



# Concerted expansion and contraction of immune receptor gene repertoires in plant genomes

Received: 4 January 2022

Accepted: 9 September 2022

Published online: 14 October 2022

 Check for updatesBruno Pok Man Ngou<sup>1,2,4</sup>, Robert Heal<sup>1,4</sup>, Michele Wyler<sup>3</sup>, Marc W. Schmid<sup>3</sup> and Jonathan D. G. Jones<sup>1</sup>  

Recent reports suggest that cell-surface and intracellular immune receptors function synergistically to activate robust defence against pathogens, but whether they co-evolve is unclear. Here we determined the numbers of cell-surface and intracellular immune receptors in 350 species. Surprisingly, the number of receptor genes that are predicted to encode cell-surface and intracellular immune receptors is strongly correlated. We suggest this is consistent with mutual potentiation of immunity initiated by cell-surface and intracellular receptors being reflected in the concerted co-evolution of the size of their repertoires across plant species.

Plants have evolved a two-tier immune system that recognizes and activates defence against pathogens<sup>1,2</sup>. Cell-surface pattern-recognition receptors (PRRs) recognize apoplastic and usually conserved pathogen-associated molecular patterns (PAMPs) and activate pattern-triggered immunity (PTI). Virulent pathogens secrete effector molecules into plant cells that suppress PTI and promote infection. Intracellular nucleotide-binding leucine-rich repeat (NLR) receptors recognize effectors and activate effector-triggered immunity (ETI). Although PTI and ETI were envisaged as two independent immune systems<sup>1</sup>, emerging evidence suggests they are inter-dependent and share multiple signalling components<sup>3–6</sup>. Thus, PTI and ETI function synergistically to provide robust immunity against pathogens. As PRRs and NLRs are functionally inter-dependent, in this Brief Communication, we investigated whether the sizes of these two receptor gene families are correlated.

Plant PRR proteins are structurally diverse but are usually receptor-like kinases (RLKs) or receptor-like proteins (RLPs). RLKs carry extracellular ectodomains and cytosolic kinase domains, while RLPs lack cytosolic kinase domains. RLKs carry multiple types of extracellular domains, such as leucine-rich repeats (LRRs), lectins and lysM motifs (LysMs)<sup>7</sup>. LRR-domain-containing RLKs (LRR-RLKs) and RLPs (LRR-RLPs) are the largest RLK- and RLP-gene families in plants<sup>8,9</sup>. LRR-RLKs can be further classified into 20 subgroups, with each subgroup involved in different biological processes<sup>10</sup> (Extended Data Fig. 1). For example, BAK1 (BRI1-ASSOCIATED RECEPTOR KINASE)

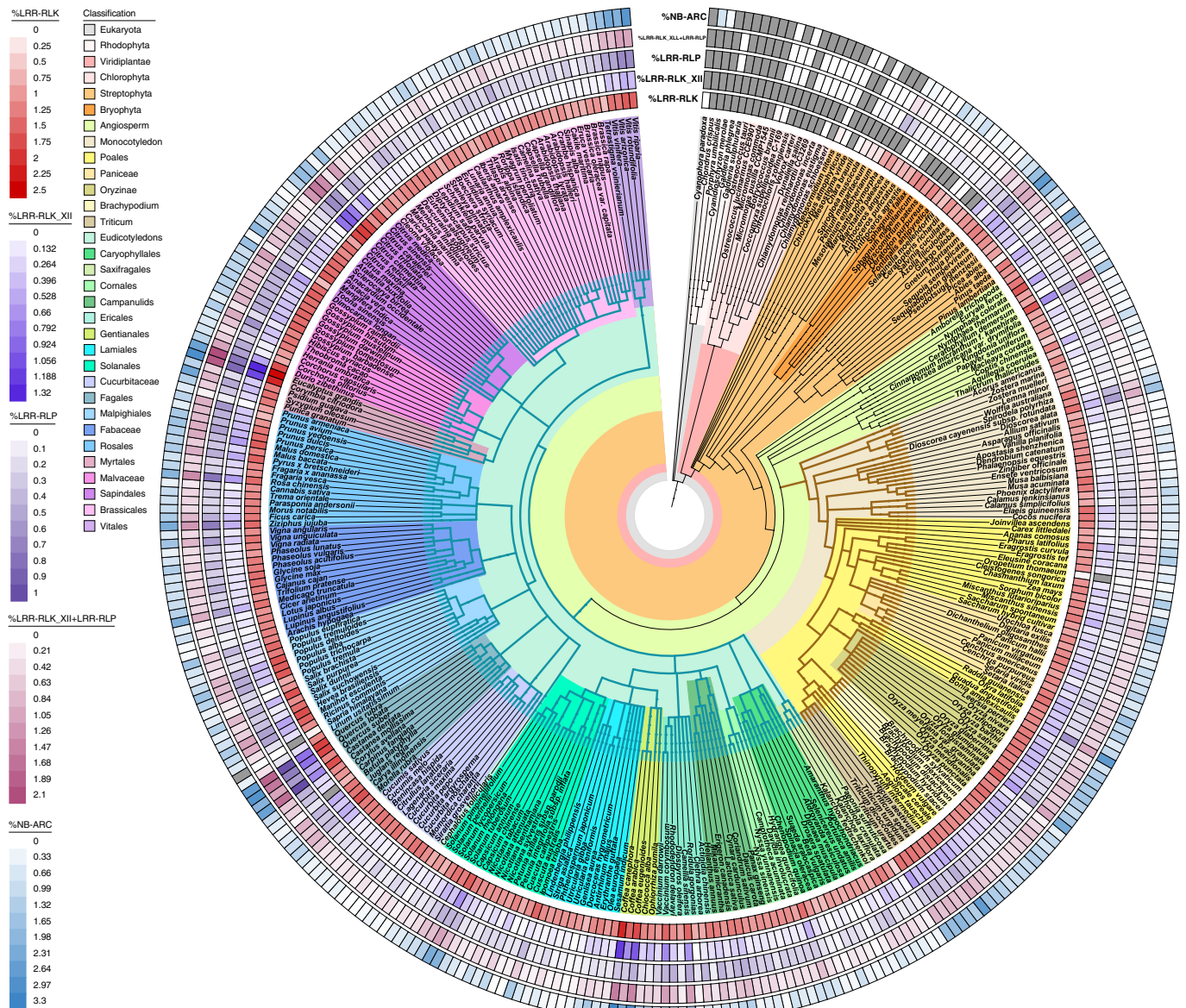
family proteins function as PRR co-receptors and belong to LRR-RLK-II (ref. <sup>11</sup>). Members of LRR-RLK-XI are involved in recognition of self-peptides<sup>12,13</sup>. Members of LRR-RLK-XII, such as FLAGELLIN-SENSITIVE 2 (FLS2), EF-TU RECEPTOR (EFR) and Xa21 (refs. <sup>2,7</sup>), are involved in detecting pathogen-derived molecules (Extended Data Fig. 1). NLRs are intracellular receptors that carry NB-ARC domains with C-terminal LRR domains and N-terminal domains, usually comprising either coiled-coil (CC), Toll/interleukin-1 receptor/resistance protein (TIR) or RPW8-like coiled-coil (RPW8) domains (hence, CC-NLRs (or CNLs), TIR-NLRs (TNLs) and RPW8-NLRs (RNLs))<sup>14,15</sup>.

## Identification of immune receptors from plant genomes

To investigate expansion or contraction of genes that encode PRR and NLR proteins, we identified these gene families in annotated proteomes from 350 publicly available genomes. These genomes include 26 algal species, 5 bryophyte species, 10 gymnosperms and 300 angiosperms (13 basal angiosperms, 79 monocots and 208 eudicots) (Extended Data Fig. 2 and Supplementary Table 1). Assembled genome sizes of these organisms range from 13 Mb to 27.6 Gb, with annotated protein counts ranging from ~5,000 to ~300,000 (Extended Data Fig. 2). To ensure consistency, we used the same pipeline to obtain primary transcripts and identify LRR-RLKs, LRR-RLPs, LysM-RLK, LysM-RLP and NB-ARCs from each of these genomes (Extended Data Fig. 3 and Methods).

<sup>1</sup>The Sainsbury Laboratory, University of East Anglia, Norwich Research Park, Norwich, UK. <sup>2</sup>RIKEN Center for Sustainable Resource Science, Yokohama, Japan. <sup>3</sup>MWSchmid GmbH, Glarus, Switzerland. <sup>4</sup>These authors contributed equally: Bruno Pok Man Ngou, Robert Heal.

✉ e-mail: [jonathan.jones@tsl.ac.uk](mailto:jonathan.jones@tsl.ac.uk)



**Fig. 1 | Immune receptor gene families in 350 genomes.** Phylogenetic tree of 350 plant species, including 300 angiosperms, 79 monocot species and 208 eudicot species. Heat maps represent the percentages (%) of LRR-RLKs, LRR-RLK\_XIIIs (red), LRR-RLPs (purple), LRR-RLK\_XIIIs + LRR-RLPs (magenta) and NB-ARCs

(blue) in their corresponding annotated proteomes. Grey boxes in heat maps indicate null values where no receptors were identified. Brown branches indicate monocots, and teal branches represent eudicots.

In total, we identified 88,020 LRR-RLKs, 28,018 LRR-RLPs, 3,500 LysM-RLKs, 1,238 LysM-RLPs and 95,127 NB-ARCs from 350 species (Supplementary Fig. 1 and Supplementary Table 2). To validate our results, we compared the number of NB-ARCs, LRR-RLKs and LRR-RLPs identified in our study with previous publications, finding they are highly similar (Extended Data Fig. 4)<sup>16–18</sup>. As expected, the number of receptors varies enormously across 300 angiosperms, with LRR-RLKs ranging from 16 to 1,129, LRR-RLPs ranging from 2 to 585, LysM-RLKs ranging from 0 to 42, LysM-RLPs ranging from 0 to 19 and NB-ARCs ranging from 0 to 3,128. To account for the effect of genome duplication and variable proteome sizes, we normalized these data using percentages (%) of LRR-RLKs, LRR-RLPs, LysM-RLK, LysM-RLP and NB-ARCs from each genome (number of identified genes/number of searched genes × 100) (Supplementary Fig. 1 and Supplementary Table 3). After adjustment, LRR-RLKs range from 0.114% to 2.464%, LRR-RLPs range from 0.00652% to 1.010%, LysM-RLKs range from 0% to 0.120%, LysM-RLPs range from 0% to 0.0559% and NB-ARCs range

from 0% to 3.266% in 300 angiosperms (Fig. 1, Supplementary Fig. 1 and Supplementary Table 3).

### Correlation between the sizes of immune receptor families

Next, we determined the correlation between the percentages of PRRs (%LRR-RLKs, %LRR-RLPs, %LysM-RLK and %LysM-RLP) and NB-ARCs in angiosperms. Surprisingly, %NB-ARC and %LRR-RLPs show a strong positive linear correlation (Pearson’s  $r = 0.759$ ), suggesting that NB-ARC and LRR-RLP gene families expand together (Fig. 2, Extended Data Figs. 5 and 6, Supplementary Fig. 7 and Supplementary Table 4). Similarly, %NB-ARC and %LRR-RLKs show a positive but weaker linear correlation (Pearson’s  $r = 0.657$ ). On the other hand, %LysM-RLKs and %LysM-RLPs show weak or no correlation with %NB-ARC (Pearson’s  $r = 0.216$  and  $-0.0430$ , respectively). We propose that PRRs involved in pathogen recognition are more likely to co-expand with NB-ARC gene families. This is consistent with the observation that characterized LRR-RLPs are

usually involved in pathogen recognition, while LRR-RLKs and LysM receptors can be involved not only in immunity but also in development, reproduction or establishing symbiosis<sup>12</sup>.

To test if the NB-ARC gene family co-expands with PRRs specifically involved in pathogen recognition, we further classified LRR-RLKs into subgroups according to their kinase domains. As mentioned, LRR-RLKs can thus be classified into 20 subgroups, with each subgroup involved in various biological processes<sup>10</sup> (Extended Data Fig. 1). Across 350 species, LRR-RLK-XII forms the largest LRR-RLK subgroup, followed by LRR-RLK-III and LRR-RLK-XI (Supplementary Figs. 2 and 3). We determined the correlation between %LRR-RLK from different subgroups and %NB-ARC in angiosperms (Extended Data Figs. 5 and 6, and Supplementary Table 4). Strikingly, only 4 out of 20 LRR-RLK subgroups show significant and positive linear correlation with %NB-ARCs (LRR-RLK-VIII\_1, LRR-RLK-VIII\_2, LRR-RLK-Xa and LRR-RLK-XII). Furthermore, LRR-RLK-XII forms much stronger positive correlation with %NB-ARCs (Pearson's  $r = 0.813$ ) compared with LRR-RLK-VIII\_1, LRR-RLK-VIII\_2 and LRR-RLK-Xa (Pearson's  $r = 0.343, 0.440$  and  $0.279$ , respectively) (Fig. 2a, Supplementary Fig. 7 and Supplementary Table 4). While LRR-RLKs involved in pathogen recognition are predominantly in subgroup XII, some members from LRR-RLK-VIII and LRR-RLK-Xa are also involved in immunity and pathogen recognition, such as CANNOT RESPOND TO DMBQ1 (CARD1/HPCA1), CELLOLIGOMER-RECEPTOR KINASE 1 (CORK1) and BAK1-INTERACTING RECEPTOR-LIKE KINASE 1 (BIR1) (refs. 19–22) (Extended Data Fig. 1). As LRR-RLK-XII forms the largest LRR-RLK subgroup, we tested if the positive correlation between %LRR-RLK (total) and %NB-ARC is predominantly caused by subgroup XII. Indeed, for all subgroups excluding XII, %LRR-RLK does not show any significant correlation with %NB-ARC (Pearson's  $r = 0.103$ ). On the other hand, %LRR-RLP combined with %LRR-RLK-XII show strong positive correlation with %NB-ARC (Pearson's  $r = 0.859$ ) (Fig. 2a,b).

We further tested the hypothesis taking into account the phylogeny of the plant species. First, we converted receptor percentages from 350 genomes into distance matrices and tested for correlation between receptor pairs with Mantel tests. %LRR-RLK-XII and %LRR-RLP show strong positive correlation with %NB-ARC (Extended Data Fig. 7a). Second, we obtained a phylogenetic tree from a previous publication<sup>23</sup> for 238 species and tested whether percentage receptor distances correlate with each other while taking into account the phylogenetic distances with a partial Mantel test. Again, %LRR-RLK-XII and %LRR-RLP show strong positive correlation with %NB-ARC (Extended Data Fig. 7b). Third, we tested for correlation between receptor percentages and phylogenetic distances directly. Whereas almost all %LRR-RLKs significantly correlate with the phylogeny, %LRR-RLK-XII, %LRR-RLPs and %NB-ARC do not (Extended Data Fig. 7c). Taken together, we conclude that PRR gene families specifically involved in pathogen recognition co-expand or co-contract with NB-ARC gene families.

## Expansion and contraction of immune receptor families

We observed a strong linear correlation between %NB-ARC and %LRR-RLK-XII and %LRR-RLP in angiosperms, monocots, eudicots and multiple plant clades (Extended Data Fig. 8). Next, we checked if NB-ARC gene family contraction coincides with PRR gene family contraction in organisms adapted to specific lifestyles, such as parasitism and carnivorous. The Alismatales and Lentibulariaceae lineages

show a reduction in the size of NB-ARC gene repertoires<sup>16</sup>, and species from these lineages also have low %PRRs (%LRR-RLP, %LRR-RLK-XII and %LysM-RLK). These include *Genislea aurea*, *Utricularia gibba*, *Utricularia reniformis*, *Zostera marina*, *Zostera muelleri*, *Lemna minor*, *Wolffia australiana* and *Spirodela polyrhiza* (Fig. 2c, Extended Data Fig. 9, Supplementary Table 3 and Supplementary Fig. 6). We infer that the %NB-ARC and %LRR-RLK correlation is not due just to co-expansion, but also co-contraction. Carnivorous, aquatic and parasitic plant genomes carry few NLRs<sup>16,24</sup>. We tested if the number of cell-surface immune receptors is also reduced in these plants. Compared with species that are not adapted to these lifestyles, carnivorous, aquatic and parasitic plant genomes have lower %NB-ARC, %LRR-RLK-XII and %LRR-RLP (Extended Data Fig. 9 and Supplementary Fig. 6). These include *Sapria himalayana*, *Cephalotus follicularis*, *Drosera spatulata*, *Dionaea muscipula* and *Aldrovanda vesiculosa*. Notably, %LRR-RLK (total) in these groups is similar to other plant species, as are most other LRR-RLK subgroups (Supplementary Fig. 6).

Some other species and genera also show lower %NB-ARC, %LRR-RLK and %LRR-RLP. For example, the Cucurbitaceae show far fewer immune receptors than the phylogenetically close Malpighiales or Fagales clades (Fig. 1, Supplementary Fig. 4 and Supplementary Table 3). Remarkably, in the monocot species *Oropetium thomaeum*, we observed only 0.0558% NB-ARC containing proteins and no LRR-RLK-XII. This contrasts with the other members of the Poales, where high %PRRs and %NB-ARCs are more frequent (Fig. 1 and Supplementary Table 3). *O. thomaeum* is an atypical member in the Poales. This drought-tolerant resurrection grass has the smallest known grass genome (245 Mb) and can survive losing 95% of cellular water<sup>25</sup>. Despite its small genome, *O. thomaeum* has a similar number of predicted proteins as other Poales species such as *Ananas comosus*, *Oryza longistaminata* and *Triticum urartu*, suggesting that the contraction of immune receptor families could be independent of the reduced genome size.

On the other hand, some plant groups show much larger immune receptor families. Many species of the order Poales show high %LRR-RLK, %LRR-RLP and %NB-ARC, most notably in the *Oryza* and *Triticum* genera (Fig. 1 and Supplementary Fig. 4). In addition, many tree species also show a high proportion of PRR and NB-ARC proteins in their proteomes. These include *Eucalyptus grandis*, *Castanea dentata*, *Corymbia citriodora*, *Quercus rubra*, *Quercus lobata*, *Coffea canephora*, *Prunus avium*, *Malus domestica*, *Theobroma cacao* and *Citrus* species (Fig. 2c, Extended Data Fig. 9 and Supplementary Table 3). Thus, some plant lifestyles might also correlate with expansion of immune receptor gene families.

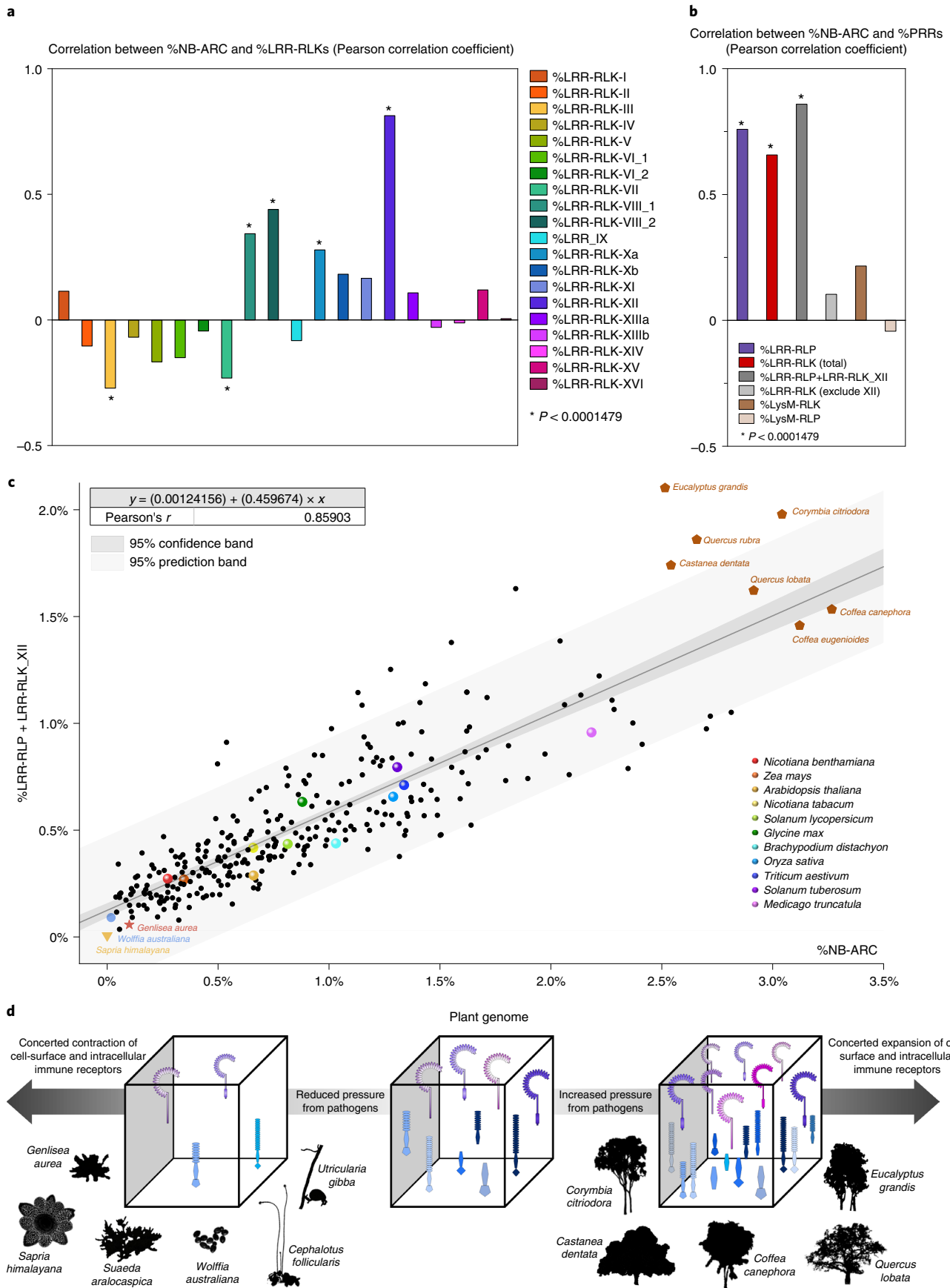
Previously, analysis of the *Solanum lycopersicum* genome has suggested that NLRs, RLPs and RLKs might form genomic clusters<sup>26</sup>. Genomic clustering could mean that expansion/contraction of a gene family could result in genes in close proximity indirectly expanding in tandem. To determine if concerted expansion/contraction of immune receptor families is due to genomic clustering, we investigated the *Solanum tuberosum*, *Zea mays* and *Oryza sativa* genomes. In all three genomes, many LRR-RLK-XII and LRR-RLP loci overlap with NB-ARC-encoding loci (Supplementary Fig. 7). To quantify this, we calculated the average distance of LRR-RLKs and LRR-RLPs to the closest NB-ARC encoding genes and compared with a distribution of randomly selected genes (for details, see Methods). Both LRR-RLK-XIIs and LRR-RLPs are located closer to NB-ARC genes than randomly selected genes. However, LRR-RLK-III and LRR-RLK-XI genes are also located

**Fig. 2 | Concerted expansion and contraction of cell-surface and intracellular immune receptor genes in plant genomes. a, b.** Correlation between %NB-ARC and %LRR-RLKs and %PRRs in 300 angiosperms. Bar chart represents the Pearson correlation coefficient, with significant values indicated with asterisks. Two-sided test of significance was performed, and Bonferroni correction was performed to adjust the  $P$  value for the multiple independent tests performed (Extended Data Fig. 6). Summary of statistical analyses and individual  $P$  values are provided in Supplementary Table 4. **c.** Scatter plot of %LRR-RLP + LRR-RLK-XII

against %NB-ARC. Black line represents the linear trend, with dark-grey shade representing the 95% confidence interval and light-grey shade representing the 95% prediction interval. Several parasitic species, carnivorous species, aquatic species and trees are indicated as yellow inverted triangles, orange stars, blue circles and brown pentagons, respectively. Model organisms are also indicated as spheres of different colours. **d.** Schematic illustration of the co-expansion and co-contraction of immune receptors in plant genomes.

nearby to NB-ARC genes (Extended Data Fig. 10). As %LRR-RLK-III and %LRR-RLK-XI do not show positive correlation with %NB-ARC, we conclude that, while NB-ARC-encoding genes can form genomic

clusters adjacent to LRR-RLK-XIIs, the co-expansion/contraction of these immune receptors is likely to be caused by mechanisms other than genomic clustering.



## Discussion

Previously it was shown that cell-surface and intracellular immune systems exhibit mutual potentiation and inter-dependency<sup>3–6</sup>. Here we show that, in addition to their functional relationship, there is also an evolutionary correlation between the numbers of cell-surface and intracellular immune receptors. Expansion and/or contraction of intracellular NLRs coincides with expansion and/or contraction of cell-surface PRRs involved in pathogen recognition (Fig. 2d). These observations are consistent with previous reports<sup>16,24,27,28</sup>.

We propose that pathogen pressure shapes the immune receptor diversity and repertoire, which, as a result, is determined by plant lifestyles and their ecological niches. We observed high %PRR and %NB-ARC in many *Oryza* and *Triticum* species. Grasses typically grow in high densities and are frequently challenged by rust and blast species that produce numerous, wind-dispersed spores with high genetic diversity. Genetic exchange by sexual reproduction and somatic hybridization drives the emergence of new virulent strains<sup>29</sup>, such as the *Ug99* strain of the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici*<sup>30</sup>. An expanded repertoire of immune receptors and increased heterogeneity in populations could be a result of high pressure from these pathogens. Conversely, it has been proposed that the reduced root system in parasitic and carnivorous plants results in fewer interactions or entry routes for pathogens<sup>24</sup>. Similarly, partial or complete submersion of aquatic species results in reduced exposure to airborne pathogenic spores, removing an interface for interaction with pathogens. Lifespan may also drive changes in the immune receptor repertoire. We found that trees generally show higher %PRR and %NB-ARC than other species. While annual plants are subject to shorter periods of pathogen pressure before reproduction, biennial or perennial plants, especially trees, must survive for much longer. Conceivably, this long-term pathogen pressure could drive the expansion of immune receptor gene families.

As parasite pressure drives the retention of sexual reproduction that reshuffles immune receptor alleles each generation<sup>31</sup>, inbreeding species may require an increased number of immune receptors compared with their outbreeding ancestors, an outcome that can also result from polyploidy. As the concerted expansion and contraction of immune receptors in plant genomes is not due to genomic clustering, further study is needed to understand the mechanism(s) underpinning these observations. As functionally inter-dependent genes often co-expand/contract together, it is likely that the functional relationship between cell-surface and intracellular immune receptors is conserved across plant species.

## Methods

### LRR-RLK identification

Protein sequences from all 350 plant proteomes were first filtered for the primary gene model. Sequences shorter than 250 amino acids (AA) were removed as they are unlikely LRR-RLKs. The remaining proteins were searched for the presence of a protein kinase domain (PFAM PF00069.26) and an LRR domain (PFAM PF18805.2, PF18831.2, PF18837.2, PF00560.34, PF07723.14, PF07725.13, PF12799.8, PF13306.7, PF13516.7, PF13855.7, PF14580.7, PF01463.25, PF08263.13 and PF01462.19) with hmmer (version 3.1b2, options -E 1e-10 for the kinase domain and -E 10e-3 for the LRR domains<sup>32</sup>). The *Arabidopsis* sequences that were previously classified into 20 LRR-RLK subgroups<sup>10</sup> were filtered likewise for the presence of LRR and kinase domains. Eleven sequences were removed because they did not pass the threshold filter for the kinase (two sequences) and LRR (nine sequences) domain searches. To classify all candidate sequences according to the *Arabidopsis* subgroups, the highest-scoring kinase domain region of each candidate was extracted and aligned to the *Arabidopsis* reference sequences using diamond<sup>33</sup> (version 0.9.26, options -e 1e-10 -k 300).

### Phylogeny

The phylogeny of each subgroup was inferred using the kinase domains. Sequences were aligned with FAMSA<sup>34</sup>. Alignments were not trimmed<sup>35</sup>

and phylogenetic trees were inferred with FastTree<sup>36</sup> (version 2.1.11 SSE3, option -lg). Trees were rooted with gotree<sup>37</sup> (v0.4.2) using the sequences belonging to the most basal species as outgroup (according to the taxonomic tree from National Center for Biotechnology Information (NCBI)).

### LRR-RLP identification

LRR-RLPs were identified similarly but filtering for proteins of a minimal length of 150 AA first. Proteins were then searched for the presence of LRR domains and the absence of a kinase domain (hmmer options as above), as well as the presence of a C3F domain (hmmer option -E 1e-10 and a minimal alignment length of 140). The hmmer profile for the C3F domain was obtained from a multiple alignment of *Arabidopsis* LRR-RLPs<sup>38</sup>. The domain was trimmed manually, starting from the conserved Y in the C2 domain (Fig. 6b in ref. <sup>39</sup>). Candidates were finally filtered for the presence of a transmembrane domain using tmhmm<sup>40</sup> with default settings (version 2.0).

### NB-ARC identification

NB-ARCs were identified using the set of proteins with a minimal length of 150 AA. Proteins were then searched for the presence of NB-ARC (PF00931.23) domains (hmmer option -E 1e-10 for NB-ARC).

### LysM identification

LysM-RLKs and LysM-RLPs were identified using the set of proteins with a minimal length of 150 AA. Proteins were filtered for the presence of a LysM domain (PF01476.21, hmmer option -max -E 1000 -incE 1000 -indomE 1000) and a transmembrane domain<sup>40</sup> (tmhmm, version 2.0). Candidates were split into LysM-RLKs and LysM-RLPs by searching for presence/absence of a kinase domain (PF00069.26, as above).

### Test for co-occurrence of NB-ARC, LRR-RLKs and LRR-RLPs

To test whether two gene groups are closer to each other than expected by chance, we used a test based on random sampling, for example, group A (LRR-RLK-XII) with  $n$  and group B (NB-ARCs) with  $m$  genes. The observed distance was calculated as the average closest distance between genes in group A and genes in group B. A distribution for the expected distance was obtained by randomly sampling  $m$  genes and calculating the average closest distance of genes in group A to the genes in the random set (1,000 times). Genes were sampled from the list of genes that was used to search for the genes in group B (Supplementary Fig. 1).

### Taxonomic tree

The taxonomic tree was obtained from NCBI (<https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi>). Phylogenetic tree of the 350 species is generated by phyloT (<https://phyloT.biobyte.de/>) based on NCBI taxonomy database. Phylogenetics trees were visualized and figures were generated by iTOL<sup>41</sup>. The tree used for testing the relationship between the fraction of candidates found and the phylogenetic distances were obtained from ref. <sup>23</sup>. The latter contained 238 out of the 350 genomes analysed.

### Test for similarities in fraction of proteins and phylogenetic relationship

To test whether the fraction of certain proteins (for example, NB-ARCs) found per species correlated with the phylogenetic relationships, we converted the fractions and the phylogenetic tree to distance matrices and tested for correlation with mantel tests (R package vegan, version 2.5-7 with 10,000 permutations). Analogously, we also tested for correlation between distance matrices obtained for two different sets of proteins (for example, LRR-RLK-XII and NB-ARCs).  $P$  values were corrected for multiple testing to reflect false discovery rates<sup>42</sup>.

### Statistical analyses

Statistical analyses were performed with OriginPro (version 2022; <https://www.originlab.com/>) and R (version 3.4.4).

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All data generated or analysed during this study are included in the article or supplementary information files. Proteomes of 350 species used in this study are downloaded from either NCBI, Phytozome13, ensemblplants, JGI, Fernbase, Penium Genome Database or the publications directly. A complete list of the proteomes and associated data used in this study are provided in Supplementary Table 1. Sequences of the identified receptors and phylogenetic analyses are available on Zenodo<sup>43</sup>.

## Code availability

All the analyses were done as described in Methods with publicly available tools (hmmer, tmhmm, diamond, FastTree, FAMSA, gotree and R). Scripts are available on [github.com/MWSchmid/Ngou-et-al.-2022](https://github.com/MWSchmid/Ngou-et-al.-2022).

## References

- Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444**, 323–329 (2006).
- Ngou, B. P. M., Ding, P. & Jones, J. D. G. Thirty years of resistance: zig-zag through the plant immune system. *Plant Cell* **34**, 1447–1478 (2022).
- Ngou, B. P. M., Ahn, H.-K., Ding, P. & Jones, J. D. G. Mutual potentiation of plant immunity by cell-surface and intracellular receptors. *Nature* **592**, 110–115 (2021).
- Yuan, M. et al. Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature* **592**, 105–109 (2021).
- Pruitt, R. N. et al. The EDS1-PAD4-ADR1 node mediates *Arabidopsis* pattern-triggered immunity. *Nature* **598**, 495–499 (2021).
- Tian, H. et al. Activation of TIR signalling boosts pattern-triggered immunity. *Nature* **598**, 500–503 (2021).
- Boutrot, F. & Zipfel, C. Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. *Annu. Rev. Phytopathol.* **55**, 257–286 (2017).
- Shiu, S. H. & Bleeker, A. B. Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in *Arabidopsis*. *Plant Physiol.* **132**, 530–543 (2003).
- Diebart, A., Gottin, C., Périn, C., Ranwez, V. & Chantret, N. Origin and diversity of plant receptor-like kinases. *Annu. Rev. Plant Biol.* **71**, 131–156 (2020).
- Lehti-Shiu, M. D., Zou, C., Hanada, K. & Shiu, S.-H. Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiol.* **150**, 12–26 (2009).
- Ma, X., Xu, G., He, P. & Shan, L. SERKing coreceptors for receptors. *Trends Plant Sci.* **21**, 1017–1033 (2016).
- Jose, J., Ghantasala, S. & Roy Choudhury, S. *Arabidopsis* transmembrane receptor-like kinases (RLKs): a bridge between extracellular signal and intracellular regulatory machinery. *Int. J. Mol. Sci.* **21**, 4000 (2020).
- Furumizu, C. et al. The sequenced genomes of nonflowering land plants reveal the innovative evolutionary history of peptide signaling. *Plant Cell* **33**, 2915–2934 (2021).
- Jones, J. D. G., Vance, R. E. & Dangl, J. L. Intracellular innate immune surveillance devices in plants and animals. *Science* **354**, aaf6395 (2016).
- Shao, Z.-Q. et al. Long-term evolution of nucleotide-binding site-leucine-rich repeat genes: understanding gained from and beyond the legume family. *Plant Physiol.* **166**, 217–234 (2014).
- Baggs, E. L. et al. Convergent loss of an EDS1/PAD4 signaling pathway in several plant lineages reveals coevolved components of plant immunity and drought response. *Plant Cell* **32**, 2158–2177 (2020).
- Dufayard, J.-F. et al. New insights on leucine-rich repeats receptor-like kinase orthologous relationships in angiosperms. *Front. Plant Sci.* **8**, 381 (2017).
- Li, P. et al. RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 852 (2016).
- Laohavisit, A. et al. Quinone perception in plants via leucine-rich-repeat receptor-like kinases. *Nature* **587**, 92–97 (2020).
- Wu, F. et al. Hydrogen peroxide sensor HPCA1 is an LRR receptor kinase in *Arabidopsis*. *Nature* **578**, 577–581 (2020).
- Tseng, Y.-H. et al. CORK1, A LRR-malectin receptor kinase, is required for cellooligomer-induced responses in *Arabidopsis thaliana*. *Cells* **11**, 2960 (2022).
- Liu, Y., Huang, X., Li, M., He, P. & Zhang, Y. Loss-of-function of *Arabidopsis* receptor-like kinase BIR1 activates cell death and defense responses mediated by BAK1 and SOBIR1. *N. Phytol.* **212**, 637–645 (2016).
- Qian, H. & Jin, Y. An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. *J. Plant Ecol.* **9**, 233–239 (2016).
- Liu, Y. et al. An angiosperm NLR Atlas reveals that NLR gene reduction is associated with ecological specialization and signal transduction component deletion. *Mol. Plant* **14**, 2015–2031 (2021).
- VanBuren, R. et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**, 508–511 (2015).
- Andolfo, G. et al. Overview of tomato (*Solanum lycopersicum*) candidate pathogen recognition genes reveals important *Solanum* R locus dynamics. *N. Phytol.* **197**, 223–237 (2013).
- Tirnaz, S. et al. Resistance gene analogs in the Brassicaceae: identification, characterization, distribution, and evolution. *Plant Physiol.* **184**, 909–922 (2020).
- Plomion, C. et al. Oak genome reveals facets of long lifespan. *Nat. Plants* **4**, 440–452 (2018).
- Stukenbrock, E. H. The role of hybridization in the evolution and emergence of new fungal plant pathogens. *Phytopathology* **106**, 104–112 (2016).
- Li, F. et al. Emergence of the Ug99 lineage of the wheat stem rust pathogen through somatic hybridisation. *Nat. Commun.* **10**, 5068 (2019).
- Hamilton, W. D., Axelrod, R. & Tanese, R. Sexual reproduction as an adaptation to resist parasites (a review). *Proc. Natl Acad. Sci. USA* **87**, 3566–3573 (1990).
- Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
- Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
- Deorowicz, S., Debudaj-Grabysz, A. & Gudyś, A. FAMSA: fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.* **6**, 33964 (2016).
- Tan, G. et al. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst. Biol.* **64**, 778–791 (2015).
- Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
- Lemoine, F. & Gascuel, O. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genom. Bioinform.* **3**, lqab075 (2021).
- Steidele, C. E. & Stam, R. Multi-omics approach highlights differences between RLP classes in *Arabidopsis thaliana*. *BMC Genomics* **22**, 557 (2021).

39. Fritz-Laylin, L. K., Krishnamurthy, N., Tör, M., Sjölander, K. V. & Jones, J. D. G. Phylogenomic analysis of the receptor-like proteins of rice and *Arabidopsis*. *Plant Physiol.* **138**, 611–623 (2005).
40. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
41. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
42. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
43. Ngou, B. P. M., Heal, R., Wyler, M., Schmid, M. W. & Jones, J. D. Genome-wide identification of cell-surface and intracellular immune receptors in 350 plant species. *Zenodo* <https://doi.org/10.5281/zenodo.7017981> (2022).

## Acknowledgements

We thank S. Fairhead, J. Rhodes, D. Weigel, A.-L. Laine, M. Moscou, Y. Tanaka and Y. Kadota for discussions and suggestions. R.H. and J.D.G.J. are supported by the Gatsby Foundation. B.P.M.N. was supported by the Norwich Research Park Biosciences Doctoral Training Partnership from the Biotechnology and Biological Sciences Research Council (BBSRC) (grant agreement BB/M011216/1).

## Author contributions

B.P.M.N., R.H. and J.D.G.J. conceived and conceptualized the study; B.P.M.N., M.W.S. and M.W. designed and performed the bioinformatic analyses; B.P.M.N., M.W. and M.W.S. performed the statistical analyses; B.P.M.N. and R.H. wrote the original draft; and B.P.M.N., R.H., M.W.S., M.W. and J.D.G.J. reviewed and edited the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41477-022-01260-5>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41477-022-01260-5>.

**Correspondence and requests for materials** should be addressed to Jonathan D. G. Jones.

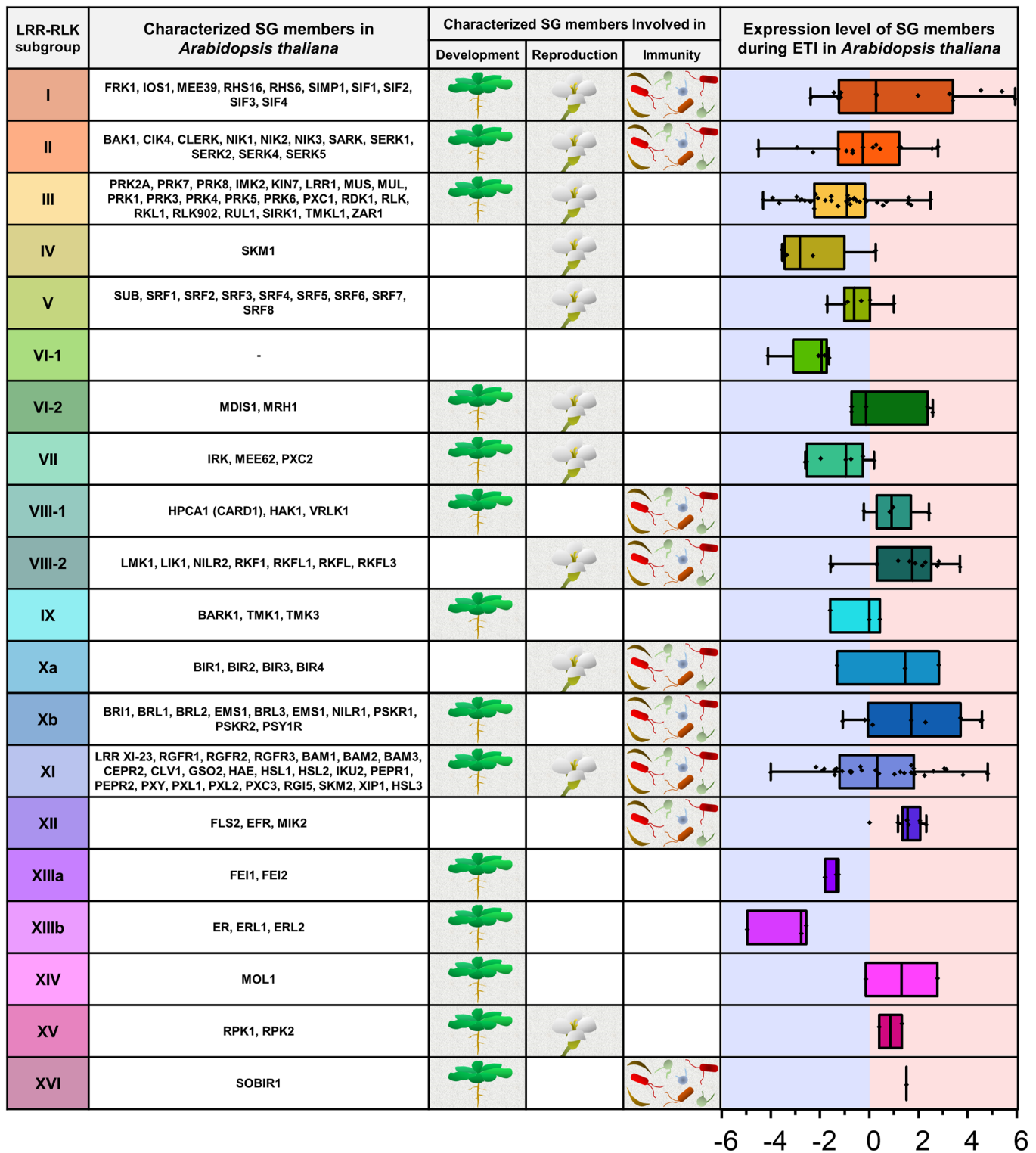
**Peer review information** *Nature Plants* thanks Guan-Zhu Han, Zhu-Qing Shao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

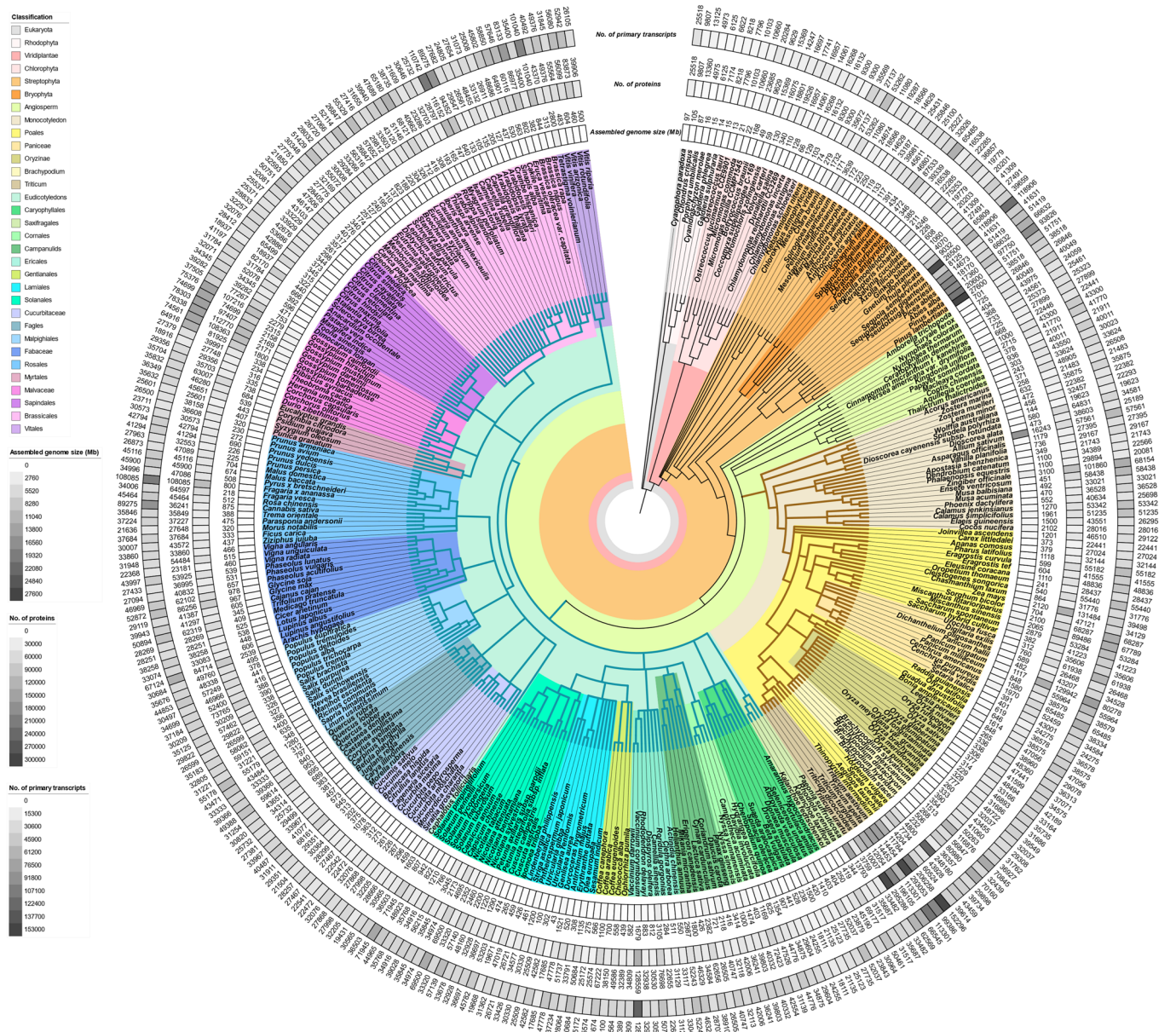
© The Author(s) 2022



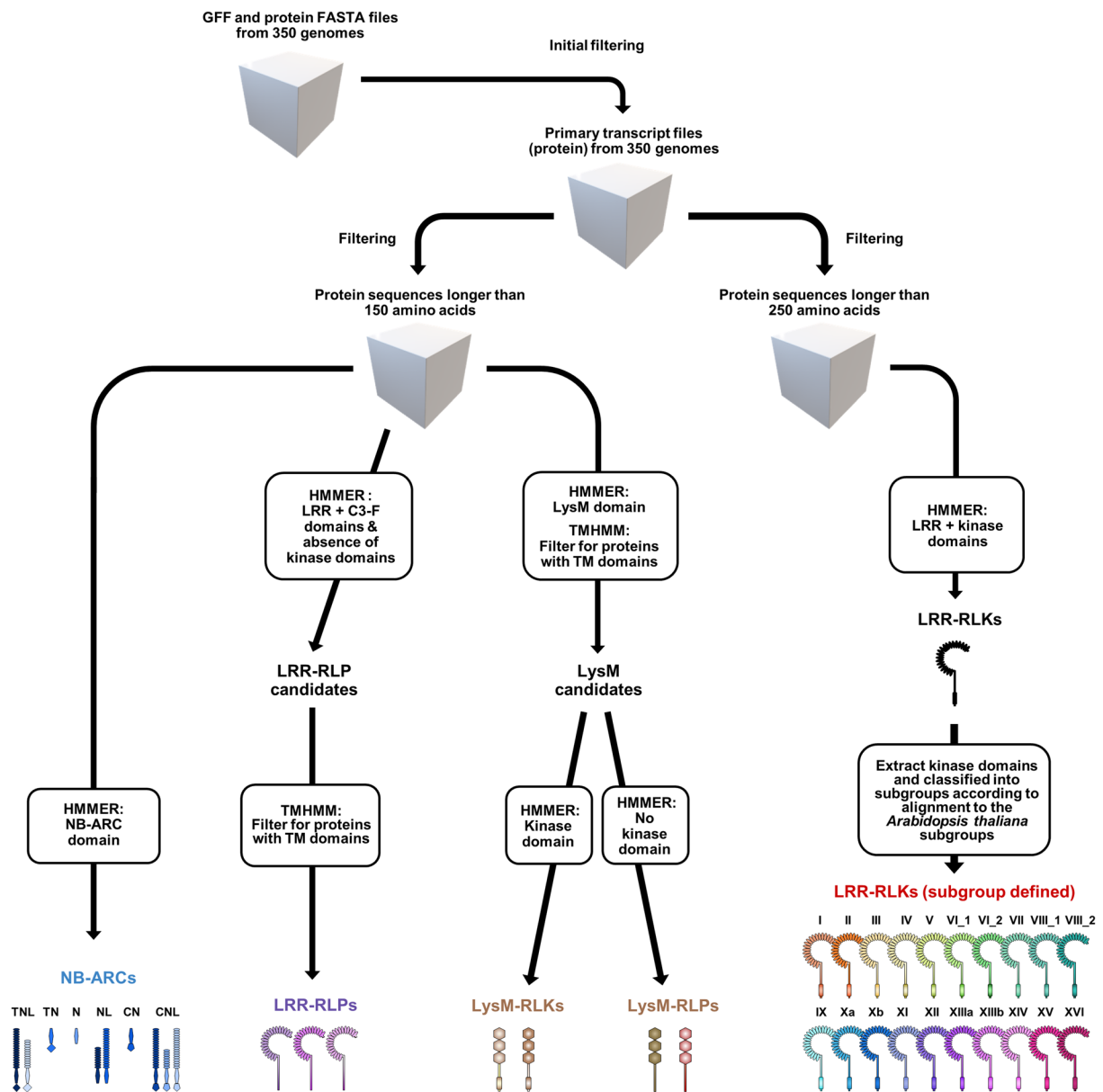
**Extended Data Fig. 1 | LRR-RLK subgroups in plants.** Table representing the characterized subgroup members in *Arabidopsis thaliana*; the biological processes of which the characterized members are involved in, and expression of subgroup members during effector-triggered immunity (ETI). Red shade represents increased expression and blue shade represents decreased expression during ETI. X-axis values represents  $\log_2$ (fold change during ETI relative to untreated samples). Box-plot elements: centre line, median; bounds of box, 25th

and 75th percentiles; whiskers,  $1.5 \times$  IQR from 25th and 75th percentiles. RNA-seq data analysed here were reported previously<sup>3</sup>. ETI is activated by estradiol-induced expression of AvrRps4 in *Arabidopsis thaliana* for 4 hours. Number of genes (n) from each LRR-RLK subgroup: I, n = 15; II, n = 15; III, n = 33; IV, n = 4; V, n = 6; VI-1, n = 4; VI-2, n = 5; VII, n = 7; VIII-1, n = 4; VIII-2, n = 12; IX, n = 3; Xa, n = 3; Xb, n = 9; XI, n = 28; XII, n = 8; XIIIa, n = 3; XIIIb, n = 3; XIV, n = 2; XV, n = 2; XVI, n = 1.



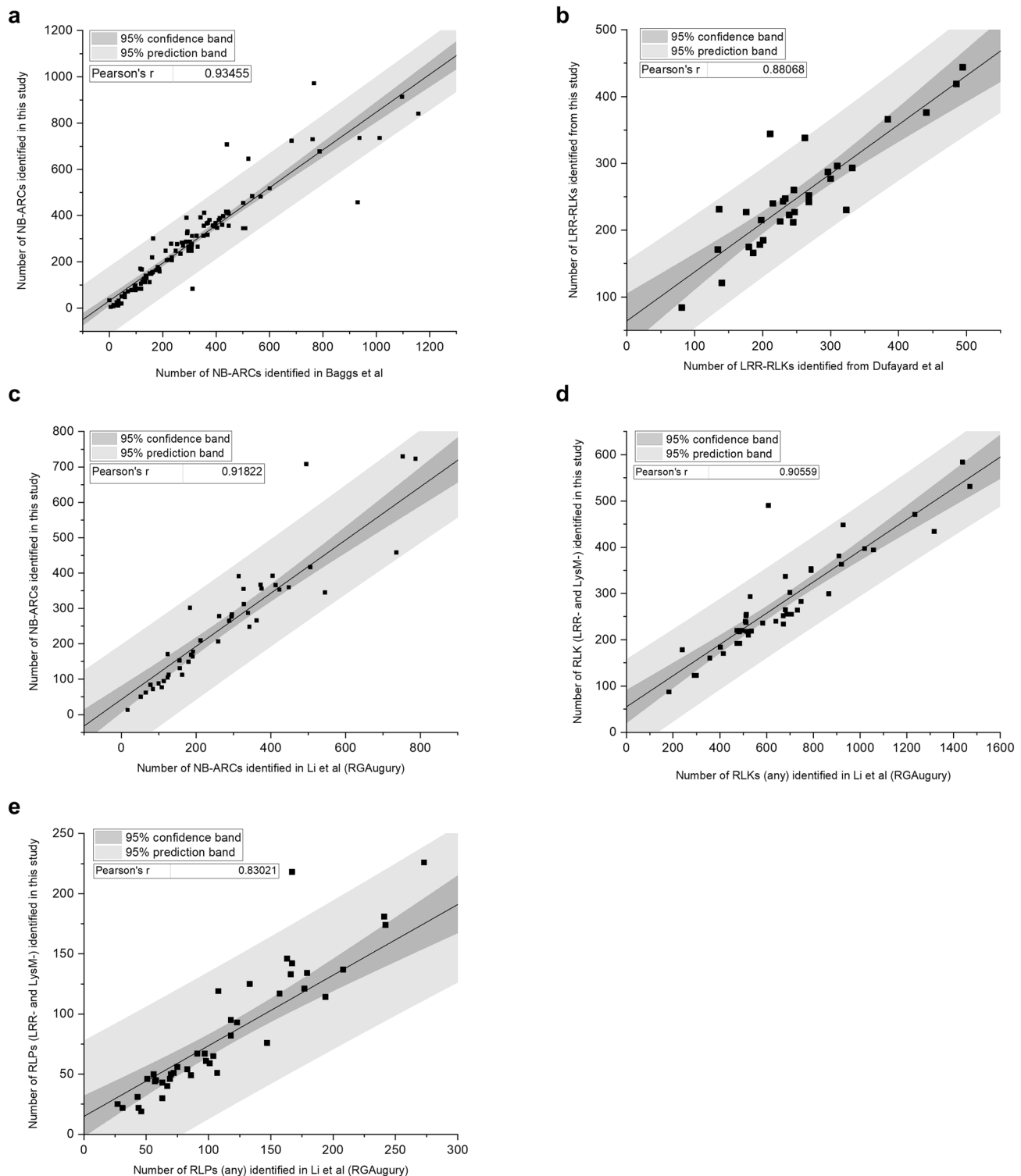


**Extended Data Fig. 2 | 350 genomes used in this study.** Phylogenetic tree of 350 plant species, with heatmaps representing the assembled genome size, number (no.) of annotated proteins and number (no.) of primary transcripts. Brown branches indicate monocots and teal branches represent eudicots.



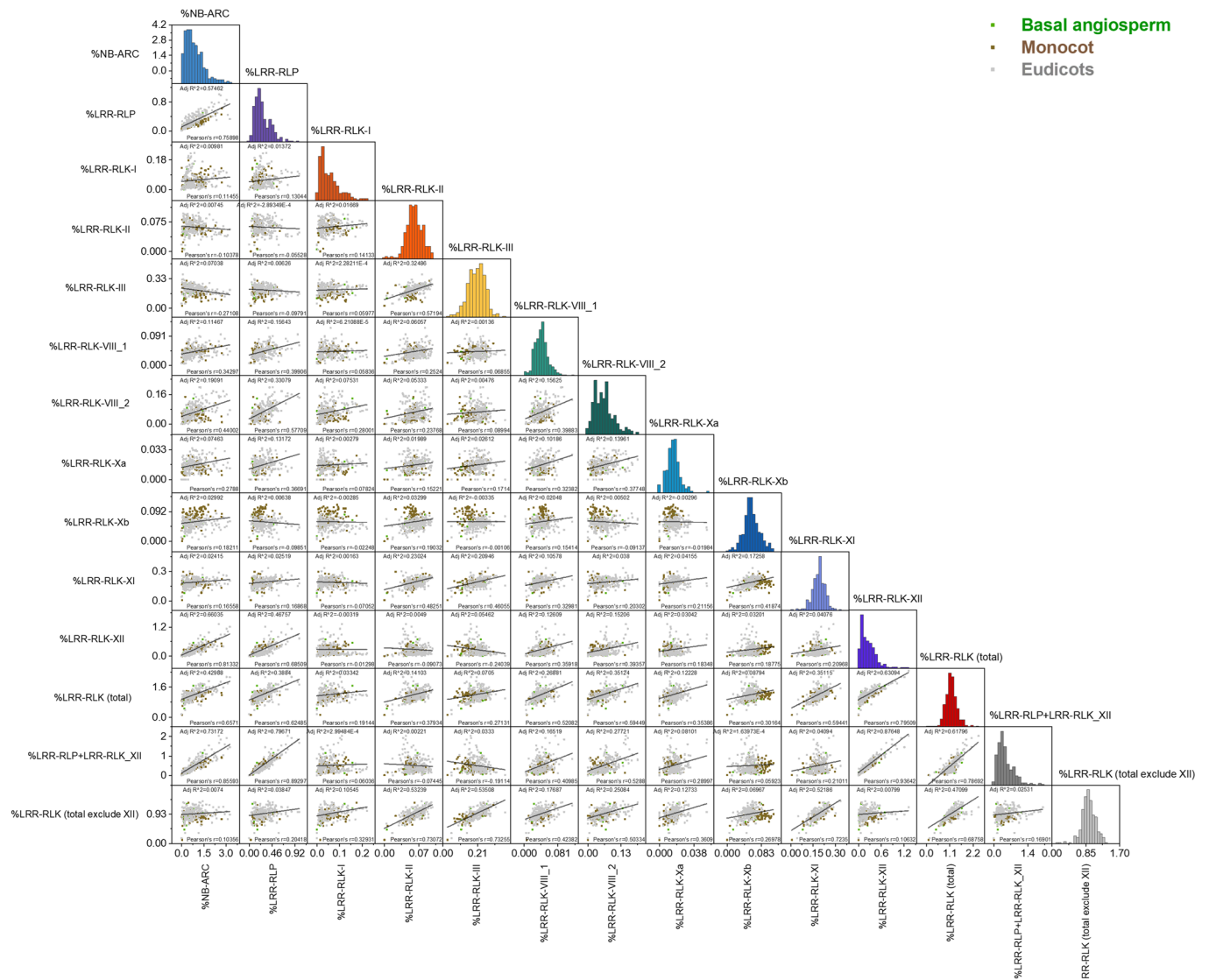
**Extended Data Fig. 3 | Identification of receptor gene families in 350 plant genomes.** Protein sequences from all 350 proteomes were first filtered for the primary gene models. Primary-transcript proteomes were then filtered and NB-ARC, LRR-RLP, LysM-RLK, LysM-RLP and LRR-RLK proteins were identified. LRR-

RLK genes were further classified into 20 subgroups according to their alignment to the *Arabidopsis thaliana* subgroups. Details of the pipeline are described in the methods section.



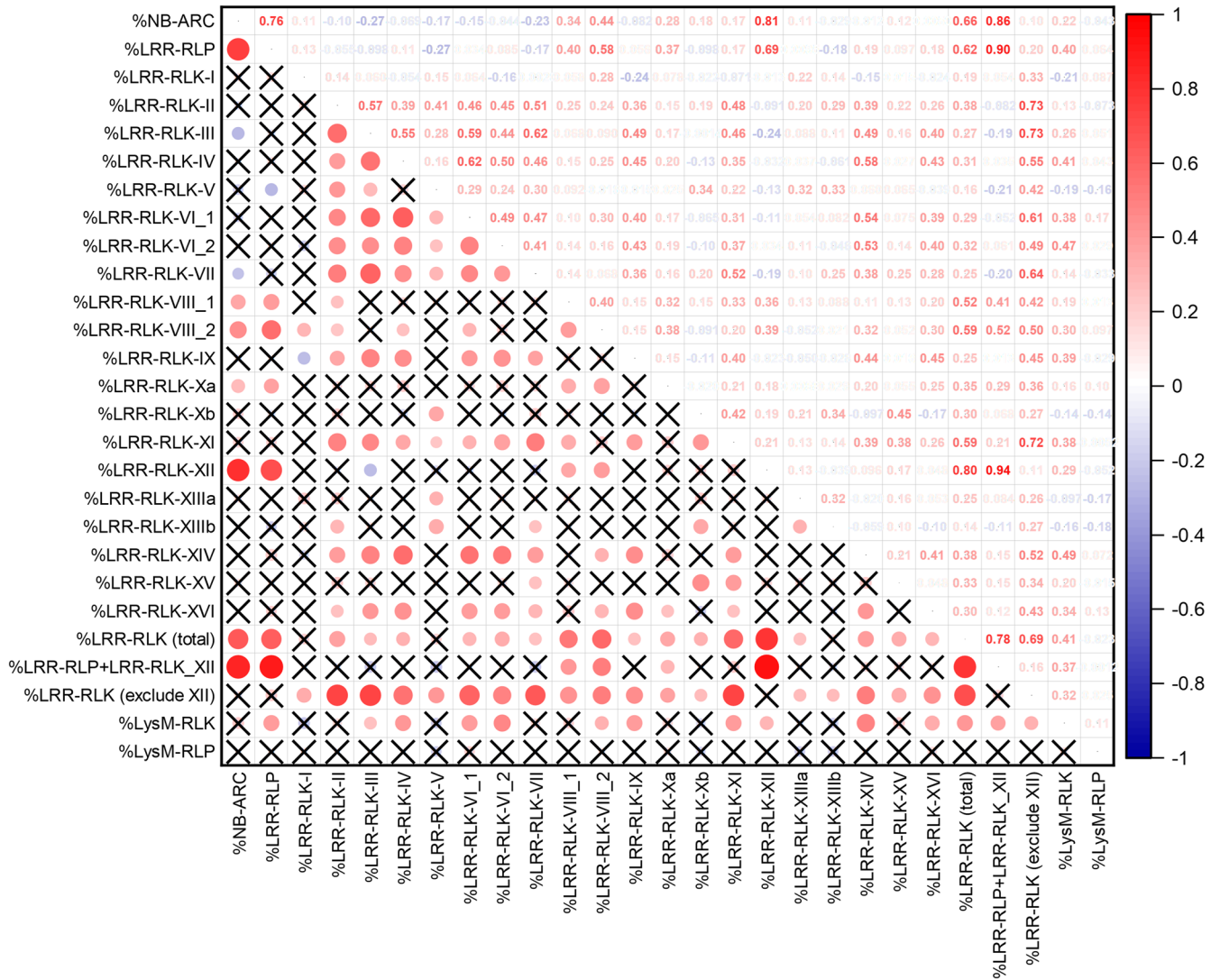
**Extended Data Fig. 4 | Validation of the methods used to identify receptor genes in this study.** **a**, Scatter plot of number of NB-ARC genes identified in this study against number of NB-ARC genes identified in Baggs et al for the corresponding plant species<sup>16</sup>. **b**, Scatter plot of number of LRR-RLKs identified in this study against number of LRR-RLKs identified in Dufayard et al for the corresponding plant species<sup>17</sup>. **c**, Scatter plot of number of NB-ARC genes identified in this study against number of NB-ARC genes identified in Li et al (RGAugury database) for the corresponding plant species (ref). **d**, Scatter plot of

number of RLK genes (LRR- and LysM-RLKs only) identified in this study against number of RLK genes (any RLKs, including LRR- and LysM-RLKs) identified in Li et al (RGAugury database) for the corresponding plant species<sup>18</sup>. **e**, Scatter plot of number of RLP genes (LRR- and LysM-RLPs only) identified in this study against number of RLP genes (any RLPs, including LRR- and LysM-RLPs) identified in Li et al (RGAugury database) for the corresponding plant species<sup>18</sup>. For each plot, the black line represents the linear trend, with dark grey shade represents the 95% confidence interval and light grey shade represents the 95% prediction interval.



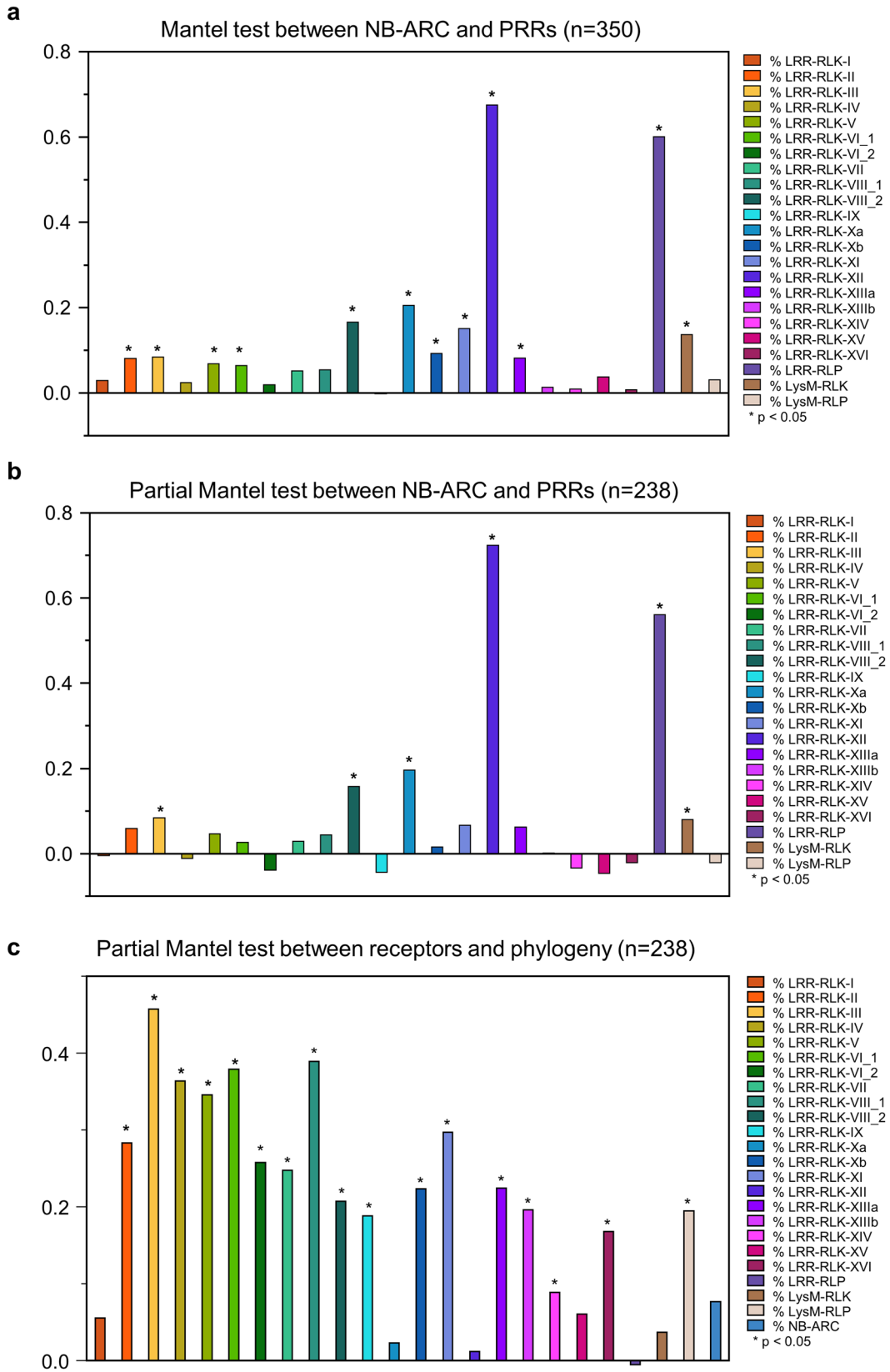
**Extended Data Fig. 5 | Scatter plot of Pearson correlation analysis between %NB-ARCs, %LRR-RLPs and %LRR-RLKs.** Bottom left boxes include scatter plot between the corresponding % receptor-gene families in 300 angiosperms. Black line represents the linear trend. Grey dots represent basal angiosperms (n = 13),

brown dots represent monocots (n = 79) and green dots represent eudicots (n = 208). The Pearson correlation coefficient (Pearson's r) is indicated below each scatter plot. The diagonal boxes include the distribution of % receptor-gene families in 300 angiosperms.



**Extended Data Fig. 6 | Pearson correlation plot of %NB-ARCs, %LRR-RLPs, %LRR-RLKs, %LysM-RLKs and %LysM-RLPs.** Bottom left squares indicate the strength of correlation between the corresponding % receptor-gene families in 300 angiosperms. Two-sided test of significance was performed, and Bonferroni correction was performed to adjust the p-value for all the independent tests. The Bonferroni-corrected p-value at  $\alpha = 0.05$  would be 0.0001479. Thus, P-values < 0.0001479 are considered as significant after correction. Red circles indicate

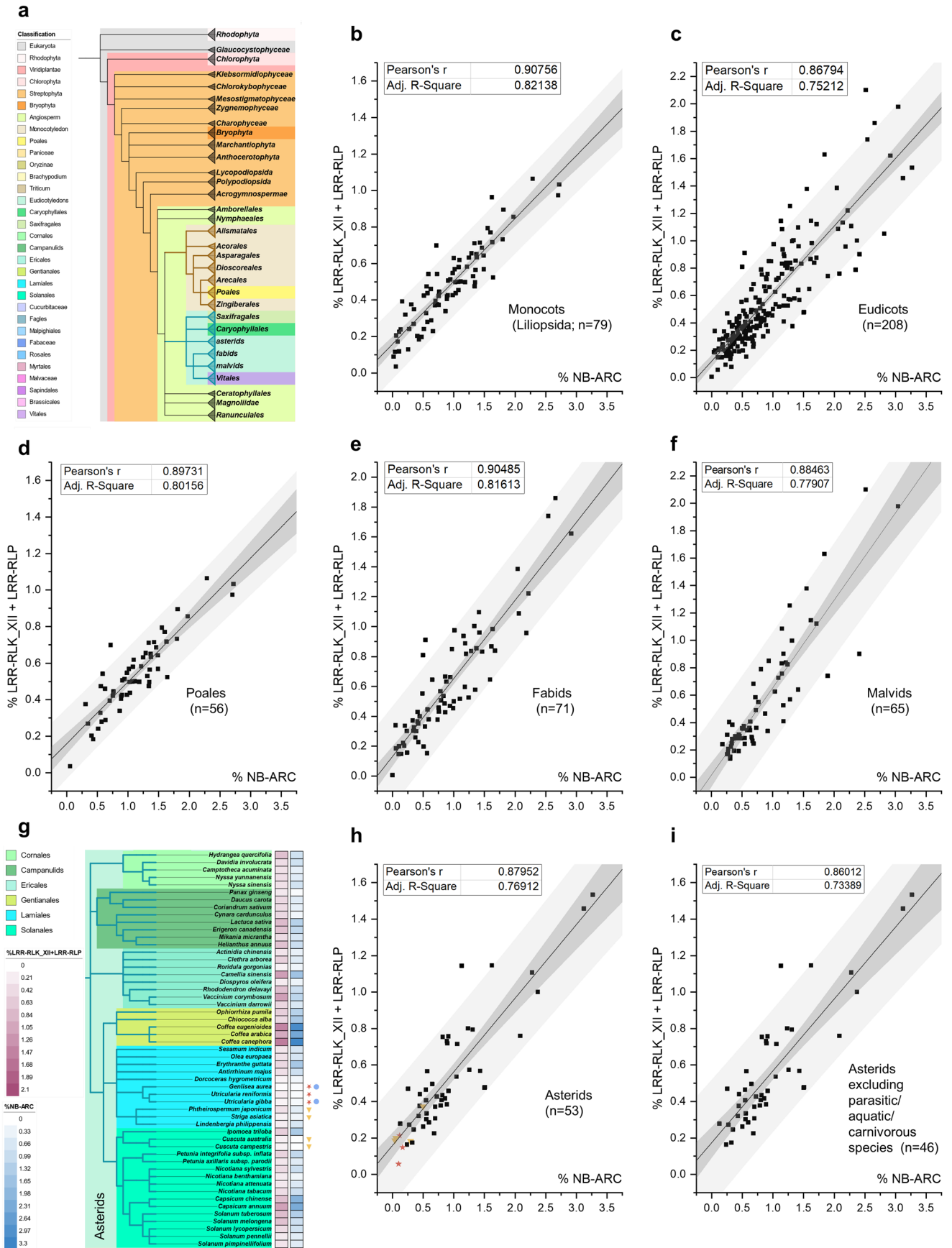
significant and positive linear correlations, blue circles indicate significant and negative linear correlations and crosses (X) indicate insignificant correlations. The sizes of circles represent the strength of the correlations. Top right squares indicate the values of Pearson correlation coefficient between the corresponding %receptor-gene families in 300 angiosperms. Red values represent positive correlations and blue values represent negative correlations. Exact p-values and summary of statistical analyses are provided in Supplementary Table 4.



Extended Data Fig. 7 | See next page for caption.

**Extended Data Fig. 7 | Test for similarities in fraction of receptor proteins and phylogenetic relationships.** To test whether the receptor percentage (%) found per species correlate with phylogenetic relationship, %receptor and the species phylogenetic tree were converted to distance matrices and tested for correlation with the Mantel tests. **a**, Mantel test for the correlation between distance matrices obtained for % receptor family proteins in 350 species. Histogram shows the correlation coefficient between % PRRs and % NB-ARCs. **b**, Partial Mantel test for the correlation between distance matrices obtained for % receptor in

238 species, taking phylogenetic distances into account. Histogram shows the correlation coefficient between % PRRs and % NB-ARCs. **c**, Mantel test for the correlation between distance matrices obtained for %receptors and phylogenetic distances in 238 species. Histogram shows the correlation coefficient between %receptor and phylogeny. One-sided test of significance was performed. P-values were corrected for multiple testing to reflect false discovery rates (FDRs, for details please refer to methods). Summary of statistical analyses are provided in Supplementary Table 4.



Extended Data Fig. 8 | See next page for caption.



**Extended Data Fig. 8 | Correlation between %PRRs and %NLRs in plant clades.**

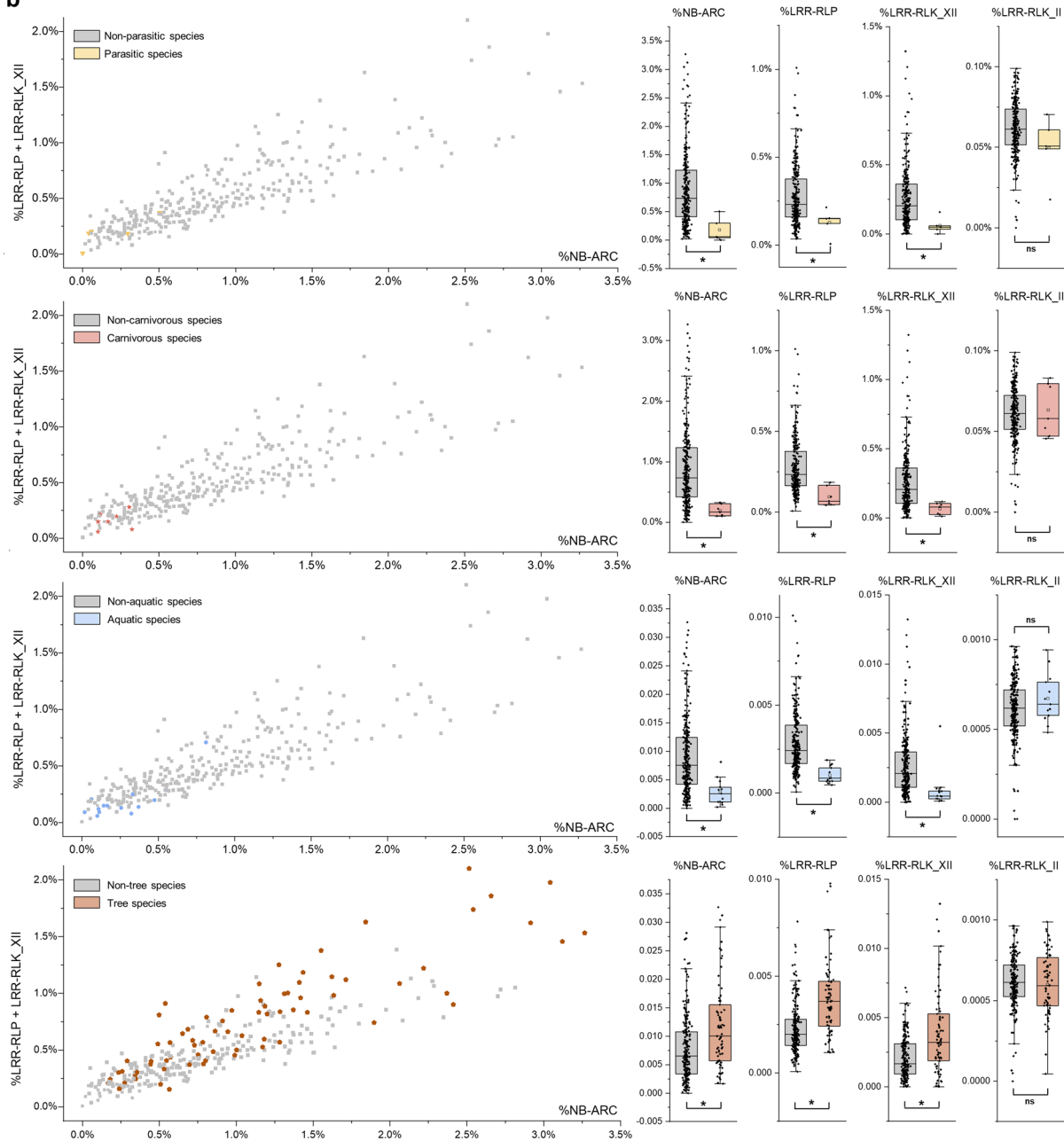
**a**, Phylogenetic tree of the species used in the analysis. Pearson correlation between %LRR-RLK\_XII+LRR-RLP and %NB-ARC in **b**, Monocots, **c**, Eudicots, **d**, Poales, **e**, Fabids, **f**, Malvids and **h**, Asterids. **g**, Phylogenetic tree of the Asterids clade used in the analysis. Carnivorous plants are marked with orange

stars; parasitic plants are marked with yellow triangles and aquatic plants are marked with blue circles. **i**, Pearson correlation between %LRR-RLK\_XII+LRR-RLP and %NB-ARC in the Asterids clade, excluding carnivorous, parasitic, and aquatic plants.

**a**

Parasitic species (n=5)	Carnivorous species (n=7)	Aquatic species (n=13)	Trees (n=75)					
<i>Striga asiatica</i>	<i>Genlisea aurea</i>	<i>Genlisea aurea</i>	<i>Persea americana</i> var. <i>drymifolia</i>	<i>Coffea canephora</i>	<i>Dimocarpus longan</i>	<i>Morella rubra</i>	<i>Salix purpurea</i>	<i>Parasponia andersonii</i>
<i>Phtheirospermum japonicum</i>	<i>Utricularia gibba</i>	<i>Utricularia gibba</i>	<i>Cinnamomum micranthum</i> f. <i>kanehirae</i>	<i>Coffea arabica</i>	<i>Sclerocarya birrea</i>	<i>Juglans regia</i>	<i>Salix dunnii</i>	<i>Pyrus x bretschneideri</i>
<i>Cuscuta australis</i>	<i>Aldrovanda vesiculosa</i>	<i>Aldrovanda vesiculosa</i>	<i>Phoenix dactylifera</i>	<i>Olea europaea</i>	<i>Anacardium occidentale</i>	<i>Carya illinoensis</i>	<i>Salix suzhouensis</i>	<i>Malus baccata</i>
<i>Cuscuta campestris</i>	<i>Utricularia reniformis</i>	<i>Zostera marina</i>	<i>Cocos nucifera</i>	<i>Punica granatum</i>	<i>Pistacia vera</i>	<i>Corylus avellana</i>	<i>Populus euphratica</i>	<i>Malus domestica</i>
<i>Sapria himalayana</i>	<i>Cephalotus follicularis</i>	<i>Zostera muelleri</i>	<i>Elaeis guineensis</i>	<i>Psidium guajava</i>	<i>Mangifera indica</i>	<i>Carpinus fangiana</i>	<i>Populus tremuloides</i>	<i>Prunus amenica</i>
	<i>Drosera spatulata</i>	<i>Nymphaea thermarum</i>	<i>Camptotheca acuminata</i>	<i>Syzygium oleosum</i>	<i>Citrus medica</i>	<i>Betula platyphylla</i>	<i>Populus deltoides</i>	<i>Prunus avium</i>
	<i>Dionaea muscipula</i>	<i>Nymphaea colorata</i>	<i>Davidia involucreta</i>	<i>Eucalyptus grandis</i>	<i>Citrus sinensis</i>	<i>Castanea mollissima</i>	<i>Populus alba</i>	<i>Prunus yedoensis</i>
		<i>Spirodela polyrhiza</i>	<i>Nyssa yunnanensis</i>	<i>Corymbia citriodora</i>	<i>Citrus cavaleriei</i>	<i>Castanea dentata</i>	<i>Populus trichocarpa</i>	<i>Prunus dulcis</i>
		<i>Euryale ferox</i>	<i>Nyssa sinensis</i>	<i>Durio zibethinus</i>	<i>Citrus trifoliata</i>	<i>Quercus rubra</i>	<i>Populus tremula</i>	<i>Prunus persica</i>
		<i>Ceratophyllum demersum</i>	<i>Diospyros oleifera</i>	<i>Theobroma cacao</i>	<i>Citrus clementina</i>	<i>Quercus lobata</i>	<i>Ziziphus jujuba</i>	<i>Chiococca alba</i>
		<i>Lemna minor</i>	<i>Camellia sinensis</i>	<i>Herrania umbriatica</i>	<i>Citrus reticulata</i>	<i>Quercus suber</i>	<i>Morus notabilis</i>	
		<i>Lepidium sativum</i>	<i>Clethra arborea</i>	<i>Corchorus olitorius</i>	<i>Citrus hindsii</i>	<i>Hevea brasiliensis</i>	<i>Ficus carica</i>	
		<i>Wolffia australiana</i>	<i>Rhododendron delavayi</i>	<i>Toona sinensis</i>	<i>Carica papaya</i>	<i>Salix brachista</i>	<i>Trema orientale</i>	

**b**



Extended Data Fig. 9 | See next page for caption.

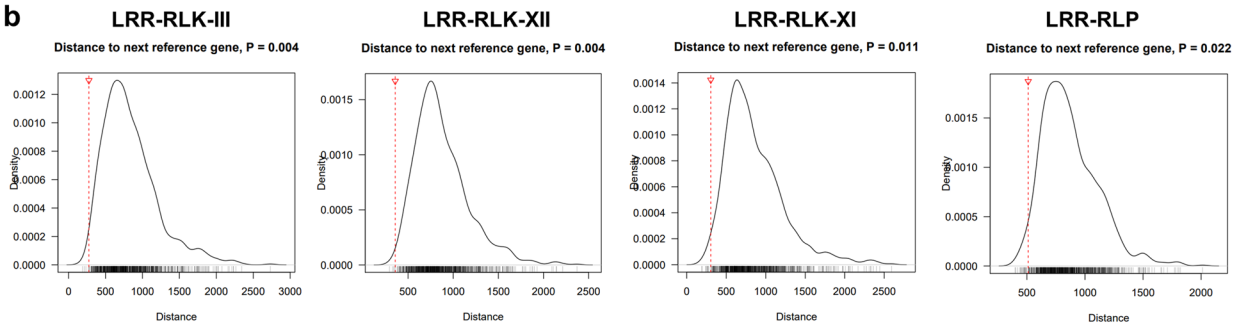
**Extended Data Fig. 9 | Expansion and contraction of PRR- and NLR-gene families in plants adapted to particular lifestyles and ecological niches. a,** List of parasitic species, carnivorous species, aquatic species and trees from this study. **b,** (left) Scatter plot of % LRR-RLP+LRR-RLK-XII against % NB-ARC in 300 angiosperms. Parasitic, carnivorous, aquatic and trees species are highlighted as yellow inverted triangles, orange stars, blue circles and brown pentagons, respectively. (right) Box-plot of % NB-ARC, LRR-RLP, LRR-RLK-XII and LRR-RLK-II in parasitic (yellow), carnivorous (orange), aquatic (blue) and trees species

(brown) compared to other plant species (gray). Box-plot elements: centre line, median; bounds of box, 25th and 75th percentiles; whiskers,  $1.5 \times$  IQR from 25th and 75th percentiles. A two-sided Welch's t-test was used to analyze significant differences between the groups ( $*P \leq 0.05$ ; ns, not significant). Exact P-values are provided in Supplementary Table 4. Number of species (n) in each category: non-parasitic species, n = 295; parasitic species, n = 5; non-carnivorous species, n = 293; carnivorous species, n = 7; non-aquatic species, n = 287; aquatic species, n = 13; tree species, n = 225; non-tree species, n = 75.

**a**

Test set ( <i>S.tuberosum</i> )	Number of genes in the subgroup	Tested distance (Observed 90%)	Sampled distance (Expected 90%)	Distance ratio	P-value
LRR-RLK-III to NB-ARC	41	275	836.69	0.329	<b>0.004</b>
LRR-RLK-XII to NB-ARC	68	358	875.45	0.409	<b>0.004</b>
LRR-RLK-XI to NB-ARC	42	304.3	858.79	0.354	<b>0.011</b>
LRR-RLK-RLP to NB-ARC	108	510.9	871.04	0.587	<b>0.022</b>

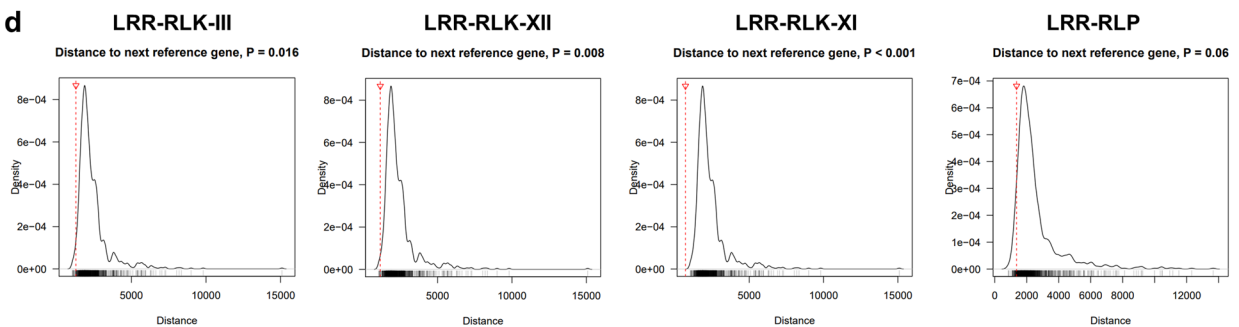
**b**



**c**

Test set ( <i>Z.mays</i> )	Number of genes in the subgroup	Tested distance (Observed 90%)	Sampled distance (Expected 90%)	Distance ratio	P-value
LRR-RLK-III to NB-ARC	43	1276.2	2375.67	0.537	<b>0.016</b>
LRR-RLK-XII to NB-ARC	43	1140.6	2375.67	0.48	<b>0.008</b>
LRR-RLK-XI to NB-ARC	43	707	2375.67	0.298	<b>&lt; 0.001</b>
LRR-RLK-RLP to NB-ARC	27	1362	2551.21	0.534	0.06

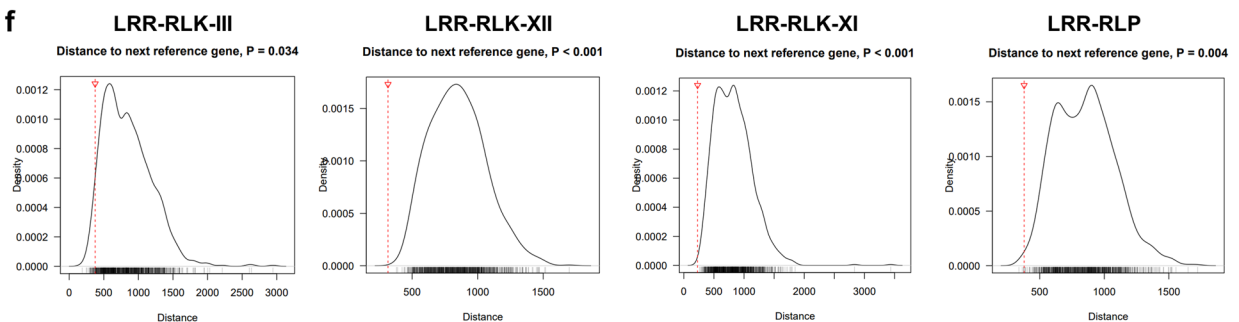
**d**



**e**

Test set ( <i>O.sativa</i> )	Number of genes in the subgroup	Tested distance (Observed 90%)	Sampled distance (Expected 90%)	Distance ratio	P-value
LRR-RLK-III to NB-ARC	33	374.8	846.58	0.443	<b>0.034</b>
LRR-RLK-XII to NB-ARC	98	315.8	850.13	0.371	<b>&lt; 0.001</b>
LRR-RLK-XI to NB-ARC	42	228.6	826.77	0.276	<b>&lt; 0.001</b>
LRR-RLK-RLP to NB-ARC	79	379.4	850.46	0.446	<b>0.004</b>

**f**



**Extended Data Fig. 10 | Genomic clustering of LRR-RLKs, LRR-RLPs and NB-ARCs in *Solanum tuberosum*, *Zea mays* and *Oryza sativa*. a, c, e.** Table summarizing the statistical analysis of genomic clustering between PRRs (LRR-RLKs and LRR-RLPs) and NB-ARCs in *Solanum tuberosum* (a), *Zea mays* (c) and *Oryza sativa* (e). The 90-percentile distance (bp) between PRR gene family members and the next closest NB-ARC genes were calculated. This is then compared to a distribution (n = 1000) of 90-percentile distances between randomly-sampled genes and the next closest NB-ARC genes. One-sided test was

performed to test the differences between tested distance (PRRs) and sampled distance (randomly-sampled). P-values are calculated based on the comparison to 1000 cases of randomly-sampled genes. Significant values are indicated in bold (p-value < 0.05 is considered as significant). **b, d, f.** Distribution (n = 1000) of 90-percentile distances (bp) between randomly-sampled genes and the next closest NB-ARC genes in *Solanum tuberosum* (b), *Zea mays* (d) and *Oryza sativa* (f). Red lines indicate the 90-percentile distance between the corresponding PRR gene family members and the next closest NB-ARC genes.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used for data collection.

Data analysis

Pearson's correlation analyses were performed with OriginPro (version 2022, <https://www.originlab.com/>).

The taxonomic tree was obtained from NCBI (<https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi>). Phylogenetic tree of the 350 species are generated by phyloT (<https://phylot.biobyte.de/>) based on NCBI taxonomy database. Phylogenetics trees are visualised and figures are generated by iTOL.

All other analyses were done as described with publicly available tools (hmmer, tmhmm, diamond, FastTree, FAMSA, gotree, R).

hmmer, version 3.1b2

tmhmm, version 2.0

diamond, version 0.9.26

FastTree, version 2.1.11 SSE3

FAMSA, version 2.1.11

R, version 3.4.4

gotree, v0.4.2

Scripts will be available on [github.com/MWSchmid/Ngou-et-al.-2022](https://github.com/MWSchmid/Ngou-et-al.-2022)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data generated or analysed during this study are included in the article or supplementary information files. Proteomes of 350 species used in this study are downloaded from either NCBI, Phyzone13, ensemblplants, JGI, Fernbase, Penium Genome Database or directly from the publications. A complete list of the proteomes and associated data used in this study are provided in Supplementary Table 1. Sequences of the identified receptors and phylogenetic analyses are available on Zenodo (<https://doi.org/10.1101/2022.01.01.474684>).

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined by the number of publicly available plant genome assemblies. 350 species with a representative genome assembly were included.
Data exclusions	No data were excluded from the analyses.
Replication	Does not apply as this study does not comprise a planned experiment with experimental groups. We selected all species for which a high-quality proteome was publicly available at the time of the study. Only bioinformatic approach was used in this study, thus we believe that replicability does not apply in this context.
Randomization	Does not apply as this study does not comprise a planned experiment with experimental groups. We selected all species for which a high-quality proteome was publicly available at the time of the study.
Blinding	We analyzed data from previously published studies and we did not include any blinding measures. We were not involved in the sample handling or sequencing of these genomes. Thus, blinding was not relevant to this study.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging