Plant Communications, Volume *1*

Supplemental Information

Variation Patterns of NLR Clusters in *Arabidopsis thaliana* Genomes

Rachelle R.Q. Lee and Eunyoung Chae

Supplemental Information

Patterns of NLR Cluster Variation in *Arabidopsis thaliana* Genomes

Rachelle R.Q. Lee, Eunyoung Chae*

Department of Biological Sciences, National University of Singapore, Singapore 117558
*Correspondence to Eunyoung Chae
(dbsce@nus.edu.sg)

List of Supplementary Figures

List of Supplementary Tables

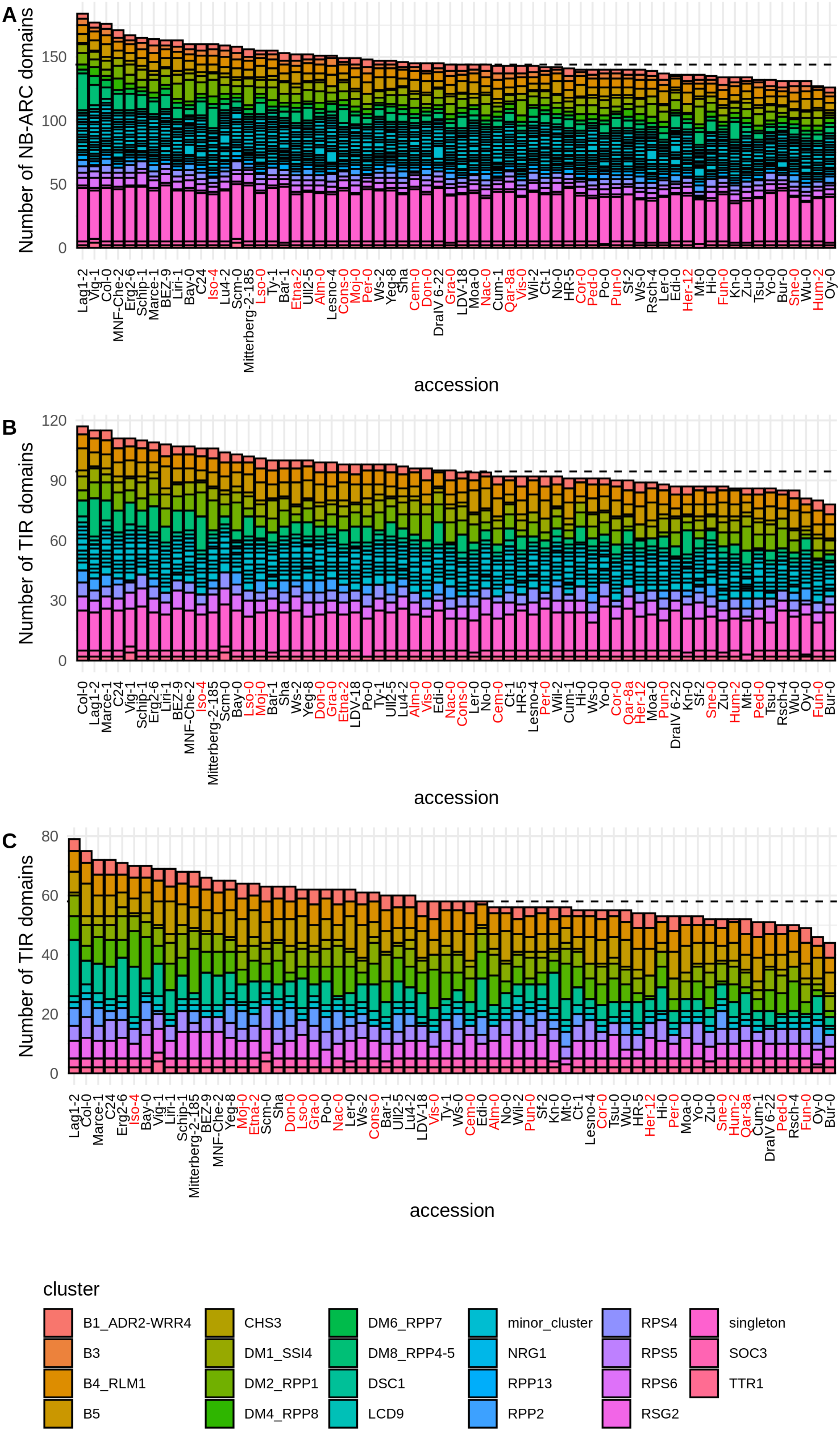
A
Number of NB-ARC domains
0
50
100
150
Lag1-2
Vig-1
Col-0
MNF-Che-2
Erg2-6
Schip-1
Marce-1
BEZ-9
Liri-1
Bay-0
C24
Iso-4
Lu4-2
Scm-0
Mitterberg-2-185
Lso-0
Ty-1
Bar-1
Etna-2
Ull2-5
Alm-0
Lesno-4
Cons-0
Moj-0
Per-0
Ws-2
Yeg-8
Sha
Cem-0
Don-0
DralV 6-22
Gra-0
LDV-18
Moa-0
Nac-0
Cum-1
Qar-8a
Vis-0
Wil-2
Ct-1
No-0
HR-5
Cor-0
Ped-0
Po-0
Pun-0
Sf-2
Ws-0
Rsch-4
Ler-0
Edi-0
Her-12
Mt-0
Hi-0
Fun-0
Kn-0
Zu-0
Tsu-0
Yo-0
Bur-0
Sne-0
Wu-0
Hum-2
Oy-0
accession
B
Number of TIR domains
0
30
60
90
120
Col-0
Lag1-2
Marce-1
C24
Vig-1
Schip-1
Erg2-6
Liri-1
BEZ-9
Iso-4
MNF-Che-2
Mitterberg-2-185
Scm-0
Bay-0
Lso-0
Moj-0
Bar-1
Sha
Ws-2
Yeg-8
Don-0
Gra-0
Etna-2
LDV-18
Po-0
Ty-1
Ull2-5
Lu4-2
Alm-0
Vis-0
Edi-0
Nac-0
Cons-0
Ler-0
No-0
Cem-0
Ct-1
HR-5
Lesno-4
Per-0
Wil-2
Cum-1
Hi-0
Ws-0
Yo-0
Cor-0
Qar-8a
Her-12
Moa-0
Pun-0
DralV 6-22
Kn-0
Sf-2
Sne-0
Zu-0
Hum-2
Mt-0
Ped-0
Tsu-0
Rsch-4
Wu-0
Oy-0
Fun-0
Bur-0
accession
C
Number of TIR domains
0
20
40
60
80
Lag1-2
Col-0
Marce-1
C24
Erg2-6
Iso-4
Bay-0
Vig-1
Liri-1
Schip-1
Mitterberg-2-185
BEZ-9
MNF-Che-2
Yeg-8
Moj-0
Etna-2
Scm-0
Sha
Don-0
Lso-0
Gra-0
Po-0
Nac-0
Ler-0
Ws-2
Cons-0
Bar-1
Ull2-5
Lu4-2
LDV-18
Vis-0
Ty-1
Ws-0
Cem-0
Edi-0
Alm-0
No-0
Wil-2
Pun-0
Sf-2
Kn-0
Mt-0
Ct-1
Lesno-4
Cor-0
Tsu-0
Wu-0
HR-5
Her-12
Hi-0
Per-0
Moa-0
Yo-0
Zu-0
Sne-0
Hum-2
Qar-8a
DralV 6-22
Cum-1
Rsch-4
Ped-0
Fun-0
Oy-0
Bur-0
accession
cluster
B1_ADR2-WRR4
B3
B4_RLM1
B5
CHS3
DM1_SSI4
DM2_RPP1
DM4_RPP8
DM6_RPP7
DM8_RPP4-5
DSC1
LCD9
minor_cluster
NRG1
RPP13
RPP2
RPS4
RPS5
RPS6
RSG2
singleton
SOC3
TTR1

Figure S1. **Predicted repertoire sizes in 64 *A. thaliana* accessions. (A)** Total NB-ARC domains predicted. **(B)** Total TIR domains predicted. **(C)** TIR domains predicted in major clusters. The median for each plot is given as a dashed line. When restricted to major clusters, the largest number of TIR domain homologues was predicted in Lag1-2 (accession ID 9100; 79 homologues), the smallest in Bur-0 (accession ID 7058, 44).

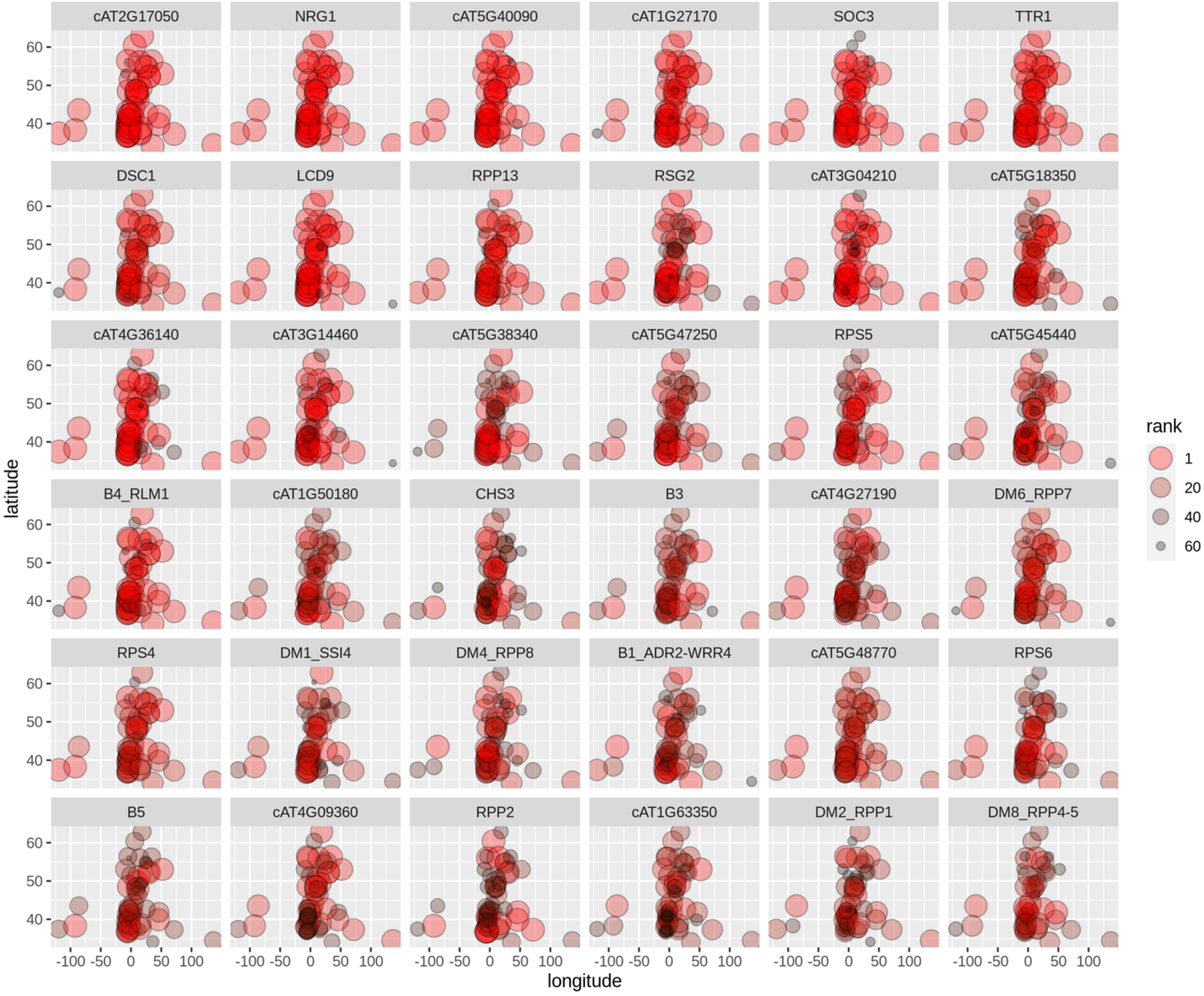

Figure S2. Cluster size plotted against coordinates. Accessions were restricted to the 56 with longitude and latitude data. Size and colour correspond to rank in terms of cluster size, with the largest, lightest circle representing the accession with the largest number of NB-ARC domains assigned to each cluster.

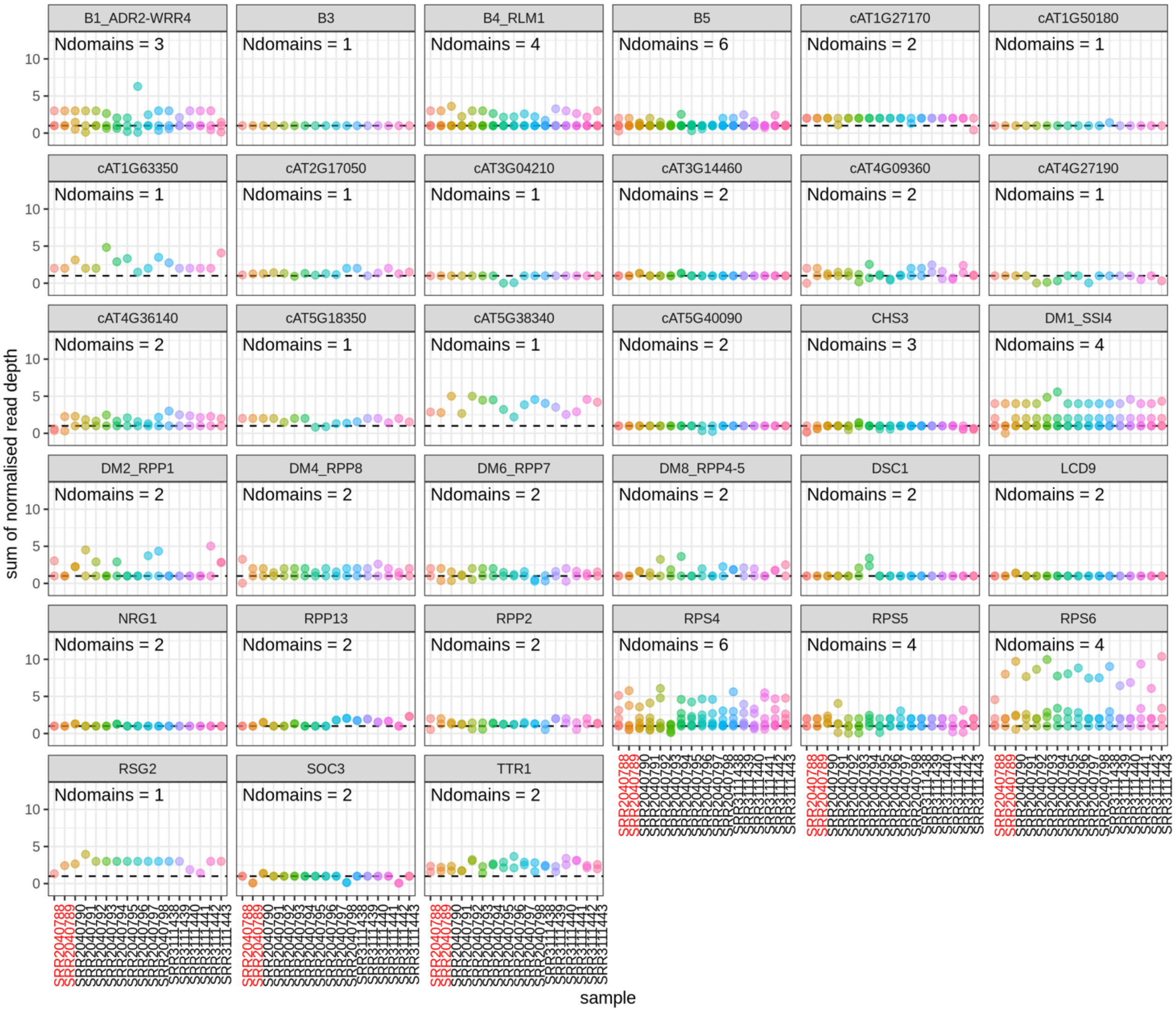

Figure S3. Normalised read depth of cluster NB-ARC homologues in 17 *A. lyrata* accessions. Short-read from whole genome shotgun sequencing generated by Novikova et al. (2016) were mapped to the reference *A. lyrata* genome and normalised read depth (nRD) was estimated using CNVnator (Abyzov et al., 2011). For all *A. lyrata* homologues assigned to each NLR NB-ARC domain in the reference *A. thaliana* genome Col-0, their nRD was summed per accession and plotted as a single point, grouped by the cluster to which the NB-ARC domain belongs in Col-0. The number of Col-0 domains with homologues in the *A. lyrata* reference genome is given for each cluster. Accessions that belong to subspecies lyrata and petraea are shown in red and black respectively.

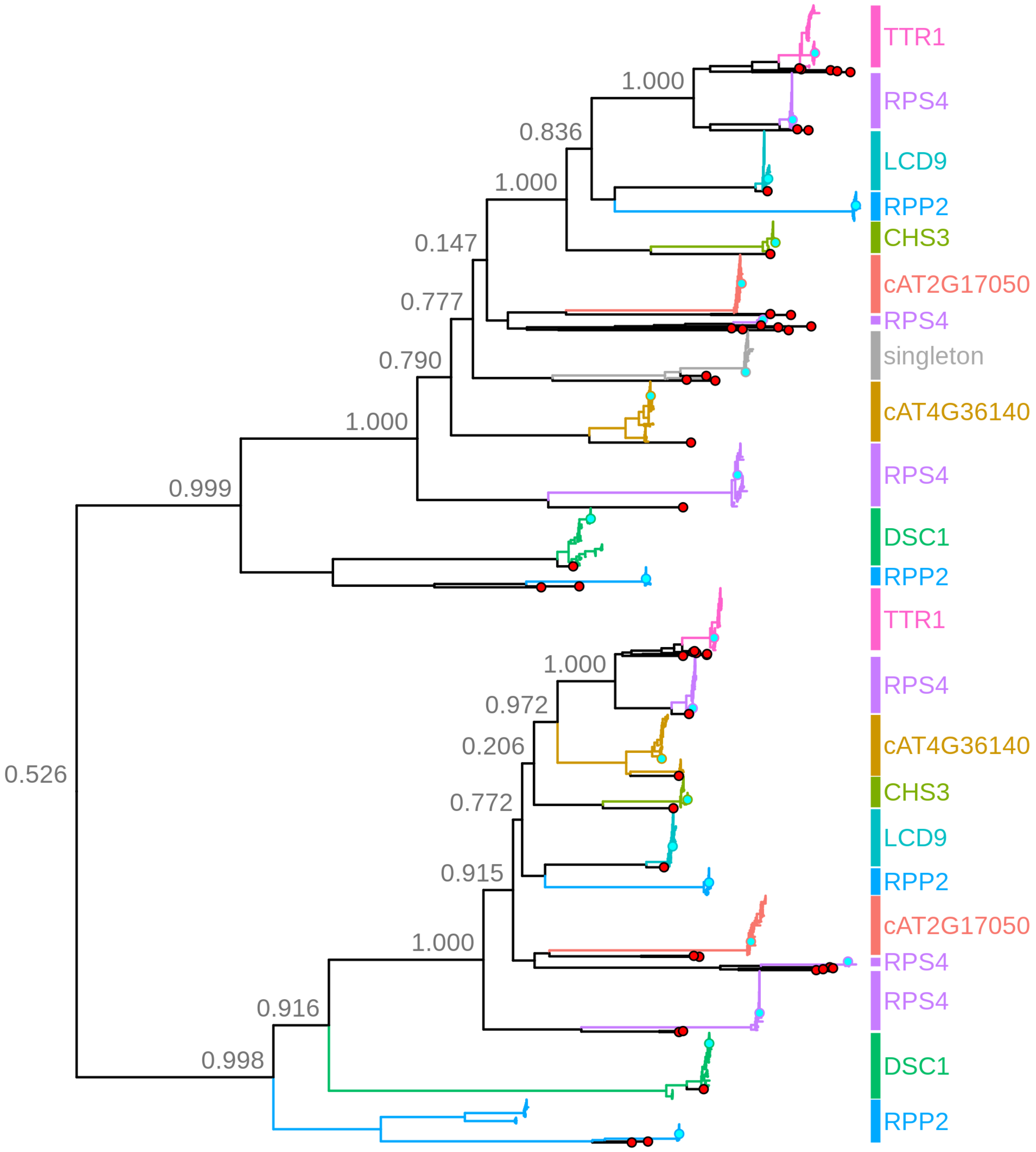

Figure S4. NB-ARC clade of paired genes showing CNV conservation. Reference *A. thaliana* (accession Col-0) and *A. lyrata* domains are marked with a cyan and red circle respectively. Clades are coloured by cluster.

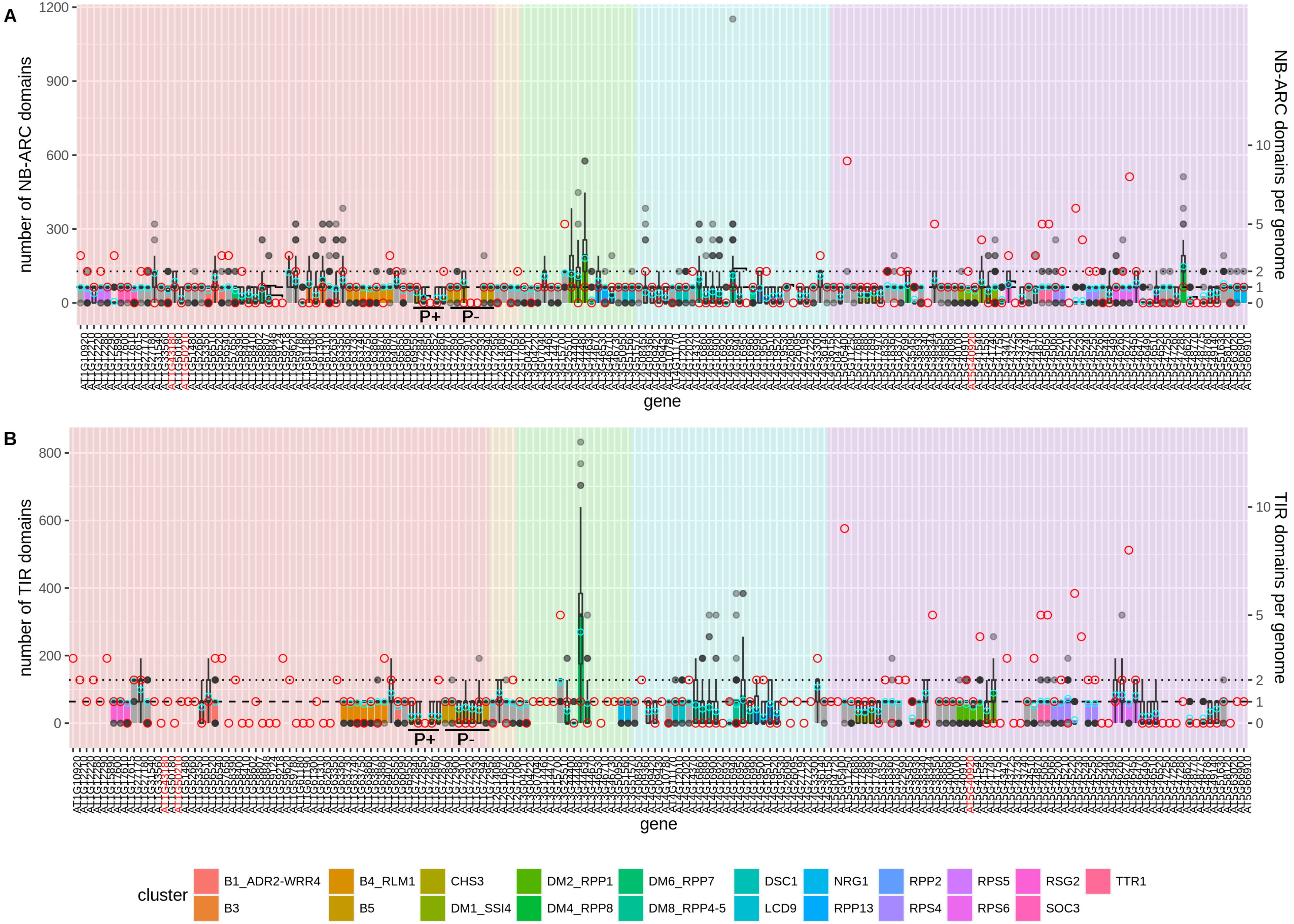
A
number of NB-ARC domains
1200
900
600
300
0
NB-ARC domains per genome
10
5
2
1
0
P+
P-
gene
B
number of TIR domains
800
600
400
200
0
TIR domains per genome
10
5
2
1
0
P+
P-
gene
cluster
B1_ADR2-WRR4
B3
B4_RLM1
B5
CHS3
DM1_SSI4
DM2_RPP1
DM4_RPP8
DM6_RPP7
DM8_RPP4-5
DSC1
LCD9
NRG1
RPP13
RPP2
RPS4
RPS5
RPS6
RSG2
SOC3
TTR1

Figure S5. Number of NB-ARC and TIR homologues by gene across all 64 accessions (left y-axis), grouped by closest Col-0 homologue, and ordered by position in genome, including a box plot with outliers in grey circles of copy number in each accession (right y-axis). Singletons are in grey. Black bars mark genes with identical NB-ARC sequences, red rings represent the number of accessions each gene is found in (left y-axis), and red numbers indicate the number of NB-ARC or TIR domains in genes with more than or fewer than 1 NB-ARC or TIR domain. Dashed line marks 1N = 64 and dotted line marks 2N = 128 along the left y-axis, and one copy and two copies of NB-ARC domains respectively per gene along the right y-axis, where N is the number of *A. thaliana* accessions surveyed.

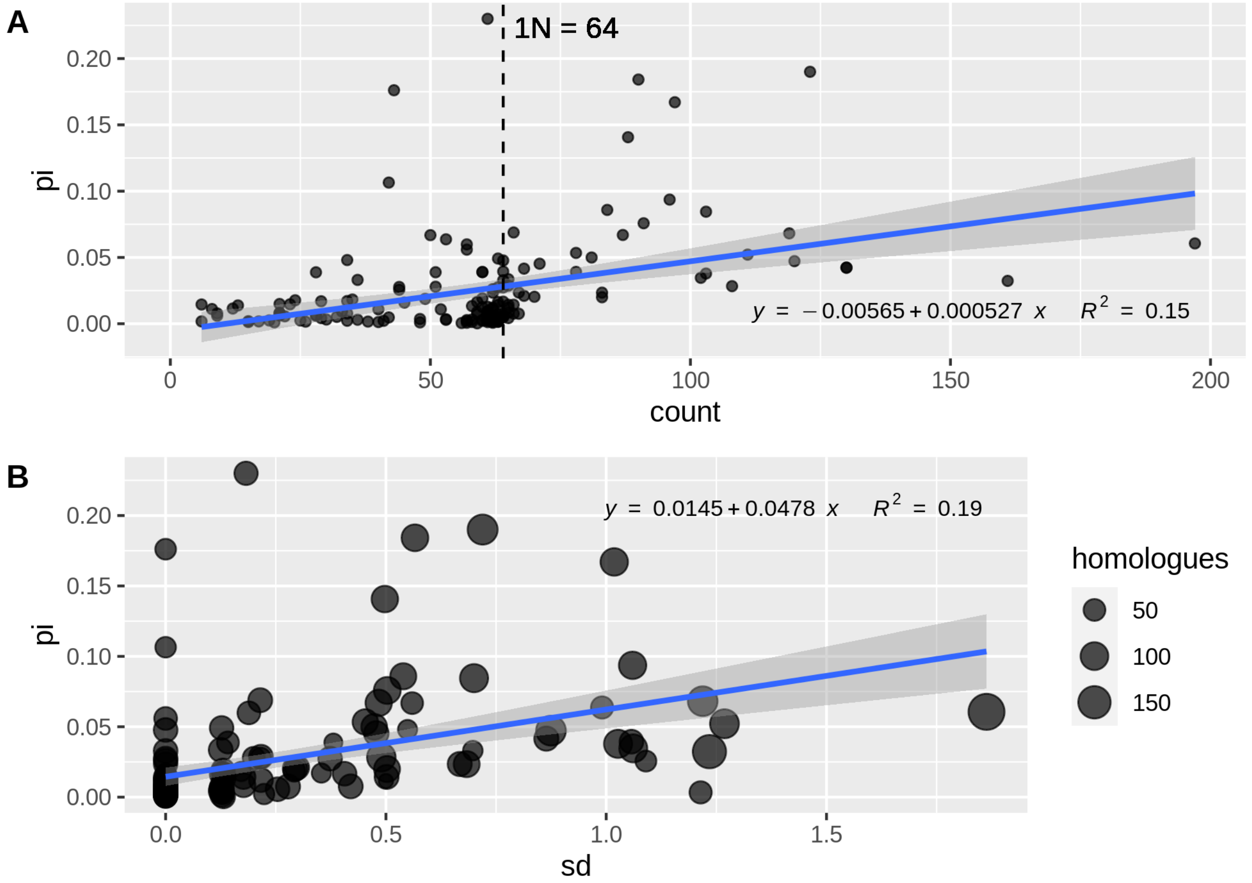

Figure S6. Nucleotide diversity of NB-ARC domains in *A. thaliana* NLRs from 64 accessions. A best fit linear model was calculated and plotted in blue, with the 95% confidence interval plotted as a grey band. **(A)** Pi as a function of number of homologues of each NB-ARC domain. **(B)** Pi as a function of standard deviation (sd) of number of homologues of each NB-ARC domain per accession, which reflects copy number variability. The size of each point in **B** reflects the number of homologues assigned to that NB-ARC domain across all 64 *A. thaliana* accessions.

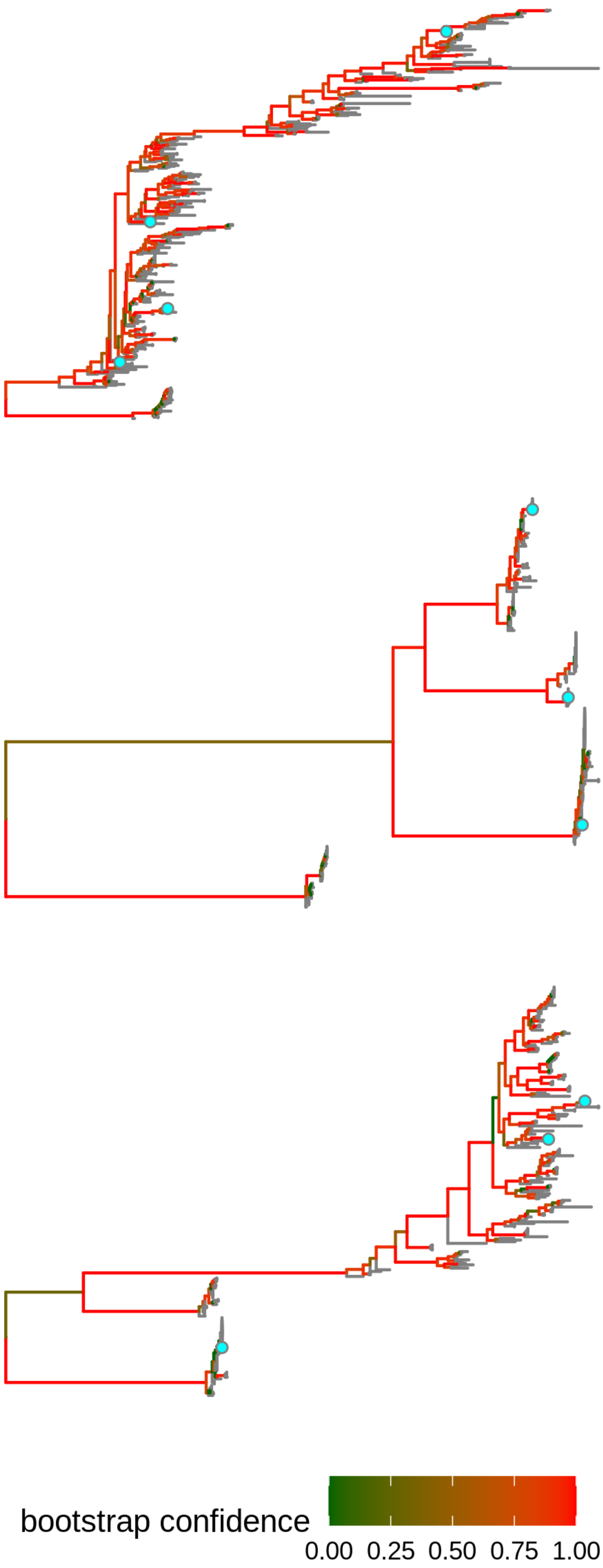

Figure S7. Bootstrap confidence of phylogenetic trees shown in Figure 6. Top to bottom: clusters *DM2*/*RPP1*, *RPP13*, and *DM4*/*RPP8*. Edges are coloured by bootstrap

confidence. Terminal edges without bootstrap values are in grey. Col-0 sequences are indicated with cyan circles.

Supplementary Methods

Bait selection

In addition to all 159 NLRs discovered by (Guo et al., 2011), the final bait set included: AT1G17920 and AT1G17930, two genes located physically within the B5 cluster that encode truncated NB-ARC domains reported to have unusual P-loop motifs (Bonardi et al., 2012); AT1G63860, a TIR-containing resistance gene known as *RLM1D* that lacks an NB-ARC domain in most accessions (including Col-0) and is found within the B1 cluster (Peele et al., 2014); AT5G45220, which lacks an NB-ARC domain but contains duplicated TIR domains and is located within the *RPS4* cluster (Meyers et al., 2002), and AT5G45490, a gene encoding both a coiled-coil domain and an NB-ARC domain (Tan et al., 2007). The *RPP13* cluster was defined by high sequence similarity between *RPP13*, which is primarily considered a singleton (Bittner-Eddy et al., 2000), and a nearby cluster of two genes consisting of AT3G46710 and AT3G4673 (Rose et al., 2004). The *DM2*/*RPP1* cluster was curated based on sequence similarities between the genes of two neighbouring clusters (Chae et al., 2014). The *DM4*/*RPP8* cluster was defined using Uniprot annotation that cited AT5G35450 and AT5G48620 as being RPP8-like.

Supplementary Results

B5 cluster P+ domain absence in more than half of accessions

AT1G72840, with twice the number of NB-ARC homologues as the other genes in the P+ group, may appear to be the exception, but an inspection of the domain tree of the B5 cluster revealed that two distinct paralogues were assigned to this gene due to the absence of the other paralogue in the reference Col-0 genome.

Explanation of specific trends in *DM4*/*RPP8* decay plot

The radiating clade (purple) splits off from the high-fidelity clades after the first iteration, resulting in the s.d. being halved from 0.00651 for the clade of the whole cluster to 0.00304 for just the radiating clade. The mean, which experienced only a slight drop from 0.00233 to 0.00227, is noticeably less affected by the split. Consistent with a clade that lacks distinct sub-clades, both mean and s.d. decay gradually after it split off from the rest of the cluster. On the other hand, the s.d. of the two high-fidelity clades (green) jumped briefly to 0.00949 due to the merging of the branch from the most basal division leading to the radiating clade and one high-fidelity clade with the branch leading to the high-fidelity clade that is sister to the radiating clade, before plummeting nearly 20-fold to 0.00053 and 0.00055 when the second iteration separated the two high-fidelity clades from each other. The mean of the high-fidelity clades exhibits the expected pattern of rapid decay to 0.00034 and 0.00046 after the second iteration when the clades are finally separated. Both mean and s.d. of high-fidelity clades remain low and decay gradually after the second iteration. Based on Fig. 6D, sequences assigned to AT5G35450 clearly form two distinct clades, and the sequences assigned to AT5G43470 and AT5G48620 are not easily separated from each other.

Supplementary References

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.

Bittner-Eddy, P.D., Crute, I.R., Holub, E.B., and Beynon, J.L. (2000). RPP13 is a simple locus in Arabidopsis thaliana for alleles that specify downy mildew resistance to different avirulence determinants in Peronospora parasitica. Plant J. *21*, 177–188.

Bonardi, V., Cherkis, K., Nishimura, M.T., and Dangl, J.L. (2012). A new eye on NLR proteins: Focused on clarity or diffused by complexity? Curr. Opin. Immunol. *24*, 41–50.

Chae, E., Bomblies, K., Kim, S.T., Karelina, D., Zaidem, M., Ossowski, S., Martín-Pizarro, C., Laitinen, R.A.E., Rowan, B.A., Tenenboim, H., et al. (2014). Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. Cell *159*, 1341–1351.

Guo, Y.-L., Fitz, J., Schneeberger, K., Ossowski, S., Cao, J., and Weigel, D. (2011). Genome-Wide Comparison of Nucleotide-Binding Site-Leucine-Rich Repeat-Encoding Genes in Arabidopsis. Plant Physiol. *157*, 757–769.

Meyers, B.C., Morgante, M., and Michelmore, R.W. (2002). TIR-X and TIR-NBS proteins: Two new families related to disease resistance TIR-NBS-LRR proteins encoded in Arabidopsis and other plant genomes. Plant J. *32*, 77–92.

Novikova, P.Y., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., Guggisberg, A., Paape, T., Schmid, K., Fedorenko, O.M., et al. (2016). Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat. Genet. *48*, 1077–1082.

Peele, H.M., Guan, N., Fogelqvist, J., and Dixelius, C. (2014). Loss and retention of resistance genes in five species of the Brassicaceae family. BMC Plant Biol. *14*, 1–11.

Rose, L.E., Bittner-Eddy, P.D., Langley, C.H., Holub, E.B., Michelmore, R.W., and Beynon, J.L. (2004). The Maintenance of Extreme Amino Acid Diversity at the Disease Resistance Gene, RPP13, in Arabidopsis thaliana. Genetics *166*, 1517–1527.

Tan, X., Meyers, B.C., Kozik, A., West, M. Al, Morgante, M., St Clair, D.A., Bent, A.F., and Michelmore, R.W. (2007). Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in Arabidopsis. BMC Plant Biol. *7*, 1–20.