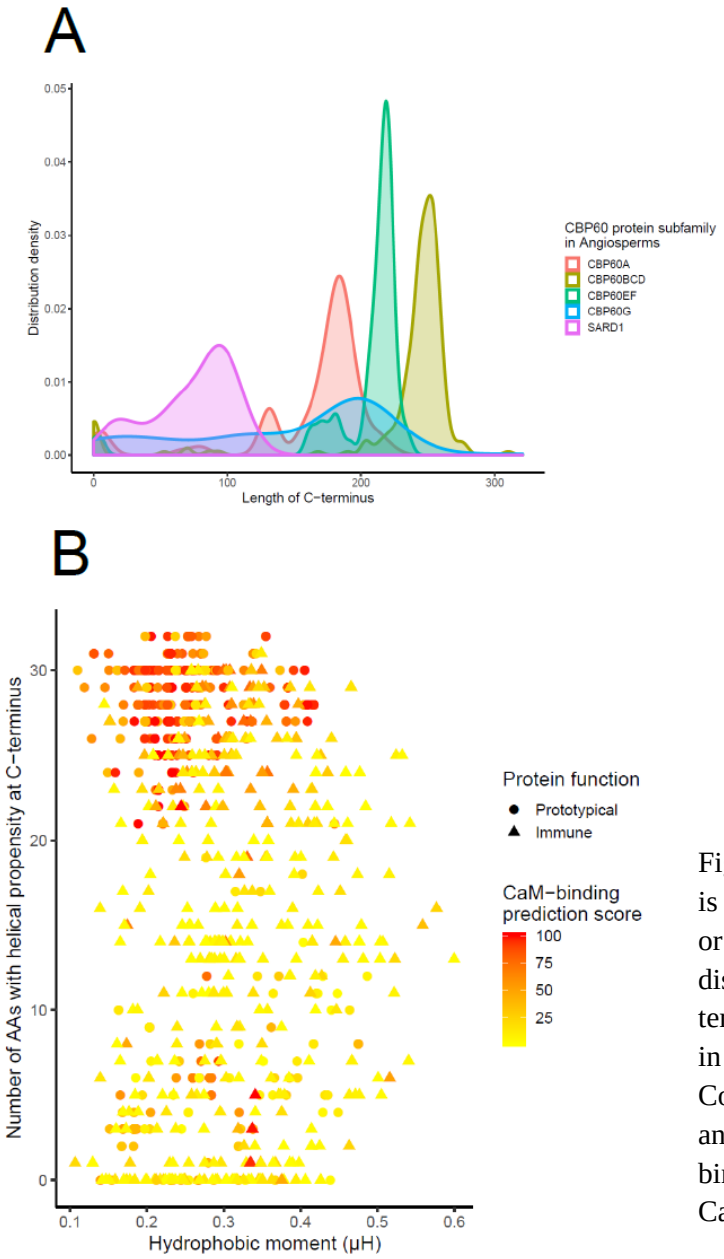


Fig 3. Loss of calmodulin-binding ability is caused by loss of the C-CaMBD region or mutations in C-CaMBD. (A) The distribution of the length of the region C-terminal to the CBP60-conserved domain in each subfamily in Angiosperms. (B) Contributions of the length of alpha-helix and its hydrophobic moment to the CaM-binding prediction score according to the CaMELS algorithm [13].

Protein Evolution Analysis in a Euclidean Space (PEAES)

Phylogenetic analysis of sequences using standard evolution models assumes stationary, reversible, and homogeneous evolution [15]. These conditions are clearly violated when evolution is under strong selection. We postulated that when a site is under strong selection, the

amino acid at the site is selected based mainly on physical-chemical characteristics of amino acids and is relatively independent of the lineage ancestry when the lineage divergence is sufficiently old. Thus, we decided to use a metric of physical-chemical characteristics of amino acids to evaluate relatedness of selection at each site. Venkatarajan and Braun described physical-chemical characteristics of 20 amino acids by five-dimensional Euclidean coordinates after applying linear dimensionality reduction to 237 physical-chemical properties [16]. We added one more dimension to include absence of an amino acid - a gap in an alignment (Table S3). This physical-chemical description of amino acids in a six-dimensional Euclidean space allows geometrical tracking of evolution at each site in a protein sequence alignment. For example, the pairwise Euclidean distance between amino acids can be defined as the amino acid dissimilarity measure (Table S4). We call this geometric analysis platform Protein Evolution Analysis in a Euclidean Space (PEAES).

In the subsequent study, we focused on CBP60 members from Core Eudicots because genome sequences (vs. transcriptome sequences) of many diverse species are available among Core Eudicots. To reduce the influence of conservation due to recently shared ancestry, a single species from each of 12 taxonomic orders of Core Eudicots was chosen for the analysis (Fig S3). To simplify the analysis, when one species has more than one member per subfamily, only one member was selected. This subfamily member selection should not introduce strong bias: in all such cases, the members in a single subfamily were paralogs diversified after divergence of the corresponding taxonomic orders. We also conducted the subsequent analyses with an alternative set of selected CBP60 members and obtained similar observations (indicated as “the alternative set” for the corresponding supplemental figures), confirming the absence of strong bias due to CBP60 member selection.

We focused on the CBP60-conserved domain because it is impossible to obtain a reliable multiple sequence alignment across the immune regulator subfamilies outside the conserved domain (Fig S4). The CBP60-conserved domain sequence that is highly conserved in the prototypical group contains 293 aa. To align the domain sequences from the immune regulator subfamilies together, three gaps were added, resulting in the multiple sequence alignment consisting of 296 sites (Fig S5). The site position within this 296-site alignment is used subsequently.

With PEAES, the measure of diversity at each site in the aligned sequences was defined as the mean of all pairwise distances among amino acids at that site. Fig 4 (Fig S6 for the alternative set) shows the conservation and diversity at each site in different CBP60 subfamilies. The CBP60bcd subfamily represents the prototypical group. It is evident that while many sites are highly conserved in the CBP60bcd subfamily, many sites in each immune regulator subfamily, particularly the CBP60g subfamily, are highly diverse. The means of the diversity measure across the sites are 0.075, 0.178, 0.291, and 0.231 for the CBP60bcd, CBP60a, CBP60g, and SARD1 subfamilies, respectively (dashed horizontal lines in Fig 4). The numbers of highly diverse sites (site diversity measure > 0.4417) are 5, 33, 81, and 58, while the numbers of subfamily-specific highly conserved sites (site diversity measure < 0.0764) are 134, 60, 13, and 33 for the CBP60bcd, CBP60a, CBP60g, and SARD1 subfamilies, respectively. These observations further demonstrate that the immune regulator subfamilies have been evolving very fast compared to the prototypical group.

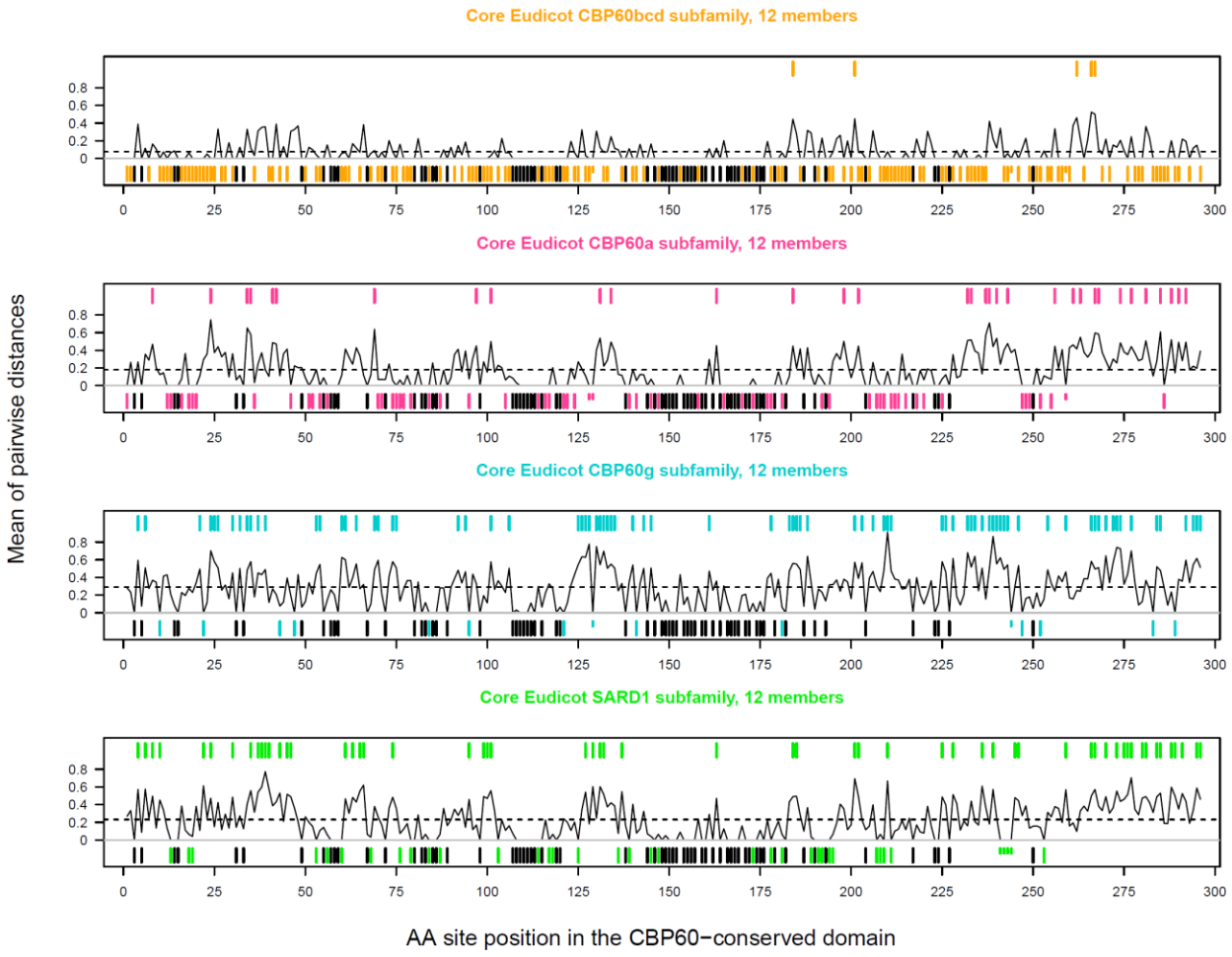


Fig 4. The CBP60-conserved domains of the immune regulator subfamilies are highly diverse among 12 diverse Core Eudicot species. The mean pairwise PEAES distance for each site across the species is shown as a line plot in each panel. One CBP60 member per subfamily was selected for each species. Highly conserved sites (pairwise distance mean ≤ 0.0764) are shown as segments at the bottom in each plot. The segments in black are the sites highly conserved among all CBP60 members of the Core Eudicot species. The subfamily-specific sites are shown in orange, salmon, cyan, and green for the CBP60bcd, CBP60a, CBP60g, and SARD1 subfamilies, respectively. Short segments represent subfamily-conserved gaps in the alignment. Highly diverse sites (pairwise distance mean > 0.4417) are shown at the top in each plot as segments colored according to the subfamily-color code.

Highly non-random, lineage-specific coevolutionary interactions across the immune regulator subfamilies are prevalent.

As they share a common ancestor as recently as immediately before divergence of Angiosperms and they have overlapping or opposing functions in immune signaling, the immune regulators CBP60a, CBP60g, and SARD1 likely form a regulatory module in the immune signaling network. We investigated the possibility that coevolution of the regulatory module components confers resilience against negative impacts of pathogen effectors targeting the module components. There may be pressure for the partially functionally-redundant, positive immune regulators, CBP60g and SARD1, to be as dissimilar as possible at critical sites, which would reduce the probability that they are both targeted by a single pathogen effector (Hypothesis 1). On the other hand, if both the negative regulator CBP60a and one of the positive regulators are targeted by a single pathogen effector, simultaneous impairment of the positive and negative regulators would moderate the negative impact of the effector on immunity [7]. If this is the case, the negative immune regulator CBP60a may be under pressure to remain similar to one or both of the positive regulators at critical sites (Hypothesis 2). Such selection imposed by pathogen effectors could be strong. In addition, the direction of selection and the sites under selection could change over short time periods within each plant species lineage. This is because pathogen effectors evolve very fast and multiple effectors from multiple pathogens may target the regulatory module of a single plant species. To detect coevolutionary interactions under rapidly changing selection, the analysis used must be specific to the site and the species lineage.

We defined the pairwise distance rank (PDR) of a single site as the rank of the pairwise PEAES distance between the members of two subfamilies of a particular species lineage among all permutations of the species lineages for the member of one of the two subfamilies (Fig 5). To allow a species lineage-specific analysis, two hypotheses about the coevolutionary interactions among the three immune regulator subfamilies were derived from Hypotheses 1 and 2 described

above.

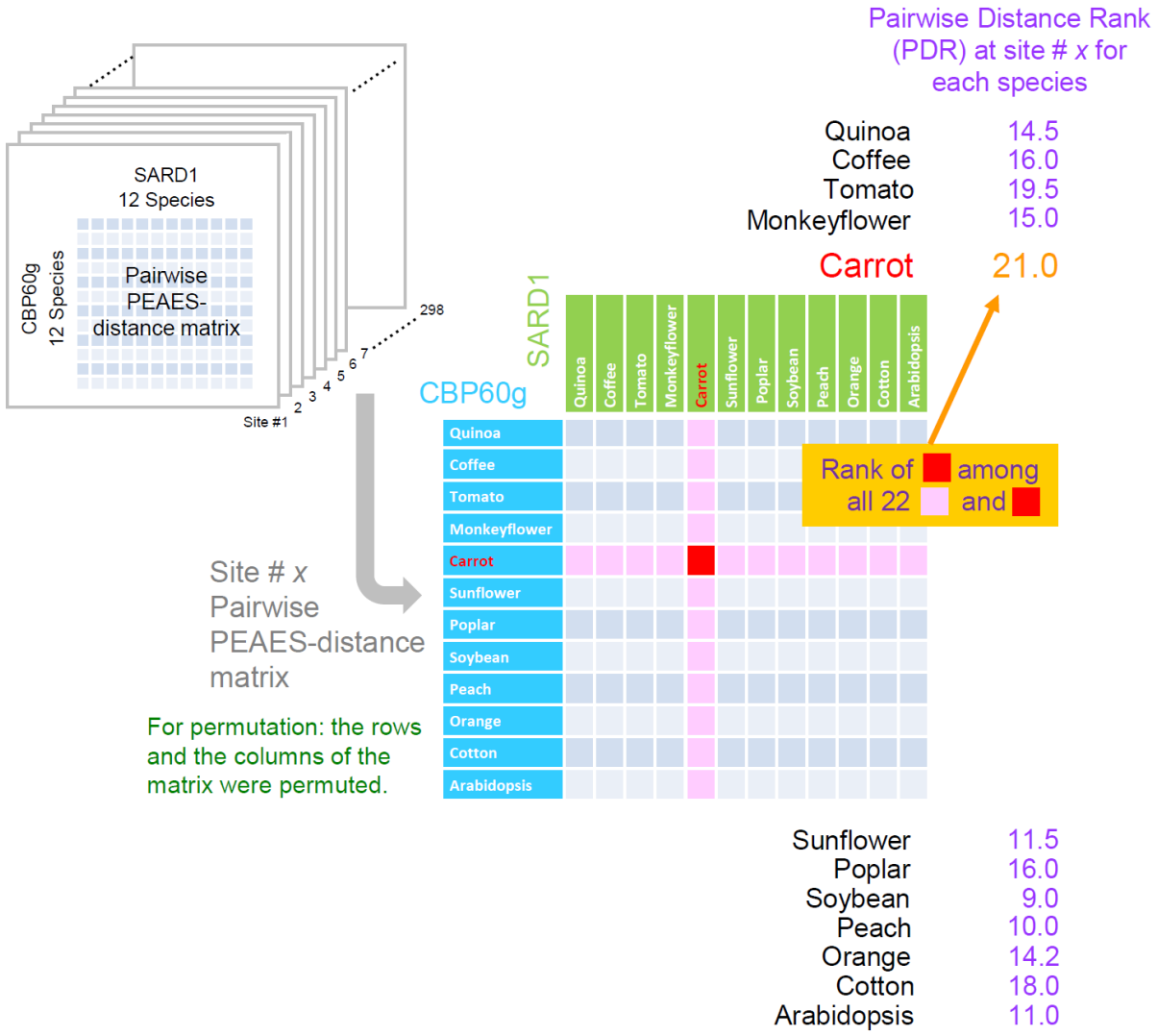
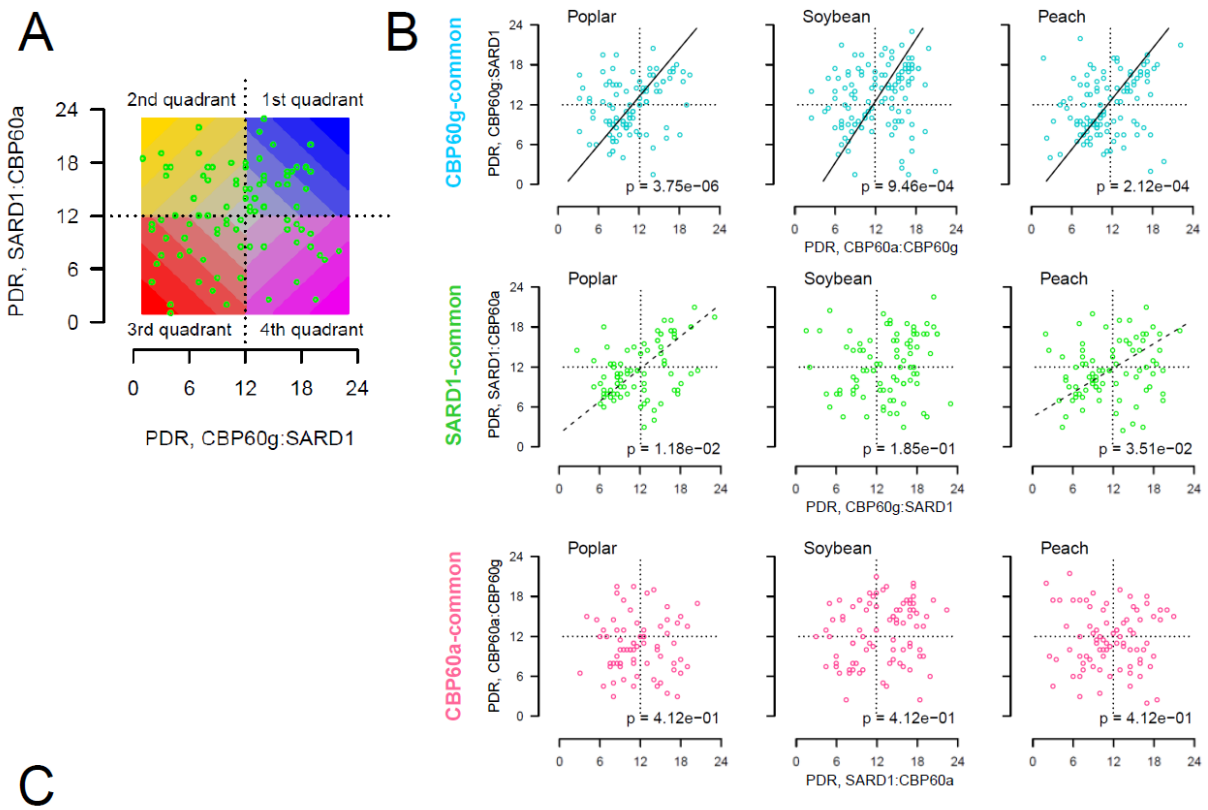


Fig 5. Pairwise distance ranks (PDRs) based on the PEAES distance metric. First, for each of the 296 amino acid sites in the multiple sequence alignment of the CBP60-conserved domain, a pairwise PEAES-distance matrix was made between members of two subfamilies (CBP60g and SARD1 in the figure as an example) in the 12 Core Eudicot species. Second, in the pairwise PEAES-distance matrix at a particular site (Site #*x*), for each species (Carrot in the figure as an example) the pairwise distance between the members of the two subfamilies of the species (red square in the matrix) was compared to the pairwise distance between the member of one subfamily of this species (Carrot) and the member of the other subfamily of a different species (pink squares). The rank of the within-species comparison was determined among these 23 pairwise distances. This rank (21.0 for Carrot) is called the pairwise distance rank (PDR) at each site in each species.



	Quinoa	Coffee	Tomato	Mokeyflower	Carrot	Sunflower	Poplar	Soybean	Peach	Oragne	Cotton	Arabidopsis
CBP60g-common	6.6×10^{-2}	1.4×10^{-3}	2.0×10^{-2}	2.9×10^{-2}	2.2×10^{-1}	1.4×10^{-4}	3.8×10^{-6}	9.5×10^{-4}	2.1×10^{-4}	3.8×10^{-6}	3.0×10^{-3}	3.6×10^{-3}
SARD1-common	5.3×10^{-2}	2.1×10^{-2}	3.5×10^{-2}	4.1×10^{-2}	1.7×10^{-3}	7.3×10^{-2}	1.2×10^{-2}	1.9×10^{-1}	3.5×10^{-2}	4.1×10^{-2}	1.1×10^{-1}	1
CBP60a-common	4.0×10^{-1}	4.0×10^{-1}	4.1×10^{-1}	4.1×10^{-1}	4.1×10^{-1}	1.4×10^{-1}	4.1×10^{-1}	4.1×10^{-1}	4.1×10^{-1}	6.5×10^{-1}	4.1×10^{-1}	3.2×10^{-2}

Fig 6. Non-random coevolutionary interactions that could protect the positive immune regulators are prevalent among the 12 Core Eudicot lineages. (A) Four quadrants in the SARD1:CBP60a PDR vs. CBP60g:SARD1 PDR plot (SARD1-common comparison) as an example. The color codes for the quadrants shown are used in Fig 7. (B) Plots for CBP60g:SARD1 PDR vs. CBP60a:CBP60g PDR (CBP60g common comparison; cyan color), SARD1:CBP60a PDR vs. CBP60g:SARD1 PDR (SARD1-common comparison; green color), and CBP60a:CBP60g PDR vs. SARD1:CBP60a PDR (CBP60a-common comparison; salmon color) are shown for three plant species, Poplar, Soybean, and Peach. Similar plots for all 12 Core Eudicot species are shown in Fig S7. The *p*-value (Fisher's exact test, 2-sided, Benjamini-Hochberg FDR-corrected) is shown in the bottom right. When $p < 0.01$ or $0.01 \leq p < 0.05$, the PC1 (principal component 1) axis is shown as a black solid line or a dashed line, respectively, indicating significant enrichment of sites in the 1st and 3rd quadrants over the 2nd and 4th quadrants. (C) The *p*-values for CBP60g-common, SARD1-common, and CBP60a-common comparisons are shown for the 12 Core Eudicot species. Red background, $p < 0.01$; yellow background, $0.01 \leq p < 0.05$.

If a site is not dissimilar between CBP60g and SARD1, CBP60a should have an amino acid similar to the amino acid in CBP60g or SARD1 to confer a protection effect (Hypothesis 3). On the other hand, at a site where CBP60a cannot provide protection through similarity to CBP60g or SARD1 (e.g., because it would impair CBP60a function), CBP60g and SARD1 should be dissimilar to avoid both proteins being targeted by a single effector (Hypothesis 4). To test these hypotheses, the PDR between CBP60g (or SARD1) and CBP60a was plotted against the PDR between CBP60g and SARD1 for all sites of a single species. The plot was divided into four quadrants at the median rank of 12 on both axes (Fig 6A). Hypotheses 3 and 4 predict that the 1st and 3rd quadrants will be significantly enriched with sites. Enrichment in the 3rd quadrant is consistent with Hypothesis 3, and enrichment in the 1st quadrant is consistent with Hypothesis 4. The significance of the site enrichment in the 1st and 3rd quadrants over the 2nd and 4th quadrants was tested by Fisher's exact test (2-sided).

Significant enrichment of the sites in the 1st and 3rd quadrants (Benjamini-Hochberg FDR = 0.05) was observed in 10 out of 12 species lineages in the comparison of CBP60g:SARD1 PDR vs. CBP60a:CBP60g PDR (Fig 6B, 6C, and S7; CBP60g-common comparison) and in 7 out of 12 species lineages in the comparison of SARD1:CBP60a PDR vs. CBP60g:SARD1 PDR (SARD1-common comparison) (Fig S8 for the alternative set). In the union of the CBP60g- and SARD1-common comparisons, 11 out of 12 species lineages showed non-random coevolutionary interactions among the three immune regulator subfamilies, which are consistent with Hypotheses 3 and 4 (and consequently with Hypotheses 1 and 2). Even Quinoa, the only species that did not show significant enrichment, the FDR-corrected *p*-values were 0.066 and 0.053 for CBP60g- and SARD1-common comparisons, respectively (Fig 6C). Only one out of 12 species lineages shows a moderately significant enrichment in the 1st and 3rd quadrants in the

comparison of CBP60a:CBP60g PDR vs. SARD1:CBP60a PDR (CBP60a-common comparison; Fig 6C), in which no enrichment was predicted by the hypotheses.

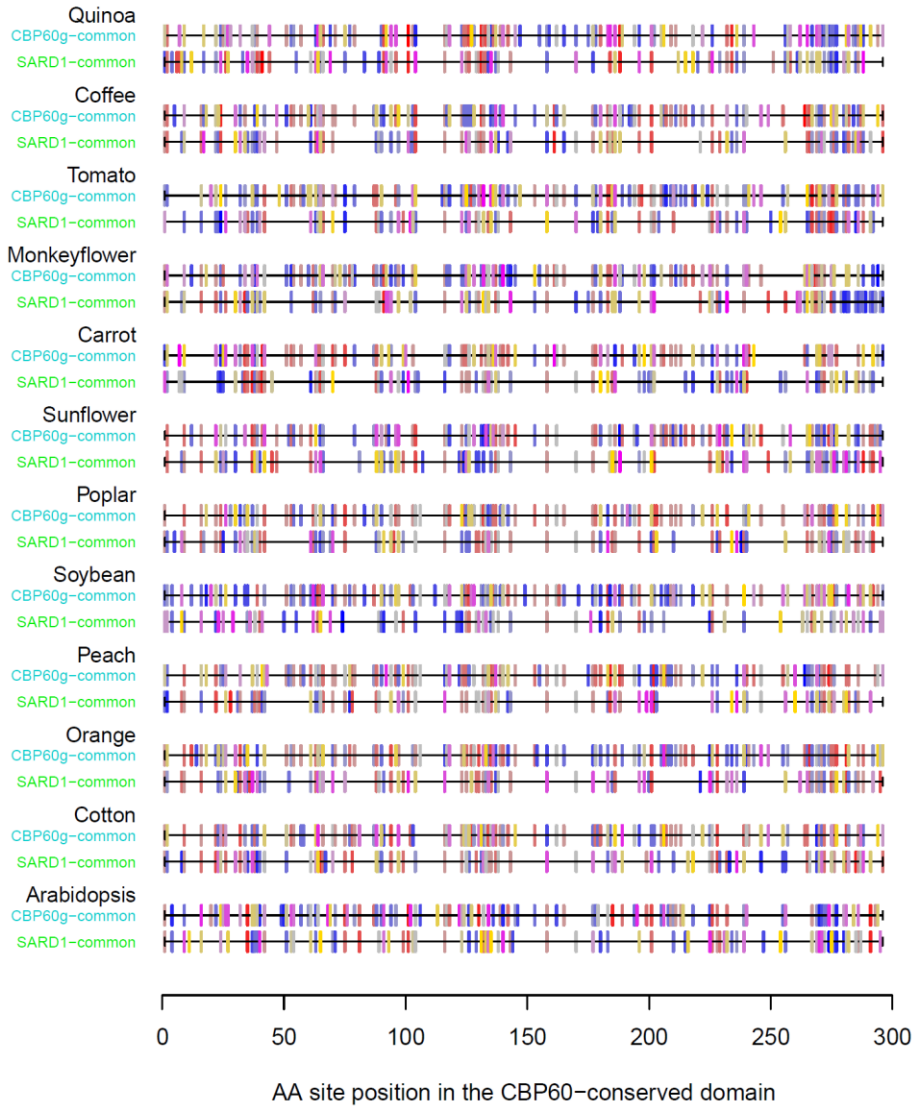


Fig 7. Site selections are largely specific to each of the Core Eudicot lineages. The sites in CBP60g-common and SARD1-common comparisons in each plant species (Figs S7A and S7B) are shown along the CBP60-conserved domain with the color-scaled quadrant information (Fig 6A).

The sites of CBP60g- and SARD1-common comparisons are shown along the CBP60-conserved domain in Fig 7 (Fig S9 for the alternative set), with the site-indicating bars color-coded according to the quadrants as in Fig 6A. The distribution of the colors of site bars does not show

strong similarity between more closely related species, consistent with our assumption that the species lineage divergences were sufficiently old and that observed selections at the sites are largely consequences of species lineage-specific selection after lineage divergence. The notion of species lineage-specific selection is further supported by the fact that the values obtained by projecting the centered data onto $y = x$ in each plot in Fig S7 show no obvious trend toward higher correlation between more closely related species pairs (Fig S10; Fig S11 for the alternative set). This projection onto $y = x$ signifies the enrichment in the 1st and 3rd quadrants (Fig S10A). We conclude that many sites in the CBP60-conserved domain of the three immune regulator subfamilies have likely been coevolving under strong and varying pressure from pathogen effectors in a species lineage-specific manner, conferring a high level of resilience on the CBP60 immune regulatory module.

DISCUSSION

One type of protein family evolution analysis focuses on emergence of conserved domain structures [17]. However, emergence of particular conserved domains may not provide information about evolution of particular biological processes and functions underlying them since some domains, such as DNA-binding domains, have versatile molecular functions that can be easily repurposed for different biological processes (e.g., [18]). Since the CBP60-conserved domains of AtCBP60g and AtSARD1 have sequence-specific DNA-binding activity [8,11], CBP60 proteins are likely DNA-binding transcription factors. A DNA-binding domain structure provides a backbone structure for physical interactions with DNA, and its binding specificity could vary in different proteins. It is possible that diversification of the immune-related clade from the prototypical group coincided with a substantial DNA-binding specificity change and

consequently resulted in a large change in the gene set regulated by CBP60 transcription factors, i.e., immune-related CBP60 transcription factors have likely been neofunctionalized for a different biological process. Thus, tracking neofunctionalization of particular subfamilies important for biological processes of interest within a larger protein family defined by the conserved domain structure provides insights into evolution of biological processes.

AtCBP60g, AtSARD1, and AtCBP60a belong to the immune-related clade and regulate immune responses, including SA signaling, in Arabidopsis. The fact that Non-Seed Plants, namely Liverworts, Mosses, Lycophytes, and Ferns, do not have a member of the immune-related clade strongly suggests that SA signaling and SA-dependent immunity are regulated differently between Angiosperms and Non-Seed Plants, or that SA-dependent immunity is absent in Non-Seed Plants. As SA is a major immune hormone in Angiosperms [19], the organization of immune signaling networks could be very different between Angiosperms and Non-Seed Plants. However, this notion of different network organization does not exclude the possibility that Angiosperms and Non-Seed Plants share some elemental immune signaling machineries [20]. Instead, it implies that the properties of the immune signaling networks, such as network resilience, are very different between Angiosperms and Non-Seed Plants.

We discovered significant coevolutionary interactions at multiple sites in the CBP60-conserved domain across the immune regulator CBP60a, CBP60g, and SARD1 subfamilies in separate Core Eudicot species lineages, using PDR with the PEAES metric (PEAES-PDR; Figs 5-7 and S7-S10). These observations indicate that although the three immune regulator subfamilies diversified at or immediately before emergence of Angiosperms, the immune regulator subfamilies have been influencing fast evolution of one another in each plant species lineage. These coevolutionary interactions appear to have been maintained through the history of

selection in the plant species lineage-specific and fast-changing pathogen effector landscape. Thus, our discovery presents a new pattern of gene evolution after gene duplication: long lasting coevolution of duplicated (or multiplicated) genes while different species lineages have experienced different selection histories, resulting in lineage-specific selection outcomes for the coevolving gene pair.

The observed coevolutionary interactions are consistent with our hypotheses of coevolution mechanisms that promote functional resilience of the positive immune regulators, CBP60g and SARD1, against pathogen effectors. Hypothesis 1- Coevolution occurs via selection for amino acid dissimilarity between functionally overlapping, positive immune regulators to prevent simultaneous targeting by a single effector. Hypothesis 2 – Coevolution occurs via selection for amino acid similarity between functionally opposing, positive and negative immune regulators to assure simultaneous targeting by a single effector. There are other important immune signaling network components in Arabidopsis that are members of subfamilies within larger protein families including AtMPK3 and AtMPK6 [21], AtEDS1, AtPAD4, and AtSAG101 [22], and AtNPR1, AtNPR3, and AtNPR4 [23,24]. It will be interesting to apply the PEAES-PDR approach to these protein subfamilies to test whether their evolutionary patterns are also consistent with these hypotheses about coevolution mechanisms that lead to increased resilience against pathogen effectors of the regulatory modules.

The results of the PEAES-PDR analysis can also generate specific, mechanistic hypotheses. First, Qin et al. reported that AtCBP60g is targeted by the fungal effector VdSCP41 [11]. Although AtSARD1 can coimmunoprecipitate VdSCP41 after overexpression in Arabidopsis protoplasts, this AtSARD1-VdSCP41 interaction appears to be much weaker than the AtCBP60g-VdSCP41 interaction (Figure 3A, Figure 4C, and Figure 3—figure supplement 2A in

[11]). These observations may suggest that AtSARD1 mostly evades being targeted by VdSCP41. VdSCP41 binds a C-terminal region of AtCBP60g (corresponding to the region C-terminal to site 130 of the CBP60-conserved domain) [11]. Our PEAES-PDR analysis detected multiple AtCBP60g sites C-terminal to site 130 potentially protected by dissimilarity between AtCBP60g and AtSARD1 (blue and magenta color bars on the line for Arabidopsis, CBP60g in Fig 7). If VdSCP41 or a very similar effector is a major driver for the dissimilarity between AtCBP60g and AtSARD1, substituting the amino acids at these sites in AtCBP60g to those at the sites in AtSARD1 may substantially weaken the interaction between AtCBP60g and VdSCP41.

Second, the sites indicated by yellow-colored bars in Fig 7 (i.e., sites in the 2nd quadrant in Fig 6A) are those that are not protected either by dissimilarity between two positive regulators or by similarity between a positive and the negative regulators. In some cases, the yellow-colored sites coincide in both CBP60g- and SARD1-common comparisons in a single species. One possibility is that such sites may not have been targeted by any effectors in a recent lineage history, and hence there is no signature of protective selection. Another possibility is that they may be protected by some other molecular mechanisms, such as immunity mediated by specific resistance (R) proteins [25]. A guard-type R protein detects some biochemical change caused by a pathogen effector in the cognate host protein and triggers immune responses. If there is protection by an R protein, keeping the amino acids at the yellow-colored sites similar between CBP60 and SARD1 may be important for function of the R protein-based protection mechanism. Identification of proteins that interact with CBP60g and SARD1 whose interactions are sensitive to substitutions of these amino acids at the yellow-colored sites to those at the same sites in other species may lead to discovery of other protection mechanisms.

We demonstrated that coevolutionary interactions across the immune regulator CBP60 subfamilies of overlapping or opposing functions are consistent with the mechanistic hypotheses for increased resilience of this particular regulatory module in the immune signaling network. It is conceivable that this concept can be more generalized among immune signaling network components. Some pathogen effectors target multiple components of the immune signaling network [1], suggesting existence of many other multiple-targeting effectors. At the same time, immune signaling network components are highly interconnected [26], and it is common for them to have overlapping or opposing effects on immunity. These two conditions, multiply targeting pathogen effectors and potential functional interactions among the effector targets, are exactly the conditions that appear to have been driving coevolution for increased resilience of the CBP60 immune regulator module. We developed a PEAES-PDR approach, which was powerful in detecting this type of coevolutionary interaction among alignable sequences and in generating specific and mechanistic hypotheses. Expanding coevolutionary interaction investigations to unalignable sequences of multiply targeted proteins will further advance our understanding of how the immune signaling network has been evolving to maintain or improve resilience in a strongly-selecting and rapidly-changing pathogenic landscape. This will be an exciting future direction in the study of biological network evolution.

METHODS

The CBP60 protein sequence set

A protein sequence database consisting of sequences from 271 species was constructed. The sequence data sources [27–42] are listed in Table S1. The database was searched by BLASTP using each of the Arabidopsis CBP60 members as a query and the union of the subject sequences

that yielded bit scores higher than the Arabidopsis CBP60 that yielded the lowest bit score was identified as the CBP60 protein sequence set, consisting of 1432 sequences. Thus, the diversity of the CBP60 member set was controlled by the level of diversity among AtCBP60 members. We call this member identification procedure a Furthest Member Selection (FMS) approach. This FMS-based CBP60 protein sequence set did not include some distantly related CBP60 homologs (Text S1). We left them out of the set because they do not allow functional interpretations of CBP60 evolution since our knowledge of CBP60 functions is limited to the Arabidopsis members and their orthologs. The CBP60 protein sequence set was subjected to quality controls to remove possible products of assembly or annotation errors and to limit sequence representation to one species per genus for balanced overall representation. These quality controls resulted in the final set consisting of 1024 sequences from 247 species (Table S2). To ease identification of plant species, their common names or genus names are used.

Phylogenetic analysis of CBP60 protein sequences

Generally, ClustalW [43], a Maximum Likelihood (ML) method [44], and iTOL [45] were used for multiple sequence alignment, tree inference, and tree visualization, respectively. ML was also used for inference of most probable ancestor sequences at branch points in trees. MEGA7 [46] was used for ClustalW and ML with the default parameters, except for 80% site coverage in ML (unless stated otherwise). We used the Most Recent Common Ancestor (MRCA)-Anchored, Clade-by-Clade Tree Inference (MACCTI) approach, in which the aligned sequence region used for tree inference is locally adjusted. Briefly, an overview tree for relationships among all major clades and MRCAs on the main CBP60 lineages was inferred using only the sequences with alignable C-terminal regions, the sequences of each clade and its

flanking MRCAs on the main CBP60 lineages were used to infer the subtree for the clade, the MRCAs of the clades were used to infer the backbone tree showing the relationships among the clade MRCAs, and the clade MRCAs in the backbone tree were replaced with the corresponding clade trees to reconstruct the entire tree. Since species phylogeny, duplication, and deletion (SPDD rule) generally explained our tree of 1024 CBP60 sequences at high levels, we concluded that the tree was sufficiently accurate to support the analysis presented here.

CaM-binding site prediction

CaMELS [13] was used for CaM-binding site prediction. The DECIPHER R Bioconductor package [47] was used to predict alpha-helical propensity. A Python script, hydrophobic_moment.py (https://gist.github.com/JoaoRodrigues/568c845915aea3efa3578babfd72423c) was used for calculation of the hydrophobic moment.

PEAES-PDR analysis

PEAES is a way to describe the physical-chemical characteristics of 20 amino acids and a gap in a 6-dimensional Euclidean space. A gap dimension was added to the 5-dimensional descriptions of physical-chemical characteristics of 20 proteinogenic amino acids by [16] (Table S3). Thus, the dissimilarity between two amino acids (or between an amino acid and a gap) is defined by the Euclidean distance between them in the 6-dimensional space (Table S4). Twelve Eudicot species with high quality genome sequences were selected to represent 12 diverse taxonomic orders (Fig S3). One sequence for each of the CBP60a, CBP60bcd, CBP60g, and SARD1 subfamilies was selected for each species (48 sequences). Since some species had more

than one paralog, two sets of 48 sequences were made by selecting different paralogs from these species. Each of these two sets was analyzed separately. The CBP60-conserved domains of 48 sequences were aligned by ClustalW [43], and then the multiple sequence alignment was manually edited to remove insertional polymorphisms represented by only one or two sequences and to make the alignments consistent between the two sets. The alignments for the two sets are provided as supplemental files. To quantify the level of diversity at each site, the mean pairwise PEAES distance was calculated for each site for each subfamily (Fig 4). PDR was calculated at each site of the CBP60-conserved domain multiple sequence alignment between two CBP60 subfamilies for each of the 12 Eudicot species based on the PEAES metric according to the procedure depicted in Fig 5. If there are no species lineage-specific coevolutionary interactions between two CBP60 subfamilies, the mean expectation for PDR is the median rank of 12 among 23 pairwise PEAES distances. Thus, when PDR values for the corresponding sites in two subfamily pairs are plotted against each other, the random expectation is that the sites would distribute equally among the four quadrants of the plot divided at the median ranks of 12 (Fig 6A). Hypotheses 3 and 4 predicted enrichment of the sites in quadrants 1 and 3. The *p*-value associated with the odds ratio for this enrichment was calculated by Fisher's exact test for each species for particular two-subfamily pairs. The *p*-value was corrected for multiple tests by Benjamini-Hochberg's False Discovery Rate across the species in each combination of two-subfamily pairs.

Detailed methods and discoveries made based on the phylogenetic relationships but not described in the main text are in Text S2.

AUTHOR CONTRIBUTIONS

F.K. conceived the project. Q.Z. assembled the protein sequence database. Q.Z., K.M., and F.K. analyzed the data. K.M. and F.K. wrote the manuscript.

ACKNOWLEDGEMENTS

We thank Gane Ka-Shu Wong and Eric Carpenter for allowing us access to the 1000 Plants transcriptomic sequence database and Fay-Wei Li for allowing us access to the Waterfern and Floating fern genome sequences, prior to their publications. We also thank Yaniv Brandvain, Ya Yang, and Emma Goldberg for technical advice and discussion and Jane Glazebrook for editing of the manuscript. Q. Z. was supported by a scholarship from China Scholarship Council and funds from the Zhiming Zhang Laboratory. This research was funded by the National Science Foundation grants MCB-1518058 and IOS-1645460 to F. K.

REFERENCES

1. Toruño, T.Y., Stergiopoulos, I., and Coaker, G. (2016). Plant-Pathogen Effectors: Cellular Probes Interfering with Plant Defenses in Spatial and Temporal Manners. Annu. Rev. Phytopathol.
2. Katagiri, F. (2018). Review: Plant immune signaling from a network perspective. Plant Sci. *276*, 14–21.
3. Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J., and Katagiri, F. (2009). Network properties of robust immunity in plants. PLoS Genet. *5*.
4. Hillmer, R.A., Tsuda, K., Rallapalli, G., Asai, S., Truman, W., Papke, M.D., Sakakibara,

H., Jones, J.D.G., Myers, C.L., and Katagiri, F. (2017). The highly buffered Arabidopsis immune signaling network conceals the functions of its components. PLoS Genet. *13*.

5. Wang, L., Tsuda, K., Sato, M., Cohen, J.D., Katagiri, F., and Glazebrook, J. (2009). Arabidopsis CaM binding protein CBP60g contributes to MAMP-induced SA accumulation and is involved in disease resistance against Pseudomonas syringae. PLoS Pathog. *5*, e1000301. Available at: https://dx.plos.org/10.1371/journal.ppat.1000301.

6. Wang, L., Tsuda, K., Truman, W., Sato, M., Nguyen, L. V., Katagiri, F., and Glazebrook, J. (2011). CBP60g and SARD1 play partially redundant critical roles in salicylic acid signaling. Plant J. *67*, 1029–1041. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21615571.

7. Truman, W., Sreekanta, S., Lu, Y., Bethke, G., Tsuda, K., Katagiri, F., and Glazebrook, J. (2013). The CALMODULIN-BINDING PROTEIN60 family includes both negative and positive regulators of plant immunity. Plant Physiol. *163*, 1741–1751.

8. Zhang, Y., Xu, S., Ding, P., Wang, D., Cheng, Y.T., He, J., Gao, M., Xu, F., Li, Y., Zhu, Z., *et al.* (2010). Control of salicylic acid synthesis and systemic acquired resistance by two members of a plant-specific family of transcription factors. Proc. Natl. Acad. Sci. U. S. A.

9. Reddy, V.S., Ali, G.S., and Reddy, A.S.N. (2002). Genes encoding calmodulin-binding proteins in the Arabidopsis genome. J. Biol. Chem.

10. Lu, Y., Truman, W., Liu, X., Bethke, G., Zhou, M., Myers, C.L., Katagiri, F., and Glazebrook, J. (2018). Different modes of negative regulation of plant immunity by calmodulin-related genes. Plant Physiol. *176*, 3046–3061.

11. Qin, J., Wang, K., Sun, L., Xing, H., Wang, S., Li, L., Chen, S., Guo, H.S., and Zhang, J.

(2018). The plant-specific transcription factors CBP60G and SARD1 are targeted by a verticillium secretory protein VDSCP41 to modulate immunity. Elife.

12. Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., Chang, X., Dong, W., Ho, S.Y.W., Liu, X., *et al.* (2020). The water lily genome and the early evolution of flowering plants. Nature.

13. Abbasi, W.A., Asif, A., Andleeb, S., and Minhas, F. ul A.A. (2017). CaMELS: In silico prediction of calmodulin binding proteins and their binding sites. Proteins Struct. Funct. Bioinforma.

14. Degrado, W.F., Erickson-Viitanen, S., Wolfe, H.R., and O'Neil, K.T. (1987). Predicted calmodulin-binding sequence in the γ subunit of phosphorylase b kinase. Proteins Struct. Funct. Bioinforma.

15. Naser-Khdour, S., Minh, B.Q., Zhang, W., Stone, E.A., Lanfear, R., and Bryant, D. (2019). The Prevalence and Impact of Model Violations in Phylogenetic Analysis. Genome Biol. Evol.

16. Venkatarajan, M., and Braun, W. (2001). New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. J. Mol. Model.

17. Domazet-Lošo, T., and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature.

18. Gordân, R., Murphy, K.F., McCord, R.P., Zhu, C., Vedenko, A., and Bulyk, M.L. (2011). Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol.

19. Vlot, A.C., Dempsey, D.A., and Klessig, D.F. (2009). Salicylic Acid, a Multifaceted

Hormone to Combat Disease. Annu. Rev. Phytopathol.

20. de Vries, S., de Vries, J., von Dahlen, J.K., Gould, S.B., Archibald, J.M., Rose, L.E., and Slamovits, C.H. (2018). On plant defense signaling networks and early land plant evolution. Commun. Integr. Biol.

21. Xu, J., Xie, J., Yan, C., Zou, X., Ren, D., and Zhang, S. (2014). A chemical genetic approach demonstrates that MPK3/MPK6 activation and NADPH oxidase-mediated oxidative burst are two independent signaling events in plant immunity. Plant J.

22. Feys, B.J., Wiermer, M., Bhat, R.A., Moisan, L.J., Medina-Escobar, N., Neu, C., Cabral, A., and Parker, J.E. (2005). Arabidopsis SENESCENCE-ASSOCIATED GENE101 stabilizes and signals within an ENHANCED DISEASE SUSCEPTIBILITY1 complex in plant innate immunity. Plant Cell.

23. Ding, Y., Sun, T., Ao, K., Peng, Y., Zhang, Y., Li, X., and Zhang, Y. (2018). Opposite Roles of Salicylic Acid Receptors NPR1 and NPR3/NPR4 in Transcriptional Regulation of Plant Immunity. Cell.

24. Fu, Z.Q., Yan, S., Saleh, A., Wang, W., Ruble, J., Oka, N., Mohan, R., Spoel, S.H., Tada, Y., Zheng, N., *et al.* (2012). NPR3 and NPR4 are receptors for the immune signal salicylic acid in plants. Nature.

25. Khan, M., Subramaniam, R., and Desveaux, D. (2016). Of guards, decoys, baits and traps: Pathogen perception in plants by type III effector sensors. Curr. Opin. Microbiol.

26. Sato, M., Tsuda, K., Wang, L., Coller, J., Watanabe, Y., Glazebrook, J., and Katagiri, F. (2010). Network modeling reveals prevalent negative regulatory relationships between signaling sectors in arabidopsis immune signaling. PLoS Pathog. *6*, 1–15.

27. Li, F.W., Brouwer, P., Carretero-Paulet, L., Cheng, S., De Vries, J., Delaux, P.M., Eily,

A., Koppers, N., Kuo, L.Y., Li, Z., *et al.* (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nat. Plants.

28. Yagi, M., Kosugi, S., Hirakawa, H., Ohmiya, A., Tanase, K., Harada, T., Kishimoto, K., Nakayama, M., Ichimura, K., Onozaki, T., *et al.* (2014). Sequence analysis of the genome of carnation (Dianthus caryophyllus L.). DNA Res.

29. Al-Dous, E.K., George, B., Al-Mahmoud, M.E., Al-Jaber, M.Y., Wang, H., Salameh, Y.M., Al-Azwani, E.K., Chaluvadi, S., Pontaroli, A.C., Debarry, J., *et al.* (2011). De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nat. Biotechnol.

30. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., *et al.* (2012). Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res.

31. Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., *et al.* (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature.

32. Zheng, Y., Wu, S., Bai, Y., Sun, H., Jiao, C., Guo, S., Zhao, K., Blanca, J., Zhang, Z., Huang, S., *et al.* (2019). Cucurbit Genomics Database (CuGenDB): A central portal for comparative and functional genomics of cucurbit crops. Nucleic Acids Res.

33. Proost, S., Bel, M. Van, Vaneechoutte, D., Van De Peer, Y., Inzé, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: An access point for plant comparative genomics. Nucleic Acids Res.

34. Gonzales, M.D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., Beavis, W.D., and Waugh, M.E. (2005). The Legume Information System (LIS): An

integrated information resource for comparative legume biology. Nucleic Acids Res.

35. Sato, S., Hirakawa, H., Isobe, S., Fukai, E., Watanabe, A., Kato, M., Kawashima, K., Minami, C., Muraki, A., Nakazaki, N., *et al.* (2011). Sequence analysis of the genome of an oil-bearing tree, jatropha curcas L. DNA Res.

36. Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., *et al.* (2008). Genome structure of the legume, Lotus japonicus. DNA Res.

37. Hiroyasu, K., Feng, L., Hideki, H., Takahiro, K., Zhongwei, Z., Yoichi, H., Kaoru, T., Sachiko, S., Aki, F., Shuji, Y., *et al.* (2014). Draft Sequences of the Radish (Raphanus sativus L.) Genome. DNA Res.

38. Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L.A., and Martin, G.B. (2012). A draft genome sequence of Nicotiana benthamiana to enhance molecular plant-microbe biology research. Mol. Plant-Microbe Interact.

39. van Bakel, H., Stout, J.M., Cote, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R., and Page, J.E. (2011). The draft genome and transcriptome of Cannabis sativa. Genome Biol.

40. Singh, R., Ong-Abdullah, M., Low, E.T.L., Manaf, M.A.A., Rosli, R., Nookiah, R., Ooi, L.C.L., Ooi, S.E., Chan, K.L., Halim, M.A., *et al.* (2013). Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature.

41. Ming, R., VanBuren, R., Liu, Y., Yang, M., Han, Y., Li, L.T., Zhang, Q., Kim, M.J., Schatz, M.C., Campbell, M., *et al.* (2013). Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). Genome Biol.

42. Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., Yang, M., He, L., Deng, T., Escalante, F.J., *et al.* (2017). Genome of wild olive and the evolution of oil biosynthesis.

Proc. Natl. Acad. Sci. U. S. A.

43. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res.

44. Guindon, S., and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst. Biol.

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

46. Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol.

47. Wright, E.S. (2016). Using DECIPHER v2.0 to analyze big biological sequence data in R. R J.

48. Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics.

FIGURE LEGENDS

Fig 1. Domain structures of AtCBP60 proteins. The amino acid lengths between the CBP60-conserved domain and the C-CaMBD in AtCBP60a-f vary from 133 to 261: in this figure, the length for AtCBP60b was used.

Fig 2. The immune-related clade diversified at or immediately before the Gymnosperm divergence and has been evolving very fast. The phylogenetic tree of 1024 CBP60 protein

sequences from 247 land plant species was inferred using RaXML [48] based on ClustalW alignment [43] and visualized using iTOL [45]. Colored dot at the leaf of the tree represents the land plant group (from Liverwort to Angiosperm) the CBP60 member originated from according to the color code for the phylogenetic clade. The inner ring with the yellow (low) to red (high) color range shows the CaM-binding score predicted by CaMELS [13]. The middle ring shows the CBP60 subfamilies according to AtCBP60 names. The outer ring shows the prototypical group (salmon) and the immune-related clade (blue). The AtCBP60 leaf positions are indicated in the outermost layer. The branch length scale of 0.1 (substitution per site) is shown at the top left.

Fig 3. Loss of calmodulin-binding ability is caused by loss of the C-CaMBD region or mutations in C-CaMBD. (A) The distribution of the length of the region C-terminal to the CBP60-conserved domain in each subfamily in Angiosperms. (B) Contributions of the length of alpha-helix and its hydrophobic moment to the CaM-binding prediction score according to the CaMELS algorithm [13].

Fig 4. The CBP60-conserved domains of the immune regulator subfamilies are highly diverse among 12 diverse Core Eudicot species. The mean pairwise PEAES distance for each site across the species is shown as a line plot in each panel. One CBP60 member per subfamily was selected for each species. Highly conserved sites (pairwise distance mean ≤ 0.0764) are shown as segments at the bottom in each plot. The segments in black are the sites highly conserved among all CBP60 members of the Core Eudicot species. The subfamily-specific sites are shown in orange, salmon, cyan, and green for the CBP60bcd, CBP60a, CBP60g, and SARD1 subfamilies,

respectively. Short segments represent subfamily-conserved gaps in the alignment. Highly diverse sites (pairwise distance mean > 0.4417) are shown at the top in each plot as segments colored according to the subfamily-color code.

Fig 5. Pairwise distance ranks (PDRs) based on the PEAES distance metric. First, for each of the 296 amino acid sites in the multiple sequence alignment of the CBP60-conserved domain, a pairwise PEAES-distance matrix was made between members of two subfamilies (CBP60g and SARD1 in the figure as an example) in the 12 Core Eudicot species. Second, in the pairwise PEAES-distance matrix at a particular site (Site #*x*), for each species (Carrot in the figure as an example) the pairwise distance between the members of the two subfamilies of the species (red square in the matrix) was compared to the pairwise distance between the member of one subfamily of this species (Carrot) and the member of the other subfamily of a different species (pink squares). The rank of the within-species comparison was determined among these 23 pairwise distances. This rank (21.0 for Carrot) is called the pairwise distance rank (PDR) at each site in each species.

Fig 6. Non-random coevolutionary interactions that could protect the positive immune regulators are prevalent among the 12 Core Eudicot lineages. (A) Four quadrants in the SARD1:CBP60a PDR vs. CBP60g:SARD1 PDR plot (SARD1-common comparison) as an example. The color codes for the quadrants shown are used in Fig 7. (B) Plots for CBP60g:SARD1 PDR vs. CBP60a:CBP60g PDR (CBP60g common comparison; cyan color), SARD1:CBP60a PDR vs. CBP60g:SARD1 PDR (SARD1-common comparison; green color), and CBP60a:CBP60g PDR vs. SARD1:CBP60a PDR (CBP60a-common comparison; salmon color) are shown for three

plant species, Poplar, Soybean, and Peach. Similar plots for all 12 Core Eudicot species are shown in Fig S7. The *p*-value (Fisher's exact test, 2-sided, Benjamini-Hochberg FDR-corrected) is shown in the bottom right. When $p < 0.01$ or $0.01 \leq p < 0.05$, the PC1 (principal component 1) axis is shown as a black solid line or a dashed line, respectively, indicating significant enrichment of sites in the 1st and 3rd quadrants over the 2nd and 4th quadrants. (C) The *p*-values for CBP60g-common, SARD1-common, and CBP60a-common comparisons are shown for the 12 Core Eudicot species. Red background, $p < 0.01$; yellow background, $0.01 \leq p < 0.05$.

Fig 7. Site selections are largely specific to each of the Core Eudicot lineages. The sites in CBP60g-common and SARD1-common comparisons in each plant species (Figs S7A and S7B) are shown along the CBP60-conserved domain with the color-scaled quadrant information (Fig 6A).

SUPPLEMENTAL INFORMATION

Text S1. Examples of CBP60 homologs that were not included in this study

Text S2. Detailed methods and the features of the CBP60 protein sequence phylogeny that are not discussed in the main text.

Fig S1. A pdf version of the CBP60 phylogenetic tree in Fig 2. This includes the name of each CBP60 member, containing the species name and the sequence ID in the originating database.

Fig S2. The three immune regulator CBP60 subfamilies diversified at or immediately before the emergence of Angiosperms. Two tree topology models, diversification of the CBP60g and SARD1 subfamilies at or before divergence of Basal Angiosperms (A) or after (B), were tested with CBP60 members from Basal Angiosperms, Magnoliids, and Chloranthales. Tree structure (A) was favored by SH-test ($p < 0.01$) [48]. Liverwort 0018s0017 was included as an outgroup. (C) SARD1 sequences were identified in recently published waterlily genome sequences [12]. The CBP60 sequence set used in (A) and (B) were supplemented with CBP60 sequences from *Nuphar advena* and *Nymphaea colorata* for the sequence phylogenetic tree. The *N. advena* and *N. colorata* SARD1 sequences are emphasized by a green dotted circle.

Fig S3. Each of 12 selected Core Eudicot species represents a different taxonomic order. The phylogeny of the Core Eudicot taxonomic orders and the 12 selected species representing diverse taxonomic orders are shown. These 12 Core Eudicot species and their lineages were used subsequently.

Fig S4. Difficulty in alignment of the immune regulator subfamilies outside the CBP60-conserved domain. Multiple sequence alignment of the regions N-terminal (A) and C-terminal (B) to the CBP60-conserved domains of CBP60a, CBP60g, SARD1, and CBP60bcd subfamily members of the 12 Core Eudicot species. The sequences from the top are 12 sequences each of the CBP60a, CBP60g, SARD1, and CBP60bcd subfamilies. Within each subfamily, the species order is the same as in Fig S3. The figures were made from screenshots of MEGA 7 alignments. The alignment was made by ClustalW in MEGA7 with the default parameter set values. No manual edits were made.

Fig S5. Multiple sequence alignment of the CBP60-conserved domains from CBP60a, CBP60g, SARD1, and CBP60bcd subfamily members of the 12 Core Eudicot species. One member per subfamily per species was selected (48 sequences). The alignment was divided into (A) sites 1-153 and (B) 154-296. The sequences from the top are 12 sequences each of the CBP60a, CBP60g, SARD1, and CBP60bcd subfamilies. Within each subfamily, the species order is the same as in Fig S3. The figures were made from screenshots of MEGA 7 alignments. The alignment FASTA file is provided as the supplemental file "CBP60.4subfam.12species.fas". To make the alignment compact, the alignment was manually edited with the following amino acids deleted. The amino acid position number is counted from the first amino acid of the CBP60-conserved domain of each sequence. SARD1_Quinoa_62033803, #N261, #G210-#K214, #Y99; CBP60g_Coffee_02g34690, #K206; CBP60bcd_Sunflower_5g00005047, #N205-#K210; SARD1_Tomato_03g119250.2.1, #I219, #T66; SARD1_Coffee_04g06970, #A131-#A132; SARD1_Peach_5G223600, #A137, #S48-#D49, #N39-#Q42; CBP60g_Arabidopsis_AT5G26920.1, #T100; SARD1_Poplar_015G045300.1, #C91; CBP60g_Monkeyflower_B00990, #Q63; SARD1_Sunflower_16g00022942, #V60; SARD1_Soybean_09G182400, #D65; SARD1_Orange_1g011961m, #I40

Fig S6. The analysis shown in Fig 4 was conducted with the alternative set. The alternative set of CBP60 members was made by selecting the member with the longest branch length, i.e., the most diverse member, among the paralogs of the species when the paralogs exist. This selection scheme was opposite from the set used in the main text. Therefore, the alternative set has all CBP60 members different from the set used for the main text where multiple members per

subfamily per species were available. The alignment FASTA file is provided as the supplemental file "CBP60.4subfam.12species.set2.fas". To make the alignment compact, the alignment was manually edited with the following amino acids deleted. CBP60g_Coffee_02g34690, #K206; SARD1_Quinoa_62029901, #N258, #E210-#K211, #C99; CBP60bcd_Sunflower_5g00005047, #N205-#K210; CBP60g_Soybean_03G232400, #L161-#K162; SARD1_Tomato_12g036390.1.1, #I127, #Y26; SARD1_Monkeyflower_I00405, #E130; SARD1_Coffee_04g06970, #A131-A132; SARD1_Peach_5G223600, #A137-#Y138, #I46-#A47, #N39-#Q42; SARD1_Arabidopsis_AT1G73805.1, #G126; CBP60g_Arabidopsis_AT5G26920.1, #T100; SARD1_Poplar_012G054900.1, #C91; SARD1_Soybean_08G044400, #Y71-#S72, #A63-#P64, #V40-#A41; SARD1_Orange_1g011961m, #I40. The amino acid position number is counted from the first amino acid of the CBP60-conserved domain of each sequence. With the alternative set: the means of the diversity measure across the sites are 0.080, 0.187, 0.301, and 0.239; the numbers of the highly diverse sites are 5, 34, 89, and 63; the numbers of the subfamily-specific highly conserved sites are 132, 54, 9, and 32; for CBP60bcd, CBP60a, CBP60g, and SARD1, respectively.

Fig S7. Non-random coevolutionary interactions that could protect the positive immune regulators are prevalent among the 12 Core Eudicot lineages. (A) CBP60g:SARD1 PDR vs. CBP60a:CBP60g PDR (CBP60g common comparison) plot for each species. (B) SARD1:CBP60a PDR vs. CBP60g:SARD1 PDR (SARD1-common comparison) plot for each species. (C) CBP60a:CBP60g PDR vs. SARD1:CBP60a PDR (CBP60a-common comparison) plot for each species. In each plot, the *p*-value (Fisher's exact test, 2-sided, Benjamini-Hochberg

FDR-corrected) is shown in the bottom right. When $p < 0.01$ or $0.01 \leq p < 0.05$, the PC1 (principal component 1) axis is shown as a black solid line or a dashed line, respectively, indicating significant enrichment of sites in the 1st and 3rd quadrants over the 2nd and 4th quadrants. Significant enrichment in the 1st and 3rd quadrants (BH-FDR = 0.05) was observed with 10 out of 12 lineages for the CBP60g-common comparison and with 7 out of 12 lineages for the SARD1-common comparison (the union of CBP60g- and SARD1-common comparisons with significant enrichment are 11 out of 12), while only one lineage with moderately significant enrichment out of 12 lineages was observed for the CBP60a-common comparison.

Fig S8. The analysis shown in Fig S7 was conducted with the alternative set. Significant enrichment in the 1st and 3rd quadrants (BH-FDR = 0.05) was observed with 8 out of 12 lineages for the CBP60g-common comparison and with 6 out of 12 lineages for the SARD1-common comparison (the union of CBP60g- and SARD1-common comparisons with significant enrichment are 9 out of 12), while only one lineage with moderately significant enrichment out of 12 lineages was observed for the CBP60a-common comparison.

Fig S9. The analysis shown in Fig 7 was conducted with the alternative set.

Fig S10. Coevolutionary selections observed in the immune regulator subfamilies are largely lineage-specific. (A) The site values projected onto $y = x$ (red line) and centered at $(x, y) = (12,12)$ as a measure of enrichment in the 1st and 3rd quadrants in a plot. Two example data points (blue) and their projected values as arrows (their sizes; green for negative and orange for positive values) are shown. (B-D) The projected values in each plot in Fig S7 were plotted

between all pairs of 12 Core Eudicot species (B-D, respectively). There is no evident tendency that more closely related species have higher correlations in selected sites, which indicates that selections at the sites largely occurred after divergence of the 12 species lineages.

Fig S11. The analysis shown in Figs S10B-S10D was conducted with the projected values from each plot in Fig S8 for the alternative set.

Table S1. The list of 271 land plant species used in the study, including the nomenclature used in the text ("Database name").

Table S2. The list of 1024 QC-ed CBP60 sequences ("Final set 1024" tab) and the list of 409 QC-omitted sequences ("QC fail" tab).

Table S3. The six-dimensional coordinates for amino acids in PEAES. The columns, E1-E6, represent the six dimensions in the Euclidean space.

Table S4. The pairwise Euclidean distance between amino acids in PEAES.

Supplemental files:

"CBP60.final.tree.nwk": The Newick format tree file for the CBP60 sequence phylogenetic tree in Figs 2 and S1.

"CBP60.4subfam.12species.fas": The FASTA multiple sequence alignment file for Fig S5.

“CBP60.4subfam.12species.set2.fas”: The FASTA multiple sequence alignment file for Fig S6 (“the alternative set”).

“peaes.pdr.figures.r”: an R script, which includes an algorithm for PEAES-PDR analysis of “CBP60.4subfam.12species.fas” and generates Figs 4, 7, and S7.

“aa.dist.mat.RData”: R object that is an equivalent of Table S4; required in “peaes.pdr.figures.r”