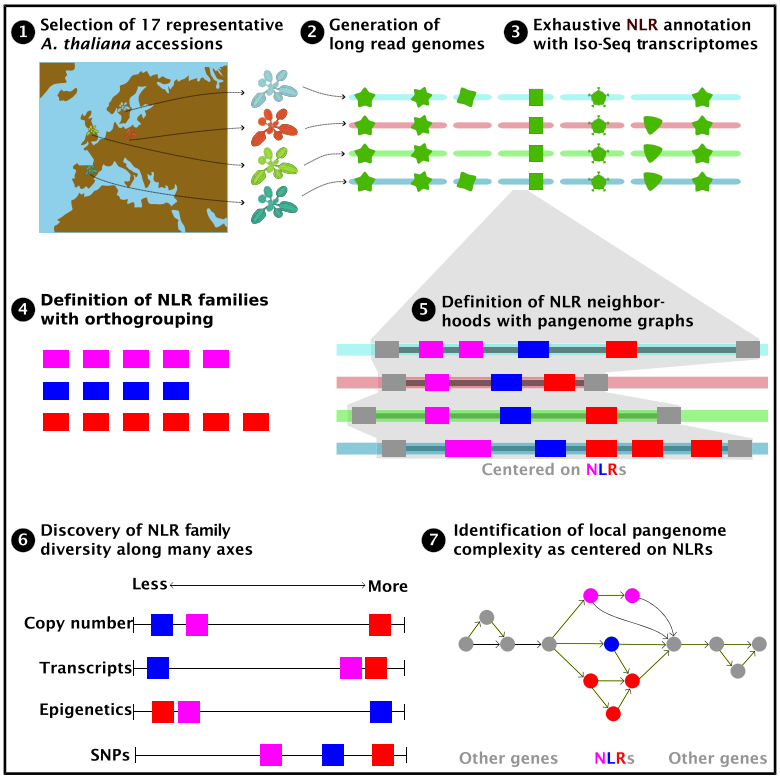
Resource

Cell Host & Microbe

Pangenomic context reveals the extent of intraspecific plant NLR evolution

Graphical abstract



Highlights

- Pangenome graphs enable nuanced analysis of NLR evolution in a genomic context
- Distinct evolutionary processes act on NLR neighborhoods defending biotrophic pathogens
- NLR diversity arises from multiple uncorrelated mutational and genomic processes
- Increased complexity in NLR neighborhoods centers specifically on NLRs

Authors

Luisa C. Teasdale, Kevin D. Murray, Max Collenberg, ..., Hajk-Georg Drost, Detlef Weigel, Gautam Shirsekar

Correspondence

weigel@tuebingen.mpg.de (D.W.), gshirsek@utk.edu (G.S.)

In brief

Individual- and population-level diversity is required for pathogen defense by nucleotide-binding site leucine-rich repeat (NLR) proteins. Teasdale et al. leverage annotated, divergent *A. thaliana* genomes and pangenome graph approaches to describe the genomic neighborhoods of NLRs, revealing evolutionary footprints in the form of "diversity in diversity" at these loci.

Teasdale et al., 2025, Cell Host & Microbe *33*, 1291–1305
August 13, 2025
https://doi.org/10.1016/j.chom.2025.07.011

CellPress

Cell Host & Microbe

CellPress
OPEN ACCESS

Resource

Pangenomic context reveals the extent of intraspecific plant NLR evolution

Luisa C. Teasdale,[1,6,9] Kevin D. Murray,[1,9] Max Collenberg,[1] Adrian Contreras-Garrido,[1,7] Theresa Schlegel,[1] Leon van Ess,[1] Justina Jüttner,[1] Christa Lanz,[1] Oliver Deusch,[1] Joffrey Fitz,[1] Regina Mencia,[1,2] Rosanne van Velthoven,[1,8] Hajk-Georg Drost,[1,3] Detlef Weigel,[1,4,10,*] and Gautam Shirsekar[1,5,*]

[1]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany
[2]Instituto de Agrobiotecnología del Litoral (CONICET-UNL), Universidad Nacional del Litoral, 3000 Santa Fe, Argentina
[3]Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee DD2 5DA, Scotland
[4]Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, 72076 Tübingen, Germany
[5]Department of Entomology and Plant Pathology, The University of Tennessee, Knoxville, TN 37996, USA
[6]Present address: Agriculture Victoria Research, Department of Energy, Environment and Climate Action, Bundoora, VIC 3083, Australia
[7]Present address: University of Lille, CNRS, 59000 Lille, France
[8]Present address: Genetwister Technologies B.V, 6709PA Wageningen, the Netherlands
[9]These authors contributed equally
[10]Lead contact
*Correspondence: weigel@tuebingen.mpg.de (D.W.), gshirsek@utk.edu (G.S.)
https://doi.org/10.1016/j.chom.2025.07.011

SUMMARY

Nucleotide-binding leucine-rich repeat (NLR) proteins are major components of the plant immune system, recognizing pathogen effectors and triggering defense responses. Because of the diversity of pathogen effector repertoires, NLRs have extraordinary sequence, structural, and regulatory variability. Although processes contributing to NLR diversity have been identified, the precise evolution of NLRs in their genomic context and along the multiple axes of diversity has been difficult to trace. We integrate genome-specific full-length transcript, homology, and transposable element information to annotate 3,789 NLRs in 17 diverse *Arabidopsis thaliana* accessions. We define 121 pangenomic NLR neighborhoods, which vary greatly in size, content, and complexity. NLRs are diverse across many axes, and multiple metrics are required to fully capture NLR variation. Based on these findings, we propose that diversity in diversity generation is fundamental to maintaining a functionally "adaptive" immune system in plants and that mechanistic studies should consider multiple axes of immune system diversity.

INTRODUCTION

All organisms must defend themselves against a multitude of adversaries, for which they use both physical and biochemical means. The latter often rely on detecting alien molecular signals that initiate countermeasures by the attacked host. While plants lack an adaptive immune system in the vertebrate sense, they benefit from extensive population-level genetic diversity of immune genes,[1] with networks of interacting sensors and downstream factors that detect infection and trigger defenses.[2,3] As a result, components of the plant immune system both evolve more rapidly than the genome at large but also maintain diversity for much longer than the rest of the genome, in co-evolution with multiple pathogens whose abundance and diversity fluctuate over space and time.[4–6] Because of a general trade-off between immunity and plant vigor, there are also inherent limits to the number and diversity of immune genes.[7] Therefore, selection on immune genes can be both strong and ephemeral at the same time, with successive waves of diversification and selection maintaining an evolutionary balance both between plants and their pathogens and between growth, defense, and avoidance of autoimmunity.

A major group of immune receptors are the NLR proteins (nucleotide-binding leucine-rich repeat proteins; alternatively NOD-like receptors), which detect effector proteins that help pathogens to evade or co-opt the primary, generalized stage of pattern-triggered immunity.[3] The diversity of NLRs appears to match the enormous diversity of pathogen effector repertoires.[8–10] NLR genes vary in their primary domain composition, with some active NLRs lacking several canonical NLR domains, and they can be found in complex and dynamic gene clusters or occur in stable paired or single gene configurations. Finally, not all NLRs are directly involved in pathogen recognition, and some function as executors of helper NLRs downstream of sensor NLRs.[11]

Early analyses noted high nucleotide-level diversity of individual NLRs[9] and highlighted the birth, divergence, and death of genes within clusters as potential key mechanisms underlying NLR evolution.[8] These patterns have been confirmed at increasing scales in the decades since,[10,12–14] including recent

insights into epigenetic and regulatory variation.[15–18] What remains less clear are the specific mechanisms by which plants generate and maintain functionally relevant NLR diversity.[11,19]

While NLRs often come to the fore when one looks for a high density of structural variants that disrupt synteny between closely related genomes,[20] one can only understand the evolution of the entire NLR family by considering both population diversity and the genomic context of all NLR genes. Here, based on evidence-based rigorous annotation and epigenetic profiles of the individual genomes, we characterize the NLR gene complement in genomic and population contexts across a diverse collection of 17 genomes that represent the range-wide haplotype diversity of *A. thaliana*. We used a pangenome graph-based principled approach for delineating the complex genomic regions surrounding NLRs, what we call "NLR neighborhoods." Importantly, our methods are robust to the rampant structural variation typical for many regions containing multiple NLRs. Using our pangenomic neighborhood approach, along with high-confidence annotations of NLRs, pseudogenes, and transposable elements (TEs), a much richer picture of immune system evolution emerges. Notably, NLR diversification and evolution defy classification by simple measures or metrics, and a comprehensive understanding requires a holistic view.

RESULTS

Defining the pangenomic context of NLRs

To characterize the dynamics of NLR evolution in *A. thaliana*, we selected 17 accessions to represent the genetic diversity across the Western Palearctic, based on geographic stratification and previously identified haplotype sharing groups[21] (Figure 1A). They represent a phenotypically diverse set of accessions based on their nested responses to 104 downy mildew disease-causing oomycete, *Hyaloperonospora arabidopsidis* (*Har*) isolates collected across the native range of *A. thaliana* in Europe (Figure 1C). We assembled and scaffolded contigs from PacBio HiFi reads into chromosomes with the aid of a BioNano optical map for one of the accessions (at9852). We annotated protein-coding genes using both homology and full-length cDNA expression evidence from *Har*-challenged plants (sequenced using PacBio isoform sequencing [Iso-Seq]) and annotated TEs using both homology and a curated collection of annotated repeat families. We annotated up to 29,321 protein-coding genes per accession (including pseudogenes but not including TE protein genes), with an average of 35.1 Mb of repeats, and a combined total of 3,789 NLR genes (Figure S1A). As expected, all genomes show a high degree of global synteny (Figure S1C).[22] We took an integrated approach to the annotation of the NLR genes, combining multiple sources of evidence including Iso-Seq data with manual curation to produce high-confidence NLR gene models (see STAR Methods and Figure S11 for full details). This annotation approach improves upon previous work that relied on single-accession references, target capture techniques, and short-read RNA sequencing (RNA-seq). We also manually annotated and verified NLR pseudogenes and partial NLR copies that could not be annotated by automated approaches, many of which likely have compromised activity, allowing us to examine mutational processes that inactivate NLRs.

Given that TEs and the epigenetic landscape can influence NLR function,[15,17] we aimed to take a principled approach that takes the entire genomic environment of NLRs beyond inter-NLR non-coding sequences into account. In addition, we wanted to include in our comparisons also NLR-free regions that were orthologous to NLR-containing regions in other accessions. To do so, we devised a method for consistent delineation of NLR-related regions that is robust to frequent presence-absence variation and that included variable flanking sequences. Compared with defining NLR clusters as regions that include a fixed amount of sequence up- and downstream of an NLR or clustering NLRs based solely on genomic proximity,[13,14,24] our approach guarantees that all structural variation is considered, including cases where NLRs are entirely absent from a region in some accessions. This allows us to study complete losses as well as *de novo* emergence of NLR loci.

We use the term NLR neighborhood for such regions that contain at least one NLR in at least one accession, bordered by pangenome-wide syntenic, conserved anchors that were determined with a pangenomic approach (Figure 1D). In a specific accession, an NLR neighborhood may contain no NLR ("nullitons"), one NLR ("singletons") or multiple NLR genes ("clusters"). These regions vary greatly in complexity in our dataset: in the simplest cases they consist of a single syntenic NLR across all accessions, but they can also be very large and diverse in NLR copy number across accessions.

Across all genomes, we identified 121 NLR neighborhoods. These generally are like other euchromatic regions with respect to protein-coding gene density and steady-state cytosine methylation (Figure S2). However, NLR neighborhoods tend to have a higher TE density, with on average younger TEs as inferred from divergence from the family consensus (Figure S2).

NLR neighborhood complexity and variability vary greatly

NLR neighborhoods differ greatly in average length, from 1.9 kb to over 900 kb, although most are smaller than 50 kb (Figure 2A). Similarly, while the length of some neighborhoods varies considerably between accessions, this is not the case for most (Figure 2A). The neighborhoods are widely variable in NLR content, ranging from neighborhoods with only NLR fragments present in just a few accessions, to neighborhoods with several NLRs that themselves differ greatly in size and sequence across accessions (Figure 2B). A numerous class, about 33%, comprises single-copy NLRs with highly conserved synteny.

There is also substantial variation in the relative proportions of NLRs, non-NLR protein-coding genes, and TEs making up NLR neighborhoods (Figures 2B and 2C). Notably, there is no simple relationship between the size of a neighborhood and the number of NLRs it contains. While larger neighborhoods on average have a higher fraction of TEs, potentially indicating that these represent a genomic context that is more tolerant to TE activity, many counterexamples exist (Figure 2C). The helper NLRs of the *Activated disease resistance 1* (*ADR1*) and *N requirement gene 1* (*NRG1*) families as well as loci conferring resistance to generalist bacterial pathogens such as *Resistance to Pseudomonas syringae pv. maculicola 1* (*RPM1*), *Resistance to Pseudomonas syringae 5* (*RPS5*), or *Resistance to Pseudomonas syringae 2* (*RPS2*), on average occur in smaller, less variable neighborhoods. The larger,

CellPress
OPEN ACCESS

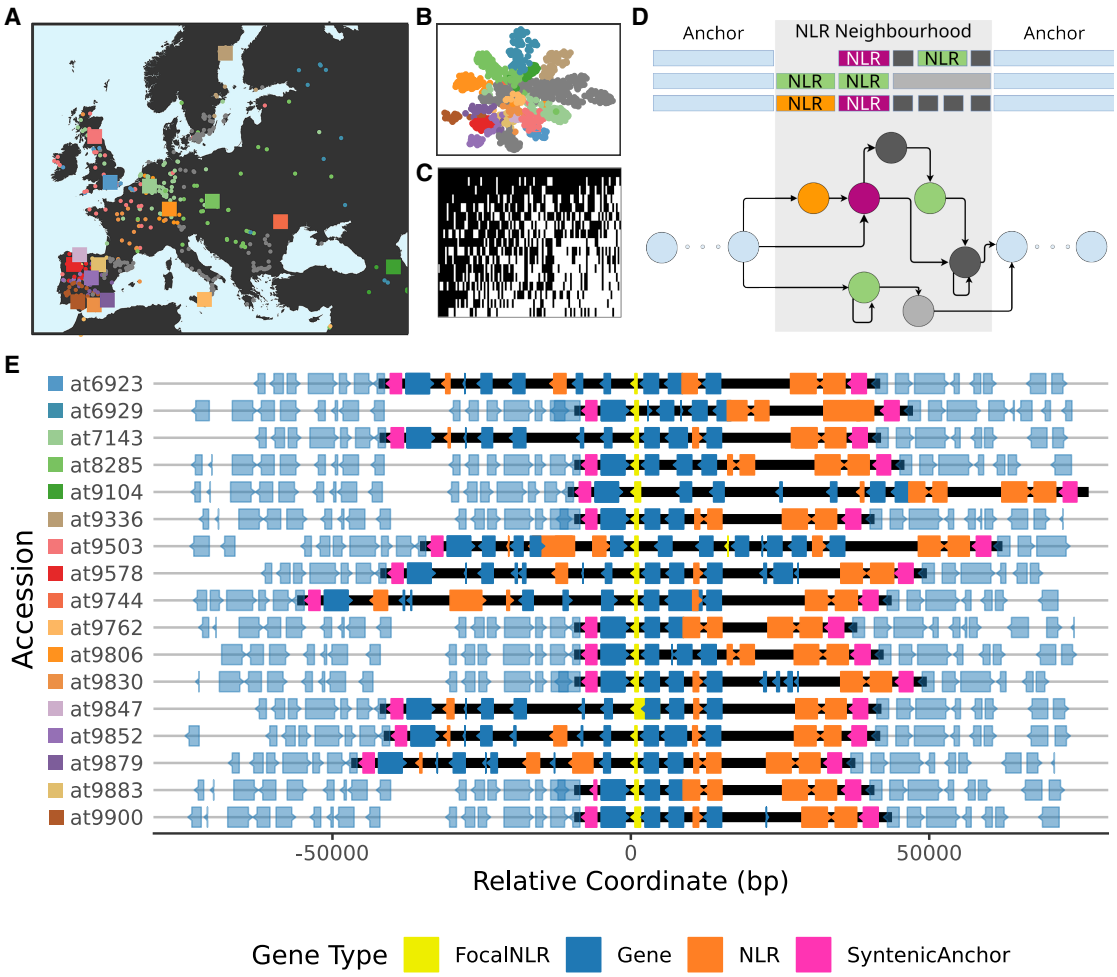


Figure 1. Defining NLR neighborhoods across 17 *Arabidopsis thaliana* genomes

(A and B) Our 17 accessions (squares) are broadly representative of both the geographic and population genetic distributions of the broader 1,001 Genomes collection (circles[23]); inset shows the UMAP embedding of 17 accessions in the co-ancestry space of all 1,135 accessions (adapted from Shirsekar et al.[21]). Colors in (A) and (B) represent the haplotype sharing groups from Shirsekar et al.[21]; samples in gray represent the three haplotype sharing groups not represented by our 17 accessions and their corresponding selected accession (see E).

(C) Phenotypic diversity of 17 accessions (rows) with respect to the oomycete pathogen *Hyaloperonospora arabidopsidis* (*Har*; columns), with resistant phenotypes denoted as black. 104 *Har* accessions were collected across Europe, as described in STAR Methods.

(D) Universal, pangenome-wide syntenic anchors define NLR neighborhoods.

(E) Representative example of an NLR neighborhood, highlighting accurate delineation of the borders of NLR-containing variable regions, regardless of structural or presence/absence variation. Black bars indicate pangenomic neighborhoods, syntenic anchors in pink. A fixed-sized window approach centered on a focal NLR (in yellow) would produce a very different outcome. Note accession codes used in this paper of the form atNNNN represent Arabidopsis ecotype IDs, see Alonso-Blanco et al.[23] and Table S1.

See also Figure S1.

more variable neighborhoods typically contain NLR genes that are co-evolving with highly specialized oomycete pathogens, such as *Resistance to Peronospora parasitica 1* (*RPP1*) and *RPP4/5*, which encode resistance to *Hyaloperonospora arabidopsidis*, but again, with notable exceptions: the neighborhood (chr2_nbh01) with the highest average length is defined by two NLR fragments in a sea of highly complex TE insertions.

Pangenomic complexity in NLR neighborhoods is centered on NLRs

The degree of variation and structural complexity varies not only among NLR neighborhoods but also within them. We quantified the local complexity within neighborhoods using a pangenome graph metric that summarizes the degree of structural variation around a sequence ("node") across all occurrences in the pangenome, what we call "node radius" (see STAR Methods; Figure S12). This metric represents the local connectedness of sequences within a pangenome graph. It increases especially with rearrangements, duplication, and translocations, as these increase the number of other sequences that can be reached from the focal node. In neighborhoods with increased local complexity, this diversity is often centered on the NLRs (Figures 2D and 3). This observation suggests that rather than diversity being a consequence of the surrounding genomic

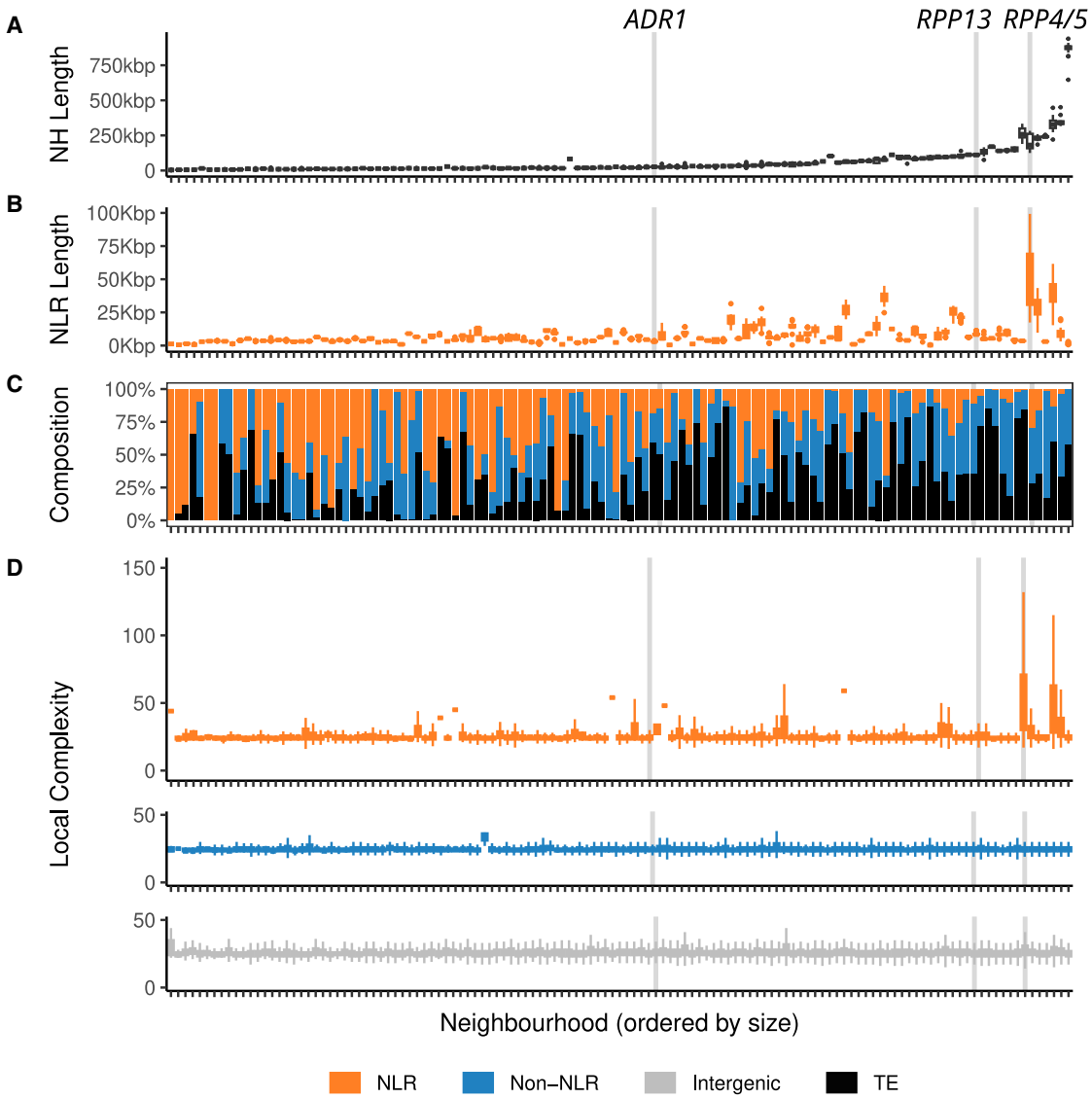


Figure 2. NLR neighborhoods across 17 genomes

(A–C) NLR neighborhoods vary across accessions in their overall length (A), their combined length of NLR sequences (B), and by the composition of their annotated space (C). See also Figure S2 (D) Pangenome local complexity ("node radius") varies within and between neighborhoods, and where elevated, is predominantly elevated in NLR genes themselves rather than non-NLR genes or intergenic space. Node radius is a genome graph-derived measure of local structural variation complexity, see Figure S12 and STAR Methods. In all subplots, neighborhoods are ordered by their mean length across accessions.

context—for example, because of DNA sequences that are particularly prone to double-strand breakage or attracting TE insertions—local complexity is associated with the duplication and rearrangement of the NLRs themselves. These patterns are exemplified by the *RPP4/5* locus, where a high degree of copy-number, domain, and single-nucleotide variation greatly increases our complexity metric over most of the numerous NLRs in this neighborhood, while TEs, including recent insertions, contribute less to pangenomic complexity (Figure 3). In contrast, in the *RPP13* neighborhood, conserved gene structure and limited copy-number variation only slightly increases local complexity, despite sequences encoding the leucine-rich repeat (LRR) domains of the NLRs varying greatly (Figure S3).

NLR diversity spans many axes

That local pangenomic complexity in NLR neighborhoods is often centered on the NLRs themselves and is consistent with rampant structural variation in many NLR genes. However, structural variation is only one of many forms of diversity. We used several complementary metrics to measure other axes of NLR diversification, such as pairwise nucleotide distance, population-wide frequency of high-impact mutations, isoform diversity, Shannon entropy of amino acid variation, diversity of NLR-associated domains (Figures S4 and S8), and copy-number variation (see STAR Methods). These metrics were calculated for homologous clusters of NLR sequences (see STAR Methods). Across these metrics, NLR gene families differ in both their main axes and degrees of

CellPress
OPEN ACCESS

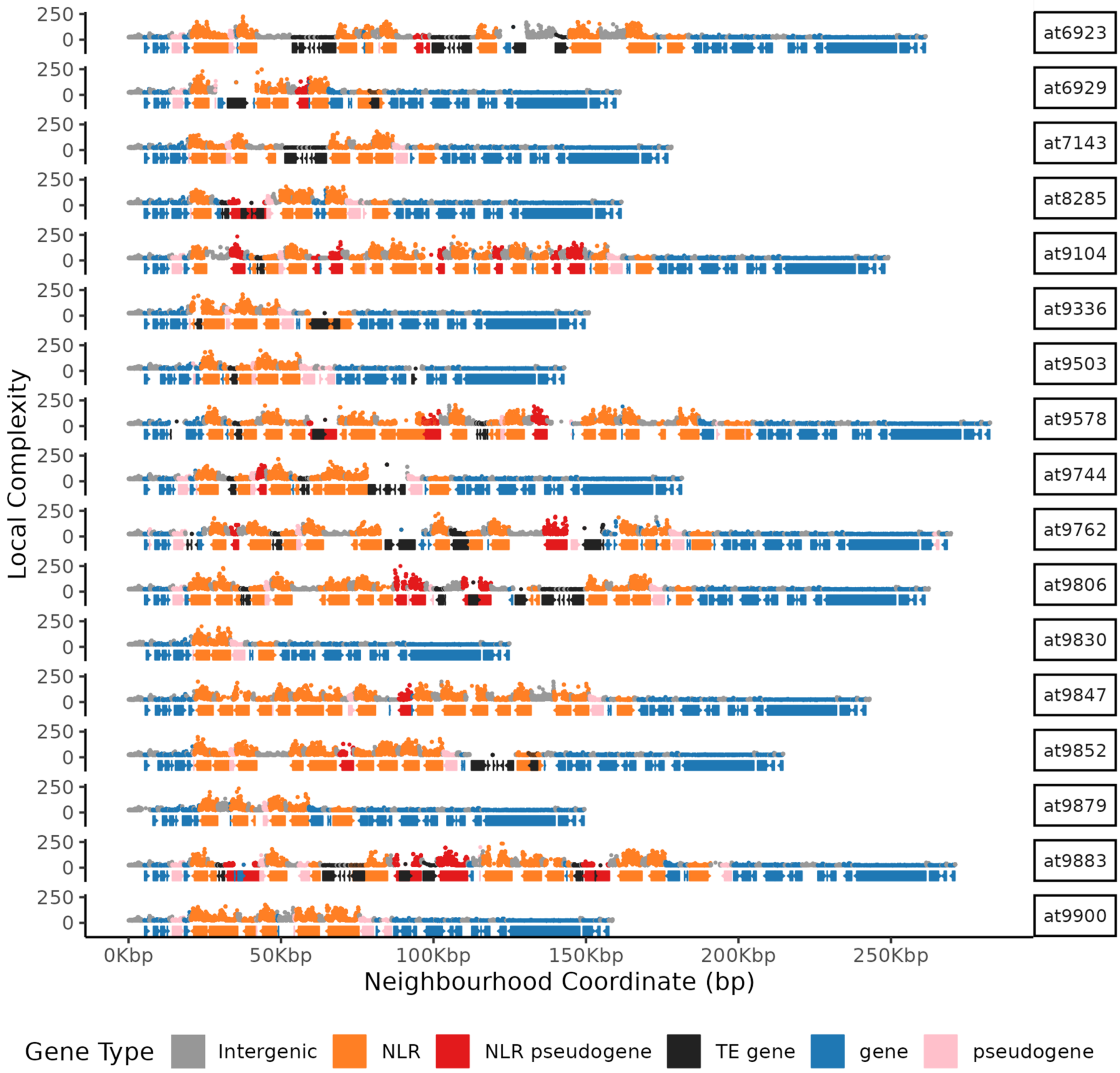


Figure 3. Pangenome complexity of the *RPP4/RPP5* NLR neighborhood

The neighborhood containing RPP4/RPP5 is among the most complex, and the highly elevated local complexity is focused on NLR genes and pseudogenes. Each per-accession track consists of both a trace of local pangenomic complexity (above the abscissae), and a representation of the gene annotation (below the abscissae). Individuals show extreme haplotypic diversity of NLRs, which is reflected in highly elevated local complexity focused on NLR genes and their immediate surrounds, while TE genes or other protein-coding genes show little elevation in complexity above the genome background. For an enlarged version contrasting two nearby neighborhoods, please see Figure S3.

diversity (Figures 4A and 4B). Importantly, no single metric captures the majority of NLR diversity on its own (Figures 4A, S4, and S5). This also applies to NLRs demonstrated to confer specific disease resistance, although they are among the more diverse NLRs by most metrics. For example, *RPP13* is highly diverse in sequence but not isoforms or copy number, whereas *DA1-related protein 4* (*DAR4*) shows the opposite pattern, while *RPP4* is consistently extreme in nearly all metrics (Figures 4A, 4B, and S4).

Transcript isoform variation is an under-appreciated axis of NLR diversity. It is only weakly correlated with domain diversity (adj. R^2 0.10, p = 1.1e−9) and gene length (adj. R^2 0.18, p = 1.3e−15) (Figure S6A). Instead, it is a specific property of individual NLR gene families (adj. R^2 0.45, $p < 2.2e{-}16$) (Figure S6B). Therefore, genes that are highly conserved based on genomic DNA-based metrics can still produce diverse proteins across accessions through the expression of alternative transcripts (e.g., the bacterial effector AvrRPS4 recognizing NLR gene *Rps4*). NLRs feature more isoform variation than non-NLR genes in the same neighborhood, indicating that this increased diversity is specific to NLRs rather than epigenetic cues affecting a broader genomic context (Figure S6D). However, we did find that methylation in NLR neighborhoods is most associated with TE presence rather than methylation of the NLRs themselves (Figure S8D).

Insights into how NLR diversity has been generated

Turning to the processes and possible molecular mechanisms generating NLR diversity, we first looked at the gain and loss

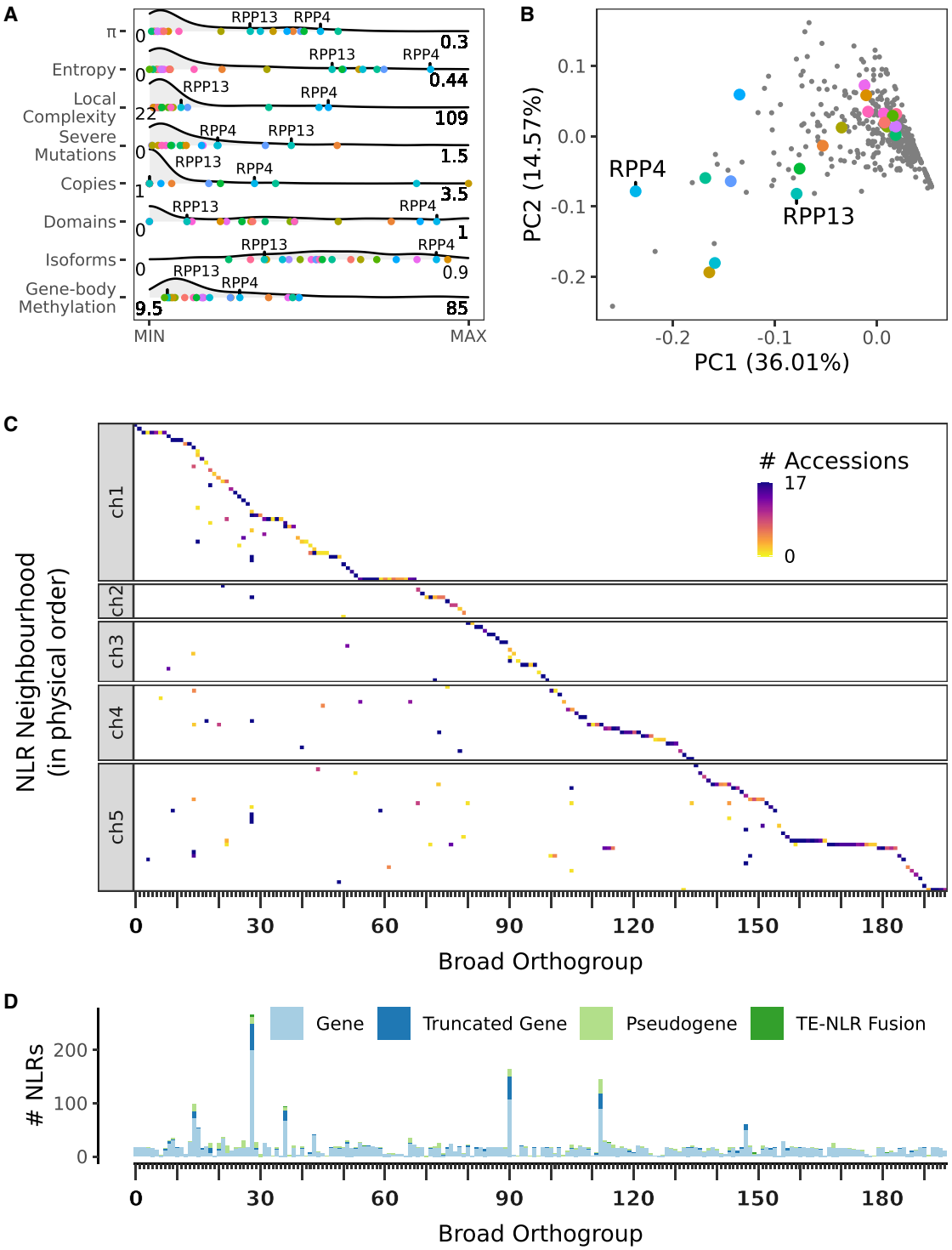


Figure 4. Linking genomic context to gene families

(A) Variation of OG70 diversity across different axes. If not stated otherwise, always for all OG70 members across all accessions (A). π, mean nucleotide pairwise distance; entropy, mean Shannon entropy of column states within an amino acid multiple sequence alignment, excluding positions with fewer than 10 non-gap characters; local complexity, mean pangenome local complexity; severe mutations, sum of allele frequencies of mutations with severe predicted consequences among the 1,001 Genomes collection for OG70s present in at9852; copies, mean number of copies per accession; domains, mean Simpson's index of diversity of NLR-associated domains detected by NLRtracker; isoforms, mean of Simpson's index of diversity of isoform variation affecting the open reading frame of each transcript; gene-body methylation, mean percentage of methylated CpG sites. Enlarged in Figure S4, and Figures S5–S8 explore each individual axis in more detail.

(legend continued on next page)

of NLR genes. We initially defined 204 more broadly related NLRs homologous clusters (broad OGs). A little under half (88) were split into more finely delineated groups with at least 70% protein sequence identity (OG70s), for a total of 371 OG70s (Figures 4C, 4D, S7D, and S7E). Some NLR neighborhoods with multiple OGs are likely evolutionarily old, as there are typically only single members of multiple OGs. In other cases, such as the *RPP1* and *RPP4/5* neighborhoods, copy-number expansions have likely occurred after speciation, and these neighborhoods generally contain few broad OGs, but with each being represented by multiple copies, indicative of an ongoing and active copy-number expansion process (Figure 4C). To be more specific, many NLR neighborhoods (45%) contain multiple broad OGs, but most broad OGs, 72%, are found in only one neighborhood (Figure 4C).

In terms of conservation of individual OGs and OG70s, fewer than half (47%) of broad OGs, and an even smaller fraction of OG70s (18%), are present in all accessions. Conversely, OG70s are more likely to be restricted to a single accession (18% of OG70s) than broad OGs (4%; Figures S7D and S7E). Some relatively rare OG70s represent long-range translocations of NLRs into new, distant neighborhoods and are likely recent and TE-mediated (Figures 4C and S10).

A special case is represented by neighborhoods with paired, yet highly divergent OGs that encode proteins forming obligate heteromers.[25] Complete duplication of such paired genes[26] and different pairs of alleles at the *Chilling sensitive 3-Constitutive shade-avoidance 1* (*CHS3-CSA1*) locus having different biochemical functions are indicative of functional linkage of paired NLR genes being under strong purifying selection. In agreement, tight genetic linkage between paired NLRs is maintained across accessions in our dataset. Finally, as in the *A. thaliana* reference genome, we did not find cases where coiled-coil NLRs (CNLs) and Toll/Interleukin-1 receptor (TIR)-NLRs (TNLs) share a neighborhood, something that has been observed in other species.[27]

An NLR neighborhood with clear examples of both NLR birth and death is the neighborhood containing *ADR2*. This neighborhood has multiple copies of a single broad OG, several of which have undergone independent duplication and pseudogenization (Figures 5A and 5B). Two separate pseudogenization events have occurred in this neighborhood, once in ADR2 homolog 2 and once in ADR2 homolog 4. In both cases intact copies still exist in some accessions. No accession has the full complement of all six homologs found in this neighborhood across accessions. Finally, there are examples of inversions and other structural variants generating novelty, such as at9879, where a TE is inserted into an NLR and the resulting new transcript encodes a fusion of the N terminus of the NLR with a DUF1985 domain containing portion of a VANDAL2 TE (Figure 5A). While just one of many examples in this dataset, this neighborhood in particular presents an ideal candidate for future studies investigating the functional consequences of gene duplication and pseudogenization, given the functional importance of ADR2 in the recognition of *Albugo* species.[28,29]

Pseudogenization is frequent but pseudogenized alleles rarely persist or spread

Compared with other genes, a major feature that makes NLR genes stand out is the large fraction of genes with mutations that interrupt the open reading frame (Figure S7C). Notably, genes that encode only some of the domains found in canonical NLRs can have biological function.[30,31]

Of all annotated NLR genes, 7.5% have a truncated open reading frame relative to allelic copies in other accessions, with single premature stop codons or frameshift mutations that are not fixed in the population. An additional 8.4% have multiple disabling mutations or produce transcripts that no longer encode any canonical NLR domain (Figure 4D). Genes in the latter class are more likely to have only limited expression compared with the genes with single large-effect mutations (Figure S6E), consistent with these genes being further along the path to pseudogenization. The causes of pseudogenization include a variety of mutational processes, such as single-nucleotide variants and small indels causing in-frame stop codons, as well as, more rarely, TE insertions or larger deletions (Figure S7B). Unsurprisingly, proximity to TEs significantly increases the odds of a gene being pseudogenic, both genome-wide (odds ratio = 2.3; 95% CI 2.25–2.37; $p < 2e{-}16$) and even more so for NLRs adjacent to TEs (odds ratio = 3.01; 95% CI 2.37–3.86; interaction term p 0.01). Importantly, this does not establish causality: are NLRs near TEs more likely to be pseudogenized, or are TEs more likely to insert or persist near pseudogenized NLRs? Single-copy conserved NLRs are less likely to be pseudogenized (Figure 4D), suggesting that, as a group, they are more likely to be essential.

While NLRs seem to have a higher burden of severe mutations than non-NLR genes, individual deleterious variants are most often found in only a single accession. To more accurately estimate the population frequency of large-effect mutations, we called single-nucleotide variants from short-read data of 1,135 individuals from the 1,001 Genomes Project[23] against one of our annotated accessions (at9852). The combined occurrence (the product of allele frequency and number of alleles) of variants with predicted severe consequences varies considerably among NLRs, with some genes with known roles in disease resistance, such as *RPP13*, having an increased burden of large-effect mutations (Figure 4A). This burden is also higher for NLRs than for other protein-coding genes (Figure S7C), albeit it only marginally, and it is not strongly correlated (R^2 = 0.065) with the overall sequence diversity within NLR OG70s (Figure S5).

(B) Principal-component analysis (PCA) of OG70 diversity across each of the axes from (A) highlights that diversity varies in both degree and class between orthogroups. In both (A) and (B), colored dots represent NLR genes previously identified as encoding resistance to a specific pathogen (see Table S2), which on the whole are distributed throughout the range of each diversity metric (A) and are relatively evenly distributed throughout the multivariate diversity space (B).
(C) The distribution of broad orthogroups across the 121 NLR neighborhoods, ordered by chromosome positions (*y* axis), and the first occurrence of the orthogroup (*x* axis). Numbers on the *x* axis refer to the row number in Table S3. See also Figure S9 for an enlarged version with neighborhoods and orthogroups names on axes, which can be used to reference individual orthogroups or neighborhoods.
(D) Total gene count per orthogroup across accessions, classified by pseudogenization state, ordered as in (C). Please note that (C) and (D) share *x* axes, ordered by the genome position of the first occurrence of the orthogroup, thus (D) represents column sums of (C). Enlarged in Figure S9.

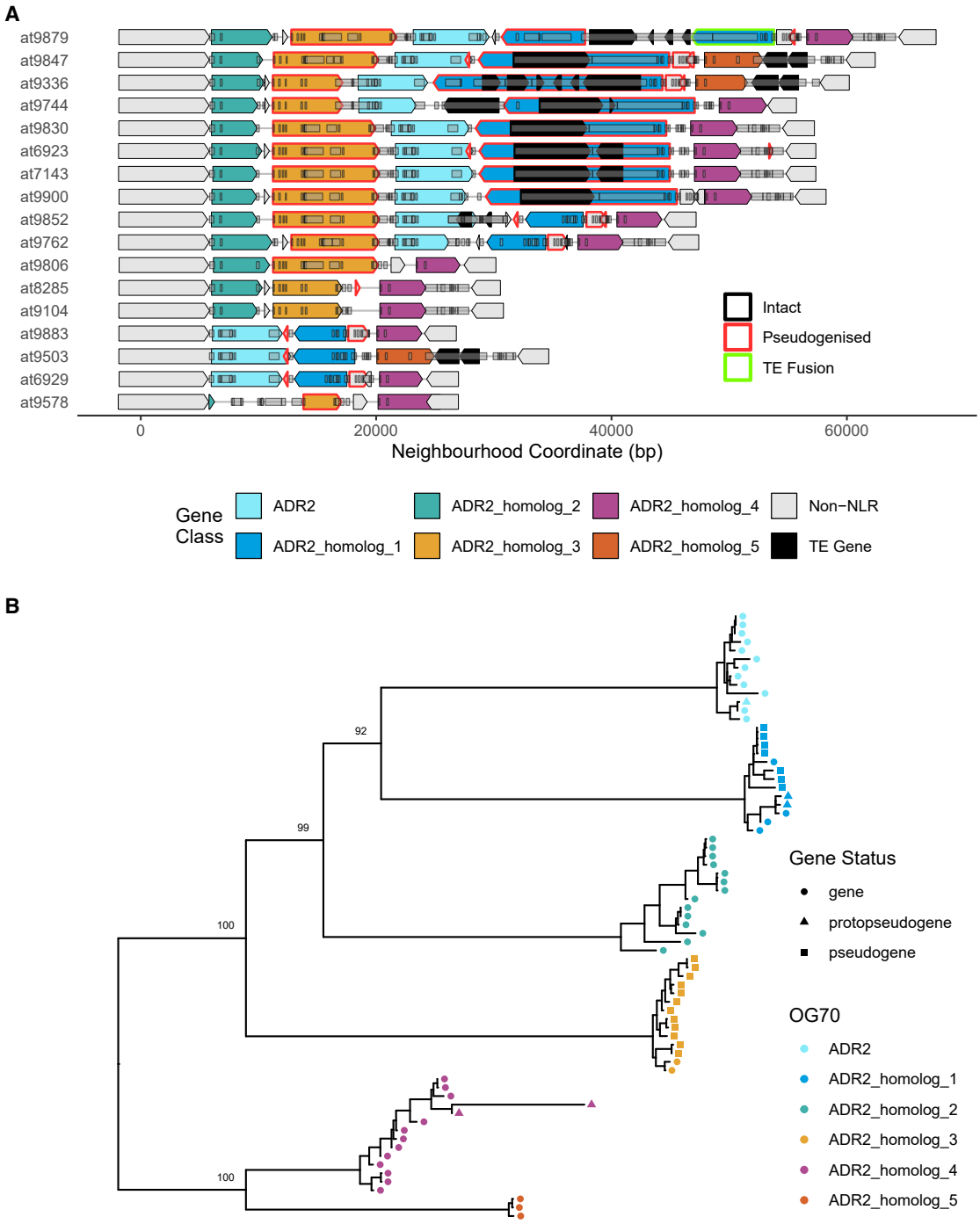


Figure 5. *ADR2* as an example of a neighborhood with birth and death of NLRs

(A) A local map of NLR genes in the neighborhood in each accession, with genes colored by OG70 membership. While some OG70s are intact in most accessions (ADR2, ADR2 homolog 4), others show presence/absence variation (ADR2 homologs 2 and 5), and/or frequent pseudogenization (ADR2 homologs 3 and 1). Most accessions have 2–3 intact NLRs and 1–2 pseudogenized NLRs, however some accessions have heavily reduced haplotypes containing no pseudogenes (e.g., at9578 and at6929).

(B) A gene tree of NLRs in this neighborhood, with annotated pseudogenization events. The gene tree shows that there have been at least two separate pseudogenization events within the neighborhood within both ADR2 homologs 3 and 1. For further examples of neighborhoods undergoing NLR birth and death, see Figure S10.

NLR neighborhoods are enriched for recent TE activity

TE sequences occur more frequently in NLR genes than in non-NLR genes. Genome-wide, just over half of intact NLR genes contain at least one TE sequence, while it is closer to a quarter for non-NLR genes. While, judging by their size, most TE sequences in NLR genes appear to be incomplete, the difference between NLR and non-NLR genes also holds when considering only intact TE insertions, with 3.9% of NLRs and 2.6% of non-NLR genes containing intact TEs, and when considering only TEs with clear homology to annotated TEs in Col-0 (NLRs, 33.1%; non-NLRs, 14.5%). While in some cases, TEs within NLRs are conserved across accessions, in most cases the TEs within an NLR vary across accessions (Figure S8B).

NLR neighborhoods are enriched for TEs, containing a wide variety of TE families. TEs within NLR neighborhoods are on average younger than outside NLR neighborhoods, as judged by formal models for dating intact LTR transposons[32] (Figure S2). We see evidence of rare long-range duplications of NLRs, likely mediated by copy-paste TE activity: at least 12 neighborhoods contain NLRs or NLR pseudogenes in only a few accessions, are TE-rich, and are not near any other NLR neighborhoods (Figures S10A–S10C). While these could represent NLR neighborhoods that have lost NLRs in all but a few accessions, the ages of linked TEs suggest that these NLR neighborhoods are most parsimoniously the result of copy-paste activity of TEs that moved partial or intact NLR genes. The presence of multiple homologous TE sequences is furthermore likely to provide a substrate for NLR gene duplication that is independent of TE copy-paste activity, but rather the result of localized processes such as replication slippage and illegitimate recombination. There are many examples where TEs seem to be directly responsible for the removal of NLRs including ADR1 a nominally essential helper NLR that is missing from one of the accessions (Figure S10D).

DISCUSSION

For three decades, studies of plant NLRs, a major class of immune receptors, have often stressed their evolutionary fluidity. This emphasis on structural and sequence variability of NLRs is not surprising, given that the functional identification of NLRs in many cases began with naturally occurring genetic variation. More recent work has suggested that extreme intraspecific diversification is a hallmark of a special class of hypervariable NLRs, while many other NLRs are evolutionarily much more stable. Here, by considering both structural and sequence variation, we show that highly conserved NLRs and hypervariable NLRs represent merely the extremes of a continuum, and that NLR diversity itself is much more complex than previously recognized. A major question for the future will be how representative our discoveries are for other types of highly variable gene families and/or immunity regulators, both in plants and in animals.

Context uncovers the extent of NLR evolution

Previous studies identified highly complex regions of the genome, a subset of which contain NLRs,[20] or studied NLRs either without genomic context,[12] or in the context of a single individual.[17] By exhaustively annotating intact and degraded NLRs in their pangenomic neighborhood contexts, we highlight that NLR loci vary considerably in structural complexity, with the majority of NLRs existing in relatively structurally conserved regions. A major surprise was that, when NLR neighborhoods are structurally complex, the complexity seems primarily centered on the NLR genes themselves. While not definitive proof, this suggests that NLRs are more likely the local drivers of structural diversification, rather than NLRs being mere passengers in regions of the genome that are generally complex.

While plants lack the somatically adaptive immune system common to vertebrates, one could consider the highly dynamic nature of NLRs in structurally complex neighborhoods to provide a similar vast adaptability as the vertebrate adaptive immune system.[33] Obviously, this analogy only holds with the proviso that plant NLR diversity and therefore adaptability plays out at the population level and is driven—at least in short-lived plants—by germline mutations rather than somatic variation that arises within individuals. It is obvious that somatic recombination and hypermutation of the loci encoding vertebrate immune receptors is the outcome of selection for a maximally effective immune system of the individual.[34] It is difficult to think that the same logic does not apply to plant immune receptors, even though the generational scale of diversification is a different one. Indeed, several—though not all—of the structurally most complex NLR neighborhoods harbor genes that are known to confer resistance to a highly co-evolving obligate biotrophic oomycete pathogen, *Hyaloperonospora arabidopsidis*.[24,35] There is, however, a fine line between adaptation and maladaptation, and several of these neighborhoods also harbor genetic loci that can cause autoimmunity.[13,20,36]

The diversity of NLRs cannot be generalized

While a few NLRs are at the same time highly diverse across multiple axes, many more NLRs have elevated diversity according to only one or a few of our seven metrics. Focusing on confirmed resistance genes, we find that they span the spectrum of diversity across all axes. Some resistance genes that are co-evolving with pathogen effectors that they directly recognize achieve extreme diversity in many axes, with many copies and a high degree of sequence variation (e.g., *RPP1* and *RPP4/5*); this is often seen as the prototypical pattern for resistance genes.[13,37,38] Most resistance genes, however, have elevated diversity in only some axes: *RPP13* exhibits a large degree of variation in its primary sequence but is otherwise a single-copy gene with comparatively low diversity along all other axes.[39] The sequence of *RPP7* is relatively conserved across accessions, with limited copy-number variation but with many independent inactivating mutations across the wider population. While one can derive simplifying metrics for NLR diversity,[17] the use of multiple metrics clearly provides a much more nuanced picture of NLR diversity. That said, single metrics have promise in highlighting candidates for bioengineering.[18] Critically, as discussed below, given our metapopulation-scale sampling of only tens of individuals, we can only form parsimonious hypotheses as to the true mechanistic underpinning of each axis of diversity, rather than tie each axis to a (combination of) specific mutagenesis mechanism(s). While it is important to understand the constraints of NLR diversification imposed by mutational mechanisms, a pathogen does not care if its receptor is disabled by point mutation, deletion, or TE insertion, only that it can now invade unrecognized.

The many axes of NLR diversification

The diversity observed at any locus is the product of local mutation rates and the specific selective landscape acting on a locus, and genomic context improves our understanding of the processes generating and maintaining NLR diversity.[12–14,17] For example, gene duplications occur more readily where local segmental duplications already exist.[40] Such duplicated copies of an NLR gene could allow additional recognition specificity to evolve, or could pose increased costs, for example, illegitimate oligomerization that impedes proper molecular functionality.[41] Whether these differences result from the underlying DNA sequence biasing mutational processes, or whether it is imposed through selection is less easy to determine. In addition, while we have treated the different axes of diversity as being largely independent, a more thorough understanding of the underlying mutational mechanisms in the future may reveal that some mutational processes lead to multiple outcomes. For example, double-strand breaks can be resolved in many ways, leading to point mutations, deletions, or complex rearrangements.[42] TEs in turn can induce double-strand breaks[43] and could therefore in principle be causal for a multitude of mutation types. Uncovering the true molecular mutagenic mechanisms underlying NLR diversification presents a challenge due to their incredible diversity, especially in retrospective range-wide studies such as this one. Future studies using regional time-series sampling of NLR diversity could make significant headway on this problem.

While NLR loci are well-known hotspots of gene duplication, inactivation, and loss,[8] relatively few NLR orthogroups are found in multiple neighborhoods, suggesting that most copy-number expansion occurs locally,[44] although we observe at least 12 apparently TE-associated distal duplications of NLRs to neighborhoods that did not previously contain an NLR. The causes of NLR inactivation by premature stop codons or frame shifts generally reflect the spectrum of mutations disabling non-NLR genes in *A. thaliana*.[45,46] Taken at face value, this would seem to indicate that inactivating mutations or pseudogenization are more common in NLR neighborhoods than elsewhere because of a greater tolerance for NLR mutations. The punctuated nature of pathogen selection may lead NLR neighborhoods to frequently experience episodes during which they are less constrained by purifying selection, and consequently we see mutations that would otherwise have been purged, especially when intact NLRs are mildly deleterious in the absence of pathogens. Whether this is unique for NLR regions or extends to other regions of the genome that contain non-essential genes will be an interesting topic for future research.

While NLR neighborhoods are enriched for TEs, TEs appear to play only a minor direct role in both gene duplication and pseudogenization events, even though they tend to be young and turn over quickly in NLR neighborhoods. We do not know whether this is due to increased rates of TE insertion, or to increased persistence of TEs, possibly due to fluctuating selection dynamics whereby TE insertions are tolerated in the absence of a pathogen, then swept to high frequency as passengers on selectively advantageous NLR haplotypes in the presence of a pathogen. It is also possible that TEs are positively selected in some NLR neighborhoods as they impact the regulatory landscape.[47,48] TE insertions likely play some role in immune system maintenance: a relatively simple example is the *CHS3/CSA1* paired NLR locus, where two distinct haplotypes are maintained across the population,[49] likely aided by the presence of highly divergent intergenic TEs that prevent recombination between the two NLRs. We also frequently observe multiple copies of NLRs and adjacent TEs within an NLR neighborhood, possibly due to segmental duplication by replication slippage or illegitimate recombination, which while mechanistically independent from TE activity may be accelerated by additional TE copies.[40]

Alternative splicing increases transcript diversity, and there are examples of isoform variation generating immune diversity in plants.[50,51] We found that individual NLR genes vary in isoform diversity, but that NLR genes from the same orthogroup have similar isoform diversity. While our study design does not investigate changes in isoform diversity during the course of infection, the observation of frequent instances of alternative splicing makes it likely that this is an important additional axis of NLR variation. NLR expression has been shown to vary among accessions[52] and across environmental clines,[53] further highlighting the need to consider regulation of expression in studies of functional NLR diversification. Similarly, structurally complex loci around NLRs are highly diverse in methylation patterns across accessions,[16] highlighting the need to consider both inter-accession variation, as well as the dynamics of methylation around NLRs over the course of pathogen invasion and defense, which our single time point data cannot evaluate.

To efficiently explore present-day diversity, we deliberately chose accessions as single representatives of regional populations, between which recent gene flow and therefore recombination is limited. An obvious next step will be a careful comparison of NLR neighborhoods across multiple populations with closely related individuals. This will allow, for example, assaying population recombination rates and to determine exactly how they are affected by structural variation.[54]

Plant immunity is not adaptive at the level of individuals but at the level of populations

We posit that ‘‘diversity in diversity’’ allows for evolutionary innovation over different time scales, enabling plants to keep pace with pathogens that themselves evolve rapidly by processes that range from two-speed genomes to the formation of minichromosomes and horizontal gene transfer.[55,56] Like all organisms, plants must survive the onslaught of a highly diverse and ever-changing set of pathogens and must also constantly balance the detection of novel pathogen signals with fitness costs that come from a hyperactive immune system—a known consequence of mal-adaptive NLR diversity.[36,57] Critically, as long as enough of the plants in a population have the right immune genes to survive that year’s wave of pathogens with only a minority of individuals having either too little or too much immunity, the population can persist. Frequency-dependent or balancing selection is a well-established concept in plant-pathogen dynamics, however the number of NLR loci for which balancing selection has been formally demonstrated is surprisingly small,[6,9,58–61] which is not surprising given the excessive diversity, including structural diversity, at NLR loci. The type of data that we have begun to present here will hopefully provide an incentive for the development of innovative population genetic approaches to prospect for signals of

balancing selection at many more NLR loci. Such studies should, of course, not be limited to NLRs, as the phenomenon of rapid coevolutionary diversification that we explore here with regards NLRs likely extends to other immune or highly diverse gene families. Another interesting question for the future is how all this plays out in long-lived plants such as trees, which will encounter a much greater diversity of pathogens over their lifetime than an ephemeral annual herb such as *A. thaliana*, and whose population-level adaptability iterates over decades rather than annually. Currently, we can only speculate whether trees might generate greater intra-individual NLR diversity through somatic mutation arising in continually dividing stem cells[62] or epigenetic variation[63,64] or whether trees rely more on NLR-independent broad-spectrum resistance.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Detlef Weigel (weigel@tue.mpg.de).

Materials availability

Seeds of *Arabidopsis thaliana* accessions used here are available from Nottingham Arabidopsis Stock Centre (NASC)/Arabidospsis Biological Resource Center (ABRC), as linked in the key resources table. *Hyaloperonospora* isolates may be available on request from the lead contact, Detlef Weigel.

Data and code availability

- Raw sequence data are publicly available as of the date of publication at the European Nucleotide Archive, ENA: PRJEB91362.
- Processed data are publicly available as of the date of publication at Zenodo: https://doi.org/10.5281/zenodo.15828128.[65]
- All original code is publicly available at GitHub: https://github.com/coevolutionlab/teasdale-2025-public and Zenodo: https://doi.org/10.5281/zenodo.15828128,[65] as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank Gal Ofir, Markéta Vlkova-Žlebková, Miriam Lucke, Sebastian Vorbrugg, Yueqi Tao, Fabrice Roux, Andrea Movilli, Fernando Rabanal, and Haim Ashkenazy for scientific discussions and/or commenting on the manuscript. We thank Katrin Fritschi and Anette Habring for assistance with preparing Iso-Seq libraries. We thank Anton Schölkopf and the students of the 2021 PBC practical at the University of Tubingen for assistance with manual NLR annotation. Computations were performed at the Max Planck Computing and Data Facility, Garching, Germany. We thank Marie Skłodowska-Curie Actions (to K.D.M.), DFG-TRR356 PlantMicrobe, the DFG-funded Excellence Cluster *Control of Microorganisms to Fight Infections* (CMFI), the Novozymes Prize of the Novo Nordisk Foundation, and the Max Planck Society (to D.W.) for research funding.

AUTHOR CONTRIBUTIONS

G.S., D.W., H.-G.D., L.C.T., and K.D.M. conceived of and designed this study; L.C.T., K.D.M., M.C., A.C.-G., T.S., L.v.E., J.J., C.L., O.D., J.F., R.M., R.v.V., and G.S. performed experiments and/or computational and statistical analyses; L.C.T., K.D.M., G.S., A.C.-G., H.-G.D., and D.W. wrote this manuscript, and all authors reviewed it.

DECLARATION OF INTERESTS

D.W. holds equity in Computomics, which advises plant breeders. D.W. has also consulted for KWS SE, a globally active plant breeder and seed producer. J.F. is an employee of Tropic TI, Lda.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Accession selection and plant growth
  - Phenotyping of Ath-Har interactions
- METHOD DETAILS
  - HMW DNA Extraction
  - PacBio HiFi Library Preparation
  - HiFi Reads and DNA Methylation
  - Long-read transcript evidence
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Genome assembly and quality assessment
  - *Ab initio* and homology-guided gene annotation
  - Gene annotation with transcript evidence
  - Evidence-weighted gene prediction
  - Isoform diversity
  - Repeat annotation
  - TE gene annotation
  - Orthogrouping and phylogenetic inference
  - Pangenome graphs
  - Definition of NLR-dense neighborhoods
  - Genotyping with data from the 1001 Genomes project
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2025.07.011.

Received: December 12, 2024
Revised: May 12, 2025
Accepted: July 15, 2025
Published: August 13, 2025

REFERENCES

1. Brown, J.K.M., and Tellier, A. (2011). Plant-parasite coevolution: bridging the gap between genetics and ecology. Annu. Rev. Phytopathol. *49*, 345–367. https://doi.org/10.1146/annurev-phyto-072910-095301.
2. Dodds, P.N., and Rathjen, J.P. (2010). Plant Immunity: Towards an Integrated View of Plant Pathogen Interactions. Nat. Rev. Genet. *11*, 539–548. https://doi.org/10.1038/nrg2812.
3. Jones, J.D.G., Staskawicz, B.J., and Dangl, J.L. (2024). The plant immune system: From discovery to deployment. Cell *187*, 2095–2116. https://doi.org/10.1016/j.cell.2024.03.045.
4. Thrall, P.H., and Burdon, J.J. (2003). Evolution of virulence in a plant host-pathogen metapopulation. Science *299*, 1735–1737. https://doi.org/10.1126/science.1080070.
5. Thompson, J.N. (2005). The Geographic Mosaic of Coevolution (University of Chicago Press). https://doi.org/10.7208/chicago/9780226118697.001.0001.
6. Koenig, D., Hagmann, J., Li, R., Bemm, F., Slotte, T., Neuffer, B., Wright, S.I., and Weigel, D. (2019). Long-term balancing selection drives evolution of immunity genes in Capsella. eLife *8*, e43606. https://doi.org/10.7554/eLife.43606.
7. Karasov, T.L., Chae, E., Herman, J.J., and Bergelson, J. (2017). Mechanisms to Mitigate the Trade-Off between Growth and Defense. Plant Cell *29*, 666–680. https://doi.org/10.1105/tpc.16.00931.
8. Michelmore, R.W., and Meyers, B.C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. *8*, 1113–1130. https://doi.org/10.1101/gr.8.11.1113.

9. Bakker, E.G., Toomajian, C., Kreitman, M., and Bergelson, J. (2006). A Genome-Wide Survey of R Gene Polymorphisms in Arabidopsis. Plant Cell *18*, 1803–1818. https://doi.org/10.1105/tpc.106.042614.

10. Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A., et al. (2007). Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science *317*, 338–342. https://doi.org/10.1126/science.1138632.

11. Barragan, A.C., and Weigel, D. (2021). Plant NLR diversity: the known unknowns of pan-NLRomes. Plant Cell *33*, 814–831. https://doi.org/10.1093/plcell/koaa002.

12. Van de Weyer, A.-L., Monteiro, F., Furzer, O.J., Nishimura, M.T., Cevik, V., Witek, K., Jones, J.D.G., Dangl, J.L., Weigel, D., and Bemm, F. (2019). A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. Cell *178*, 1260–1272.e14. https://doi.org/10.1016/j.cell.2019.07.038.

13. Prigozhin, D.M., and Krasileva, K.V. (2021). Analysis of intraspecies diversity reveals a subset of highly variable plant immune receptors and predicts their binding sites. Plant Cell *33*, 998–1015. https://doi.org/10.1093/plcell/koab013.

14. Lee, R.R.Q., and Chae, E. (2020). Patterns of NLR cluster variation in Arabidopsis thaliana genomes. Plant Commun. 100089.

15. Tsuchiya, T., and Eulgem, T. (2013). An alternative polyadenylation mechanism coopted to the Arabidopsis RPP7 gene through intronic retrotransposon domestication. Proc. Natl. Acad. Sci. USA *110*, E3535–E3543. https://doi.org/10.1073/pnas.1312545110.

16. Kawakatsu, T., Huang, S.C., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M. A., Castanon, R., Nery, J.R., Barragan, C., He, Y., et al. (2016). Epigenomic Diversity in a Global Collection of Arabidopsis Thaliana Accessions. Cell *166*, 492–505. https://doi.org/10.1016/j.cell.2016.06.044.

17. Sutherland, C.A., Prigozhin, D.M., Monroe, J.G., and Krasileva, K.V. (2024). High allelic diversity in Arabidopsis NLRs is associated with distinct genomic features. EMBO Rep. *25*, 2306–2322. https://doi.org/10.1038/s44319-024-00122-9.

18. Brabham, H.J., Hernández-Pinzón, I., Yanagihara, C., Ishikawa, N., Komori, T., Matny, O.N., Hubbard, A., Witek, K., Numazawa, H., Green, P., et al. (2023). Rapid Discovery of Functional NLRs Using the Signature of High Expression, High-Throughput Transformation, and Large-Scale Phenotyping. https://doi.org/10.2139/ssrn.4446759.

19. Sutherland, C.A., Stevens, D.M., Seong, K., Wei, W., and Krasileva, K.V. (2025). The resistance awakens: Diversity at the DNA, RNA, and protein levels informs engineering of plant immune receptors from Arabidopsis to crops. Plant Cell *37*, koaf109. https://doi.org/10.1093/plcell/koaf109.

20. Jiao, W.-B., and Schneeberger, K. (2020). Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat. Commun. *11*, 989. https://doi.org/10.1038/s41467-020-14779-y.

21. Shirsekar, G., Devos, J., Latorre, S.M., Blaha, A., Queiroz Dias, M., González Hernando, A., Lundberg, D.S., Burbano, H.A., Fenster, C.B., and Weigel, D. (2021). Multiple Sources of Introduction of North American Arabidopsis thaliana from across Eurasia. Mol. Biol. Evol. *38*, 5328–5344. https://doi.org/10.1093/molbev/msab268.

22. Lian, Q., Huettel, B., Walkemeier, B., Mayjonade, B., Lopez-Roques, C., Gil, L., Roux, F., Schneeberger, K., and Mercier, R. (2024). A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. Nat. Genet. *56*, 982–991. https://doi.org/10.1038/s41588-024-01715-9.

23. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at; 1001 Genomes Consortium (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell *166*, 481–491. https://doi.org/10.1016/j.cell.2016.05.063.

24. Holub, E.B. (2001). The arms race is ancient history in Arabidopsis, the wildflower. Nat. Rev. Genet. *2*, 516–527. https://doi.org/10.1038/35080508.

25. van Wersch, S., and Li, X. (2019). Stronger When Together: Clustering of Plant NLR Disease resistance Genes. Trends Plant Sci. *24*, 688–699. https://doi.org/10.1016/j.tplants.2019.05.005.

26. Saucet, S.B., Ma, Y., Sarris, P.F., Furzer, O.J., Sohn, K.H., and Jones, J. D.G. (2015). Two linked pairs of Arabidopsis TNL resistance genes independently confer recognition of bacterial effector AvrRps4. Nat. Commun. *6*, 6338. https://doi.org/10.1038/ncomms7338.

27. Ameline-Torregrosa, C., Wang, B.-B., O'Bleness, M.S., Deshpande, S., Zhu, H., Roe, B., Young, N.D., and Cannon, S.B. (2008). Identification and Characterization of Nucleotide-Binding Site-Leucine-Rich Repeat Genes in the Model Plant Medicago truncatula. Plant Physiol. *146*, 5–21. https://doi.org/10.1104/pp.107.104588.

28. Borhan, M.H., Gunn, N., Cooper, A., Gulden, S., Tör, M., Rimmer, S.R., and Holub, E.B. (2008). WRR4 encodes a TIR-NB-LRR protein that confers broad-spectrum white rust resistance in Arabidopsis thaliana to four physiological races of Albugo candida. Mol. Plant Microbe Interact. *21*, 757–768. https://doi.org/10.1094/MPMI-21-6-0757.

29. Redkar, A., Cevik, V., Bailey, K., Zhao, H., Kim, D.S., Zou, Z., Furzer, O.J., Fairhead, S., Borhan, M.H., Holub, E.B., et al. (2023). The Arabidopsis WRR4A and WRR4B paralogous NLR proteins both confer recognition of multiple Albugo candida effectors. New Phytol. *237*, 532–547. https://doi.org/10.1111/nph.18378.

30. Liang, W., van Wersch, S., Tong, M., and Li, X. (2019). TIR-NB-LRR immune receptor SOC3 pairs with truncated TIR-NB protein CHS1 or TN2 to monitor the homeostasis of E3 ligase SAUL1. New Phytol. *221*, 2054–2066. https://doi.org/10.1111/nph.15534.

31. Nishimura, M.T., Anderson, R.G., Cherkis, K.A., Law, T.F., Liu, Q.L., Machius, M., Nimchuk, Z.L., Yang, L., Chung, E.-H., El Kasmi, F., et al. (2017). TIR-only protein RBA1 recognizes a pathogen effector to regulate cell death in Arabidopsis. Proc. Natl. Acad. Sci. USA *114*, E2053–E2062. https://doi.org/10.1073/pnas.1620973114.

32. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. Nat. Genet. *20*, 43–45. https://doi.org/10.1038/1695.

33. Boehm, T., and Swann, J.B. (2014). Origin and evolution of adaptive immunity. Annu. Rev. Anim. Biosci. *2*, 259–283. https://doi.org/10.1146/annurev-animal-022513-114201.

34. Giorgetti, O.B., O'Meara, C.P., Schorpp, M., and Boehm, T. (2023). Origin and evolutionary malleability of T cell receptor α diversity. Nature *619*, 193–200. https://doi.org/10.1038/s41586-023-06218-x.

35. Nemri, A., Atwell, S., Tarone, A.M., Huang, Y.S., Zhao, K., Studholme, D. J., Nordborg, M., and Jones, J.D.G. (2010). Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. Proc. Natl. Acad. Sci. USA *107*, 10302–10307. https://doi.org/10.1073/pnas.0913160107.

36. 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. https://doi.org/10.1016/j.cell.2014.10.049.

37. van der Biezen, E.A., Freddie, C.T., Kahn, K., Parker, J.E., and Jones, J. D.G. (2002). Arabidopsis RPP4 is a member of the RPP5 multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signalling components. Plant J. *29*, 439–451. https://doi.org/10.1046/j.0960-7412.2001.01229.x.

38. Yi, H., and Richards, E.J. (2007). A cluster of disease resistance genes in Arabidopsis is coordinately regulated by transcriptional activation and RNA silencing. Plant Cell *19*, 2929–2939. https://doi.org/10.1105/tpc.107.051821.

39. 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. https://doi.org/10.1046/j.1365-313x.2000.00664.x.

40. Reams, A.B., and Roth, J.R. (2015). Mechanisms of gene duplication and amplification. Cold Spring Harb. Perspect. Biol. *7*, a016592. https://doi.org/10.1101/cshperspect.a016592.

41. Stirnweis, D., Milani, S.D., Brunner, S., Herren, G., Buchmann, G., Peditto, D., Jordan, T., and Keller, B. (2014). Suppression among alleles encoding nucleotide-binding-leucine-rich repeat resistance proteins interferes with resistance in F1 hybrid and allele-pyramided wheat plants. Plant J. *79*, 893–903. https://doi.org/10.1111/tpj.12592.

42. Scully, R., Panday, A., Elango, R., and Willis, N.A. (2019). DNA double-strand break repair-pathway choice in somatic mammalian cells. Nat. Rev. Mol. Cell Biol. *20*, 698–714. https://doi.org/10.1038/s41580-019-0152-0.

43. Huefner, N.D., Mizuno, Y., Weil, C.F., Korf, I., and Britt, A.B. (2011). Breadth by depth: expanding our understanding of the repair of transposon-induced DNA double strand breaks via deep-sequencing. DNA Repair *10*, 1023–1033. https://doi.org/10.1016/j.dnarep.2011.07.011.

44. Baumgarten, A., Cannon, S., Spangler, R., and May, G. (2003). Genome-level evolution of resistance genes in Arabidopsis thaliana. Genetics *165*, 309–319. https://doi.org/10.1093/genetics/165.1.309.

45. Monroe, J.G., Powell, T., Price, N., Mullen, J.L., Howard, A., Evans, K., Lovell, J.T., and McKay, J.K. (2018). Drought adaptation in nature by extensive genetic loss-of-function. Preprint at bioRxiv. https://doi.org/10.1101/372854.

46. Monroe, J.G., McKay, J.K., Weigel, D., and Flood, P.J. (2021). The population genomics of adaptive loss of function. Heredity *126*, 383–395. https://doi.org/10.1038/s41437-021-00403-2.

47. Zervudacki, J., Yu, A., Amesefe, D., Wang, J., Drouaud, J., Navarro, L., and Deleris, A. (2018). Transcriptional control and exploitation of an immune-responsive family of plant retrotransposons. EMBO J. *37*, e98482. https://doi.org/10.15252/embj.201798482.

48. Panda, K., and Slotkin, R.K. (2020). Long-Read cDNA Sequencing Enables a "Gene-Like" Transcript Annotation of Transposable Elements. Plant Cell *32*, 2687–2698. https://doi.org/10.1105/tpc.20.00115.

49. Yang, Y., Kim, N.H., Cevik, V., Jacob, P., Wan, L., Furzer, O.J., and Dangl, J.L. (2022). Allelic variation in the Arabidopsis TNL CHS3/CSA1 immune receptor pair reveals two functional cell-death regulatory modes. Cell Host Microbe *30*, 1701–1716.e5. https://doi.org/10.1016/j.chom.2022.09.013.

50. Yang, S., Tang, F., and Zhu, H. (2014). Alternative splicing in plant immunity. Int. J. Mol. Sci. *15*, 10424–10445. https://doi.org/10.3390/ijms150610424.

51. Kufel, J., Diachenko, N., and Golisz, A. (2022). Alternative splicing as a key player in the fine-tuning of the immunity response in Arabidopsis. Mol. Plant Pathol. *23*, 1226–1238. https://doi.org/10.1111/mpp.13228.

52. Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., et al. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature *477*, 419–423. https://doi.org/10.1038/nature10414.

53. MacQueen, A., and Bergelson, J. (2016). Modulation of R-gene expression across environments. J. Exp. Bot. *67*, 2093–2105. https://doi.org/10.1093/jxb/erv530.

54. Choi, K., Reinhard, C., Serra, H., Ziolkowski, P.A., Underwood, C.J., Zhao, X., Hardcastle, T.J., Yelina, N.E., Griffin, C., Jackson, M., et al. (2016). Recombination Rate Heterogeneity within Arabidopsis Disease Resistance Genes. PLOS Genet. *12*, e1006179. https://doi.org/10.1371/journal.pgen.1006179.

55. Möller, M., and Stukenbrock, E.H. (2017). Evolution and genome architecture in fungal plant pathogens. Nat. Rev. Microbiol. *15*, 756–771. https://doi.org/10.1038/nrmicro.2017.76.

56. Bertazzoni, S., Williams, A.H., Jones, D.A., Syme, R.A., Tan, K.-C., and Hane, J.K. (2018). Accessories make the outfit: Accessory chromosomes and other dispensable DNA regions in plant-pathogenic fungi. Mol. Plant Microbe Interact. *31*, 779–788. https://doi.org/10.1094/MPMI-06-17-0135-FI.

57. Bomblies, K., and Weigel, D. (2007). Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. Nat. Rev. Genet. *8*, 382–393. https://doi.org/10.1038/nrg2082.

58. Stahl, E.A., Dwyer, G., Mauricio, R., Kreitman, M., and Bergelson, J. (1999). Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature *400*, 667–671. https://doi.org/10.1038/23260.

59. Tian, D., Araki, H., Stahl, E., Bergelson, J., and Kreitman, M. (2002). Signature of balancing selection in Arabidopsis. Proc. Natl. Acad. Sci. USA *99*, 11525–11530. https://doi.org/10.1073/pnas.172203599.

60. 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. https://doi.org/10.1534/genetics.166.3.1517.

61. Caicedo, A.L., Schaal, B.A., and Kunkel, B.N. (1999). Diversity and molecular evolution of the RPS2 resistance gene in Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA *96*, 302–306. https://doi.org/10.1073/pnas.96.1.302.

62. Johannes, F. (2025). Somatic evolution of stem cell mutations in long-lived plants. Mol. Biol. Evol. msaf165. https://doi.org/10.1093/molbev/msaf165.

63. Johannes, F. (2024). Allometric scaling of somatic mutation and epimutation rates in trees. Evolution *79*, 1–5. https://doi.org/10.1093/evolut/qpae150.

64. Tobias, P.A., and Guest, D.I. (2014). Tree immunity: growing old without antibodies. Trends Plant Sci. *19*, 367–370. https://doi.org/10.1016/j.tplants.2014.01.011.

65. Murray, K. (2025). Teasdale et al 2025 Public Git Annex. Zenodo. https://doi.org/10.5281/ZENODO.15828128.

66. Wlodzimierz, P., Rabanal, F.A., Burns, R., Naish, M., Primetis, E., Scott, A., Mandáková, T., Gorringe, N., Tock, A.J., Holland, D., et al. (2023). Cycles of satellite and transposon evolution in Arabidopsis centromeres. Nature *618*, 557–565. https://doi.org/10.1038/s41586-023-06062-z.

67. Ni, P., Nie, F., Zhong, Z., Xu, J., Huang, N., Zhang, J., Zhao, H., Zou, Y., Huang, Y., Li, J., et al. (2023). DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. Nat. Commun. *14*, 4054. https://doi.org/10.1038/s41467-023-39784-9.

68. Baid, G., Cook, D.E., Shafin, K., Yun, T., Llinares-López, F., Berthet, Q., Belyaeva, A., Töpfer, A., Wenger, A.M., Rowell, W.J., et al. (2023). DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. Nat. Biotechnol. *41*, 232–238. https://doi.org/10.1038/s41587-022-01435-7.

69. Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572. https://doi.org/10.1093/bioinformatics/btr167.

70. Cheetham, S.W., Kindlova, M., and Ewing, A.D. (2022). Methylartist: tools for visualizing modified bases from nanopore sequence data. Bioinformatics *38*, 3109–3112. https://doi.org/10.1093/bioinformatics/btac292.

71. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

72. Li, H. (2021). New strategies to improve minimap2 alignment accuracy. Bioinformatics *37*, 4572–4574. https://doi.org/10.1093/bioinformatics/btab705.

73. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

74. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

75. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

76. Laetsch, D.R., and Blaxter, M.L. (2017). BlobTools: Interrogation of genome assemblies. F1000Res *6*, 1287.

77. Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., and Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. *23*, 258. https://doi.org/10.1186/s13059-022-02823-7.

78. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

79. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

80. Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. *20*, 277. https://doi.org/10.1186/s13059-019-1911-0.

81. Nachtweide, S., and Stanke, M. (2019). Multi-Genome Annotation with AUGUSTUS. Methods Mol. Biol. *1962*, 139–160. https://doi.org/10.1007/978-1-4939-9173-0_8.

82. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. *34*, W435–W439. https://doi.org/10.1093/nar/gkl200.

83. Shumate, A., and Salzberg, S.L. (2021). Liftoff: Accurate Mapping of Gene Annotations. Bioinformatics *37*, 1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016.

84. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

85. Kuo, R.I., Cheng, Y., Zhang, R., Brown, J.W.S., Smith, J., Archibald, A.L., and Burt, D.W. (2020). Illuminating the dark side of the human transcriptome with long read transcript sequencing. BMC Genomics *21*, 751. https://doi.org/10.1186/s12864-020-07123-7.

86. Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. *31*, 5654–5666. https://doi.org/10.1093/nar/gkg770.

87. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

88. Li, H. (2023). Protein-to-genome alignment with miniprot. Bioinformatics *39*, btad014. https://doi.org/10.1093/bioinformatics/btad014.

89. Steuernagel, B., Witek, K., Krattinger, S.G., Ramirez-Gonzalez, R.H., Schoonbeek, H.-J., Yu, G., Baggs, E., Witek, A.I., Yadav, I., Krasileva, K.V., et al. (2020). The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire. Plant Physiol. *183*, 468–482. https://doi.org/10.1104/pp.19.01273.

90. Kourelis, J., Sakai, T., Adachi, H., and Kamoun, S. (2021). RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family. PLOS Biol. *19*, e3001124. https://doi.org/10.1371/journal.pbio.3001124.

91. Lee, E., Helt, G.A., Reese, J.T., Munoz-Torres, M.C., Childers, C.P., Buels, R.M., Stein, L., Holmes, I.H., Elsik, C.G., and Lewis, S.E. (2013). Web Apollo: a web-based genomic annotation editing platform. Genome Biol. *14*, R93. https://doi.org/10.1186/gb-2013-14-8-r93.

92. Robinson, J.T., Thorvaldsdottir, H., Turner, D., and Mesirov, J.P. (2023). igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Bioinformatics *39*, btac830. https://doi.org/10.1093/bioinformatics/btac830.

93. Murray, K.D. (2024). kdm9/blindschleiche, (Version 0.3.1). Zenodo. https://doi.org/10.5281/ZENODO.10049825.

94. Murray, K.D., Borevitz, J.O., Weigel, D., and Warthmann, N. (2024). Acanthophis: a comprehensive plant hologenomics pipeline. J. Open Source Softw. *9*, 6062. https://doi.org/10.21105/joss.06062.

95. Dainat, J., Hereñú, D., Davis, E., Crouch, K., LucileSol, A., Pascal-Git, N., Zollman, Z., and Tayyrov, A. (2023). NBISweden/AGAT: AGAT, (v1.1.0). Zenodo. https://doi.org/10.5281/ZENODO.3552717.

96. Pardo-Palacios, F.J., Arzalluz-Luque, A., Kondratova, L., Salguero, P., Mestre-Tomás, J., Amorín, R., Estevan-Morió, E., Liu, T., Nanni, A., McIntyre, L., et al. (2024). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat. Methods *21*, 793–797. https://doi.org/10.1038/s41592-024-02229-2.

97. Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *10*.

98. Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. *20*, 275. https://doi.org/10.1186/s13059-019-1905-y.

99. Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9*, 18. https://doi.org/10.1186/1471-2105-9-18.

100. Ou, S., and Jiang, N. (2019). LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mobile DNA *10*, 48. https://doi.org/10.1186/s13100-019-0193-0.

101. Xiong, W., He, L., Lai, J., Dooner, H.K., and Du, C. (2014). HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc. Natl. Acad. Sci. USA *111*, 10263–10268. https://doi.org/10.1073/pnas.1410068111.

102. Su, W., Gu, X., and Peterson, T. (2019). TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. Mol. Plant *12*, 447–460. https://doi.org/10.1016/j.molp.2019.02.008.

103. Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. *38*, e199. https://doi.org/10.1093/nar/gkq862.

104. Hu, K., Xu, K., Wen, J., Yi, B., Shen, J., Ma, C., Fu, T., Ouyang, Y., and Tu, J. (2019). Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. BMC Bioinformatics *20*, 354. https://doi.org/10.1186/s12859-019-2945-8.

105. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

106. Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., and Ma, Y. (2022). TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic. Res. *9*, uhac017. https://doi.org/10.1093/hr/uhac017.

107. Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633.

108. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

109. Lovell, J.T., Sreedasyam, A., Schranz, M.E., Wilson, M., Carlson, J.W., Harkess, A., Emms, D., Goodstein, D.M., and Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation

across multiple genomes. eLife *11*, e78526. https://doi.org/10.7554/eLife.78526.

110. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.
111. Katoh, K., Misawa, K., Kuma, K.-I., and Miyata, T. (2002). MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. Nucleic Acids Res. *30*, 3059–3066. https://doi.org/10.1093/nar/gkf436.
112. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.
113. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.
114. Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. Curr. Protoc. Bioinformatics *69*, e96. https://doi.org/10.1002/cpbi.96.
115. Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., et al. (2024). Building pangenome graphs. Preprint at bioRxiv. 2023.04.05.535718. https://doi.org/10.1101/2023.04.05.535718.
116. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., and Garrison, E. (2022). ODGI: understanding pangenome graphs. Bioinformatics *38*, 3319–3326. https://doi.org/10.1093/bioinformatics/btac308.
117. Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88. https://doi.org/10.1186/s13104-016-1900-2.
118. Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. Preprint at arXiv. https://arxiv.org/abs/1303.3997.
119. Danecek, P., and McCarthy, S.A. (2017). BCFtools/csq: haplotype-aware variant consequences. Bioinformatics *33*, 2037–2039. https://doi.org/10.1093/bioinformatics/btx100.
120. Garrison, E., and Marth, G. (2012) Haplotype-Based Variant Detection from Short-Read Sequencing.Preprint at arXiv. https://arxiv.org/abs/1207.3907.
121. McDowell, J.M., Hoff, T., Anderson, R.G., and Deegan, D. (2011). Propagation, storage, and assays with Hyaloperonospora arabidopsidis: A model oomycete pathogen of Arabidopsis. Methods Mol. Biol. *712*, 137–151. https://doi.org/10.1007/978-1-61737-998-7_12.
122. Mencia, R., Arce, A.L., Houriet, C., Xian, W., Contreras, A., Shirsekar, G., Weigel, D., and Manavella, P.A. (2025). Transposon-triggered epigenetic chromatin dynamics modulate EFR-related pathogen response. Nat. Struct. Mol. Biol. *32*, 199–211. https://doi.org/10.1038/s41594-024-01440-1.
123. Yaffe, H., Buxdorf, K., Shapira, I., Ein-Gedi, S., Moyal-Ben Zvi, M., Fridman, E., Moshelion, M., and Levy, M. (2012). LogSpin: a simple, economical and fast method for RNA isolation from infected or healthy plants and other eukaryotic tissues. BMC Res. Notes *5*, 45. https://doi.org/10.1186/1756-0500-5-45.
124. Yuan, W., Beitel, F., Srikant, T., Bezrukov, I., Schäfer, S., Kraft, R., and Weigel, D. (2023). Pervasive under-dominance in gene expression underlying emergent growth trajectories in Arabidopsis thaliana hybrids. Genome Biol. *24*, 200. https://doi.org/10.1186/s13059-023-03043-3.
125. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
126. Li, H. (2012). Seqtk – Toolkit for Processing Sequences in FASTA/Q Formats. GitHub *767*, 69. https://github.com/lh3/seqtk.
127. Zdobnov, E.M., Kuznetsov, D., Tegenfeldt, F., Manni, M., Berkeley, M., and Kriventseva, E.V. (2021). OrthoDB in 2020: evolutionary and functional annotations of orthologs. Nucleic Acids Res. *49*, D389–D393. https://doi.org/10.1093/nar/gkaa1009.
128. Arabidopsis Genome Initiative (2000). Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana. Nature *408*, 796–815. https://doi.org/10.1038/35048692.
129. Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom. Bioinform. *3*, lqaa108. https://doi.org/10.1093/nargab/lqaa108.
130. Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. *89*, 789–804. https://doi.org/10.1111/tpj.13415.
131. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2017). TransDecoder (Broad Institute & CSIRO). https://github.com/TransDecoder/TransDecoder.
132. Rabanal, F.A., Gräff, M., Lanz, C., Fritschi, K., Llaca, V., Lang, M., Carbonell-Bejerano, P., Henderson, I., and Weigel, D. (2022). Pushing the limits of HiFi assemblies reveals centromere diversity between two Arabidopsis thaliana genomes. Nucleic Acids Res. *50*, 12309–12327. https://doi.org/10.1093/nar/gkac1115.
133. Ou, S., and Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. *176*, 1410–1422. https://doi.org/10.1104/pp.17.01310.
134. Neumann, P., Novák, P., Hoštáková, N., and Macas, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA *10*, 1. https://doi.org/10.1186/s13100-018-0144-1.
135. Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science *327*, 92–94. https://doi.org/10.1126/science.1180677.
136. Katoh, K., Kuma, K.-I., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. *33*, 511–518. https://doi.org/10.1093/nar/gki198.
137. Murray, K.D. (2024). kdm9/raugraf, (Version 0.0.5). Zenodo. https://doi.org/10.5281/ZENODO.13144148.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Organisms/strains		
at6923	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76940
at6929	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76532
at7143	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76492
at8285	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76815
at9104	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS77001
at9336	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76715
at9503	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76640
at9578	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS77229
at9744	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76944
at9762	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76487
at9806	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS77223
at9830	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76736
at9847	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76854
at9852	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS76945
at9879	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS77169
at9883	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS77179
at9900	1001 Genomes Consortium	https://abrc.osu.edu/stocks/number/CS77386
Har1032-1	This manuscript	Har1032-1
Har1094-1	This manuscript	Har1094-1
Har1109-1	This manuscript	Har1109-1
Har1123-1	This manuscript	Har1123-1
Har1192-1	This manuscript	Har1192-1
Har1199-1	This manuscript	Har1199-1
Har1206-1	This manuscript	Har1206-1
Har1207-1	This manuscript	Har1207-1
Har1233-1	This manuscript	Har1233-1
Har1239-1	This manuscript	Har1239-1
Har124-1	This manuscript	Har124-1
Har1245-1	This manuscript	Har1245-1
Har1251-1	This manuscript	Har1251-1

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Har1263-1	This manuscript	Har1263-1
Har1272-1	This manuscript	Har1272-1
Har1284-1	This manuscript	Har1284-1
Har129-1	This manuscript	Har129-1
Har1297-1	This manuscript	Har1297-1
Har1312-1	This manuscript	Har1312-1
Har1317-1	This manuscript	Har1317-1
Har1323-1	This manuscript	Har1323-1
Har133-1	This manuscript	Har133-1
Har1350-1	This manuscript	Har1350-1
Har1354-1	This manuscript	Har1354-1
Har136-1	This manuscript	Har136-1
Har136-2	This manuscript	Har136-2
Har1373-1	This manuscript	Har1373-1
Har140-1	This manuscript	Har140-1
Har1407-1	This manuscript	Har1407-1
Har14-1	This manuscript	Har14-1
Har1422-1	This manuscript	Har1422-1
Har1424-1	This manuscript	Har1424-1
Har1440-1	This manuscript	Har1440-1
Har1448-1	This manuscript	Har1448-1
Har1456-1	This manuscript	Har1456-1
Har1457-1	This manuscript	Har1457-1
Har1464-1	This manuscript	Har1464-1
Har1469-1	This manuscript	Har1469-1
Har147-1	This manuscript	Har147-1
Har1471-1	This manuscript	Har1471-1
Har1477-1	This manuscript	Har1477-1
Har14OHML04	This manuscript	Har14OHML04
Har1504-1	This manuscript	Har1504-1
Har1505-1	This manuscript	Har1505-1
Har156-1	This manuscript	Har156-1
Har157-1	This manuscript	Har157-1
Har15INRC55	This manuscript	Har15INRC55
Har162-3	This manuscript	Har162-3
Har1625-1	This manuscript	Har1625-1
Har1629-1	This manuscript	Har1629-1
Har163-1	This manuscript	Har163-1
Har164-1	This manuscript	Har164-1
Har1645-1	This manuscript	Har1645-1
Har166-4	This manuscript	Har166-4
Har168-1	This manuscript	Har168-1
Har16INRC059	This manuscript	Har16INRC059
Har16MIKE048	This manuscript	Har16MIKE048
Har16MIMB003	This manuscript	Har16MIMB003
Har1713-1	This manuscript	Har1713-1
Har1715-1	This manuscript	Har1715-1
Har1719-1	This manuscript	Har1719-1
Har1736a-1	This manuscript	Har1736a-1

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Har1749-1	This manuscript	Har1749-1
Har1751-1	This manuscript	Har1751-1
Har17KABG005	This manuscript	Har17KABG005
Har1837-1	This manuscript	Har1837-1
Har2002-1	This manuscript	Har2002-1
Har211-1	This manuscript	Har211-1
Har257-1	This manuscript	Har257-1
Har320-1	This manuscript	Har320-1
Har325-1	This manuscript	Har325-1
Har458-1	This manuscript	Har458-1
Har459-1	This manuscript	Har459-1
Har466-1	This manuscript	Har466-1
Har471-1	This manuscript	Har471-1
Har477-1	This manuscript	Har477-1
Har495-1	This manuscript	Har495-1
Har511-1	This manuscript	Har511-1
Har527-3	This manuscript	Har527-3
Har538-1	This manuscript	Har538-1
Har55-1	This manuscript	Har55-1
Har564-1	This manuscript	Har564-1
Har566-1	This manuscript	Har566-1
Har660-1	This manuscript	Har660-1
Har661-1	This manuscript	Har661-1
Har682-1	This manuscript	Har682-1
Har695-1	This manuscript	Har695-1
Har708-1	This manuscript	Har708-1
Har711-1	This manuscript	Har711-1
Har724a-1	This manuscript	Har724a-1
Har733-1	This manuscript	Har733-1
Har736-2	This manuscript	Har736-2
Har745-1	This manuscript	Har745-1
Har877-1	This manuscript	Har877-1
Har884-1	This manuscript	Har884-1
Har885-1	This manuscript	Har885-1
Har891-1	This manuscript	Har891-1
Har892-1	This manuscript	Har892-1
Har900-1	This manuscript	Har900-1
Har919-1	This manuscript	Har919-1
Har920-1	This manuscript	Har920-1
Har944-1	This manuscript	Har944-1
HarEmCo5	This manuscript	HarEmCo5
HarHiks1	This manuscript	HarHiks1
Deposited data		
18 HiFi Ath genomes (reads + assemblies)	Wlodzimierz et al.[66]; this manuscript	PRJEB91362
17 HiFi IsoSeq library reads	this manuscript	PRJEB91362
1 Illumina bisulphite library reads	this manuscript	PRJEB91362

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Processed data archive	this manuscript	https://dl20.coevolutionlab.org/datarelease/; https://github.com/coevolutionlab/teasdale-2025-public
Source code for analyses	this manuscript	https://github.com/coevolutionlab/teasdale-2025-public
Software		
PacBio CCS v 6.0.0	Pacific Biosciences of California Inc., Menlo Park, CA, USA	https://github.com/PacificBiosciences/ccs/releases/tag/v6.0.0
PacBio Bam2fastx v1.3.0	Pacific Biosciences of California Inc., Menlo Park, CA, USA	https://github.com/PacificBiosciences/bam2fastx/releases/tag/1.3.0
Ccsmeth 0.3.0	Ni et al.[67]	https://github.com/PengNi/ccsmeth/releases/tag/0.3.0
DeepConsensus	Baid et al.[68]	https://github.com/google/deepconsensus/releases/tag/v1.0.0
Bismark v0.24.0	Krueger et al.[69]	https://github.com/FelixKrueger/Bismark/releases/tag/0.24.0
methylartist	Cheetham et al.[70]	https://github.com/adamewing/methylartist
hifiasm v0.15.4-r343	Cheng et al.[71]	https://github.com/chhylp123/hifiasm/releases/tag/0.15.4
minimap2 (various versions)	Li[72]; Li[73]	https://github.com/lh3/minimap2/releases
samtools (various versions)	Danecek et al.[74]	https://github.com/samtools/samtools/releases
diamond2 (various versions)	Buchfink et al.[75]	https://github.com/bbuchfink/diamond/releases/tag/v2.1.7
blobtools v1.1.1	Laetsch et al.[76]	https://github.com/DRL/blobtools/releases/tag/blobtools_v1.1.1
ragtag v2.1.0	Alonge et al.[77]	https://github.com/malonge/RagTag/releases/tag/v2.1.0
BUSCO v4.0.6	Simão et al.[78]	https://gitlab.com/ezlab/busco/-/releases/4.0.6
QUAST v5	Gurevich et al.[79]	https://quast.sourceforge.net/quast.html
SyRi	Goel et al.[80]	https://github.com/schneebergerlab/syri/releases/tag/v1.3
AUGUSTUS	Stanke et al.[81,82]	https://github.com/Gaius-Augustus/Augustus/releases/tag/v3.4.0
Liftoff	Shumate et al.[83]	https://github.com/agshumate/Liftoff/releases/tag/1.6.0
InterProScan (various versions >= 5.44)	Jones et al.[84]	https://github.com/ebi-pf-team/interproscan/releases
Hydra	This Manuscript	https://github.com/Lnve/HYDRA
lima 2.6.0	Pacific Biosciences of California Inc., Menlo Park, CA, USA	https://github.com/PacificBiosciences/barcoding
Tama	Kuo et al.[85]	https://github.com/GenomeRIK/tama
PASA	Haas et al.[86]; Haas et al.[87]	https://github.com/PASApipeline/PASApipeline/releases/tag/pasa-v2.5.2
TransDecoder (v5.5.0)	(no published manuscript)	https://github.com/TransDecoder/TransDecoder/releases/tag/TransDecoder-v5.5.0
miniprot	Li[88]	https://github.com/lh3/miniprot/releases/tag/v0.5
NLRannotator (v2)	Steuernagel et al.[89]	https://github.com/steuernb/NLR-Annotator

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
NLRtracker (custom version)	Kourelis et al.[90]; this manuscript	https://github.com/kdm9/NLRtracker
WebApollo	Lee et al.[91]	https://github.com/GMOD/Apollo/releases/
igv-js (2.12.6; via Blindschleiche)	Robinson et al.[92]	https://github.com/kdm9/blindschleiche/blob/main/blsl/genigvjs.py
Blindschleiche (various versions)	Murray et al.[93,94]; this manuscript	https://github.com/kdm9/blindschleiche
AGAT	Dainat et al.[95]	https://github.com/NBISweden/AGAT
SQUANTI 3	Pardo-Palacios et al.[96]	https://github.com/ConesaLab/SQANTI3
RepeatMasker (v4)	Chen[97]	https://github.com/Dfam-consortium/RepeatMasker
EDTA v1.9.7	Ou et al.[98]	https://github.com/oushujun/EDTA/releases/tag/v1.9.6
LTRharvest 1.5.10	Ellinghaus et al.[99]	https://www.zbh.uni-hamburg.de/en/forschung/gi/software/ltrharvest.html
LTR_FINDER_parallel 1.0	Ou et al.[100]	https://github.com/oushujun/LTR_FINDER_parallel
HelitronScanner 1.0	Xiong et al.[101]	https://sourceforge.net/projects/helitronscanner/
TIR-Learner 1.23	Su et al.[102]	https://github.com/WeijiaSu/TIR-element-annotation/tree/master/TIR-Learner
MITE-Hunter 1.0	Han et al.[103]	http://target.iplantcollaborative.org/mite_hunter.html
EAhelitron	Hu et al.[104]	https://github.com/dontkme/EAHelitron
CD-hit (various versions)	Li et al.[105]	https://www.bioinformatics.org/cd-hit/
TEsorter	Zhang et al.[106]	https://github.com/zhangrengang/TEsorter
Ape	Paradis et al.[107]	https://cran.r-project.org/web/packages/ape/index.html
Orthofinder 2	Emms et al.[108]	https://github.com/davidemms/OrthoFinder/releases/tag/2.5.4
Genespace	Lovell et al.[109]	https://github.com/jtlovell/GENESPACE
NCBI Blast	Camacho et al.[110]	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
mafft (various versions)	Katoh et al.[111,112]	https://mafft.cbrc.jp/alignment/software/
IQtree (various versions)	Minh et al.[113]	https://github.com/iqtree/iqtree2
ggtree	Yu[114]	https://bioconductor.org/packages/release/bioc/html/ggtree.html
PGGB 0.5.0	Garrison et al.[115]	https://github.com/pangenome/pggb/releases/tag/v0.5.0
Odgi	Guarracino et al.[116]	https://github.com/pangenome/odgi
raugraf 0.1.0	This Manuscript	https://github.com/kdm9/raugraf
Acanthophis v0.2.0	Murray et al.[94]	https://github.com/kdm9/Acanthophis/releases/tag/0.2.0
AdapterRemoval 2.3.1	Schubert et al.[117]	https://github.com/MikkelSchubert/adapterremoval/releases/tag/v2.3.1
BWA MEM	Li[118]	https://github.com/lh3/bwa/releases/tag/v0.7.17
bcftools (various versions)	Danecek et al.[74,119]	https://github.com/samtools/bcftools/
freebayes	Garrison et al.[120]	https://github.com/freebayes/freebayes/releases/tag/v1.3.6
Seqtk	(no published manuscript)	https://github.com/lh3/seqtk

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Accession selection and plant growth

We selected 17 accessions of *Arabidopsis thaliana* (Ath) from the 1001 Genomes Consortium[23] diversity set based on geographic stratification, seed availability, and previously identified haplotype sharing groups.[21] We grew plants in potting mix at 23°C with 16 hours light and 65% humidity until 30 days old. We harvested whole rosettes after 30 hours of dark treatment to reduce starch, and stored them at -80°C.

Phenotyping of Ath-Har interactions

We inoculated a Europe-wide collection of 104 *Hyaloperonospora arabidopsidis* (Har) isolates individually on five to ten 8-day-old seedlings of each of the 17 accessions following a standard protocol.[121] Briefly, we propagated isolates on *eds-1* Ath by infecting 8-day old eds-1 seedlings with spores extracted from sporulating tissue (initially from field samples, thereafter the previous generation of *eds-1* infected with the respective isolate) every approximately 10 days. We grew infected plants at 16 C and approximately 100 μE light intensity in self-contained growth cabinets. We performed infection assays under similar conditions, using 8-day old seedlings of each respective Ath isolate. We replicated inoculations 3-5 times per accession and scored the immune response of the accessions after 11 days of the inoculation as resistant (no Har sporulation to mild Har sporulation) or susceptible (moderate to profuse Har sporulation).

METHOD DETAILS

HMW DNA Extraction

To extract high molecular weight (HMW) nuclear DNA with minimal organellar contamination, we first isolated nuclei. Approximately 40 grams of fresh plant tissue was ground to an ultra-fine powder with a mortar and pestle cooled with liquid nitrogen. Working at 4°C, we resuspended about 20 grams of ground tissue in 200mL of nuclei isolation buffer (NIB) and gently stirred for 15 minutes. The solution was filtered through Miracloth and incubated with 10 mL NIB-Triton20 for 15 minutes. We centrifuged the solution to collect the nuclei and washed the pellet twice with NIB-Triton20 and finally centrifuged again to collect the final clean nuclei pellet.

We extracted HMW DNA by lysing nuclei in G2 lysis buffer at 37°C, treating with RNase for 30 minutes at 37°C and then proteinase K at 50°C for 3 hours. We centrifuged this lysate at 4°C to remove debris. We eluted and precipitated the HMW DNA in the genomic tip device, then physically separated clumped DNA with a sterile glass hook and eluted in Qiagen EB buffer. We left the eluted HMW DNA at room temperature until it fully dissolved and stored it at 4°C. We measured HMW DNA fragment sizes with the Femto® Pulse system (Agilent), and quantitated DNA with the dsDNA-HS Qubit Fluorometer kit (Life Technologies). Typical yields were 30-80 μg of HMW DNA, with median fragment lengths of over 60-80 kb.

PacBio HiFi Library Preparation

We prepared the extracted genomic DNA for PacBio Circular Consensus Sequencing (CCS) with the "Procedure & Checklist - Preparing HiFi SMRTbell Libraries using SMRTbell Express Template Prep Kit 2.0 (PN 101-853-100 Version 01 (September 2019))" protocol, with the following modifications. We sheared the genomic DNA with the Megaruptor 2 device (Diagenode) to a target size of 20 kb to 25 kb. To reduce clogging of the Megaruptor hydropore, we sheared 15 μg HMW genomic DNA in a volume of 700 μl to an average size of between 15-20 kb. We purified sheared DNA with AMPure PB Beads (Pacific Biosciences) and checked the concentration and size with the Qubit Fluorometer and the Femto Pulse System. We used the Express Template Prep Kit 2.0 for library construction, except that we doubled the suggested amount of input DNA (10 μg of sheared DNA) and accordingly increased all reagent volumes two-fold. The library was size selected to >15 kb using a BluePippin instrument. Immediately before sequencing on a Pacific Biosciences Sequel II instrument, we performed a final bead cleanup, annealed the templates with Sequencing Primer v2 and bound it to the sequencing polymerase.

HiFi Reads and DNA Methylation

We generated HiFi consensus reads from raw read files using the PacBio Circular Consensus Sequencing tool (v6.0.0) and bam2fasta, keeping reads $\geq$ 10 kb and with average quality $\geq$Q20 for genome assembly. We re-generated HiFi consensus reads with kinetics information with ccsmeth call_hifi.[67] We generated an Illumina bisulfite sequencing (BS-seq) library as described[122] for a single accession (at9852) using a portion of input DNA from which PacBio CCS reads had been generated. We used Bismark[69] to generate estimates of cytosine methylation from the BS-seq reads. We trained a ccsmeth model (v0.3.0; [67]) to call methylation based on HiFi read kinetics, using the Bismarck methylation estimates as ground truth. Briefly, genomic positions with 100% methylation and at least 6x BS-seq read depth were considered methylated and those with 0% methylation were considered non-methylated. During cross-validation, our fine-tuned ccsmeth CG model had 98% accuracy, but CHG and CHH sites could not be accurately predicted. We then used ccsmeth to estimate methylation status for each HiFi reads for all accessions. We used methylartist[70] to aggregate read-level methylation data for individual cytosines in each genome.

Long-read transcript evidence

To produce transcript evidence for gene annotation, we infected 10-day old seedlings of each of the 17 accessions with three isolates of the specialist oomycete pathogen *Hyaloperonospora arabidopsidis* (166-4, 495-1, 527-3). We harvested two replicates each at 2 hours, 24 hours, and 4 days post infection, and extracted RNA with a silica column-based protocol.[123,124] We quantified the RNA using a Nanodrop instrument and pooled all samples for each accession in an approximately equimolar fashion. We prepared Iso-Seq libraries using the "Procedure & Checklist - Preparing HiFi SMRTbell ® Libraries using SMRTbell ® Express Template Prep Kit 3.0 (PN102-141-700 Rev: 06 (2022))" protocol. Libraries were individually indexed using SMRTbell barcoded adaptors during the adapter ligation step of the library prep protocol. The libraries were multiplexed and sequenced across 3 SMRTcells (8-10 libraries per SMRTcell) using the Sequel II binding kit 3.1 recommended for libraries with fragments shorter than 3 kb. The ten libraries with the lowest yields were re-pooled and re-sequenced on a fourth SMRTcell.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genome assembly and quality assessment

We assembled reads into primary contigs with hifiasm (v0.15.4-r343).[71] We mapped HiFi reads back to primary contigs with minimap2[72,73] and SAMtools,[74,125] and searched each contig against the NCBI's non-redundant protein database with diamond blastx.[75] We used blobtools[76] to combine these data and generate summary plots that we used to identify contaminant sequences, which we removed with seqtk.[126] To increase the accuracy of scaffolding, we used an optical map (produced by BioNano Genomics) of accession at9852. We scaffolded contaminant-filtered primary contigs of at least 100 kb by aligning them to the BioNano map of at9852 using ragtag scaffold (v2.0.1;[77]). We assessed the completeness and correctness of each de novo assembly with BUSCO (v4.0.6)[78] using the odb10_embryophyta database.[127] We calculated continuity, GC content, overall assembly length and reference coverage of assembled contigs with Quast (v5.0.2).[79] We detected structural variation between each scaffolded genome and the *A. thaliana* reference genome TAIR10[128] with SyRi (v1.3).[80]

Ab initio and homology-guided gene annotation

We created initial gene annotations using both homology evidence and statistical prediction. We created *ab initio* gene predictions with Augustus,[81] using a BUSCO-refined *A. thaliana* model similar to the BREAKER workflow.[129] We used liftoff[83] to transfer the Araport11[130] annotation of the Col-0 reference accession to homologous regions of each assembly. We also produced a protein domain annotation, for all putative gene models using InterProScan (interproscan-5.51-85.0[84]). We then translated these matches into GFF files using pfam2gff.py (https://github.com/wrf/genomeGTFtools/blob/master/pfam2gff.py).

Gene annotation with transcript evidence

To predict isoforms from the Iso-Seq reads we used a custom pipeline that can handle multiple reference genomes (https://github.com/Lnve/HYDRA). Briefly, Pacbio CCS reads were demultiplexed and barcodes removed using Lima (2.6.0) with the Iso-Seq mode. We clipped poly-A tails using refine (3.8.1) and removed apparent concatemers. We merged samples that were sequenced twice with samtools (1.16.1). We mapped reads to the respective reference genome using minimap2 (2.17-r941 -ax splice:hq;[72,73]) and collapsed the reads into predicted isoforms using TAMA (tc_version_date_2020_12_14;[85]). We set the 5' threshold (-a 10) to 10 bp, the 3' threshold to 5 bp (-z 5) and did not allow for any differences at the splice junctions (-m 0), resulting in variation at the ends, but capturing all possible splice junctions of each gene. Only sequences with 99% identity were collapsed (-i 99). We then updated the initial gene predictions from AUGUSTUS with PASA (2.5.2;[86]) using default settings. We predicted open reading frames for all isoforms using transdecoder (5.7.0;[131]) with the complete ORF setting.

Evidence-weighted gene prediction

To integrate *ab initio*, liftoff and RNA supported gene predictions, we combined all predictions into non-overlapping loci likely representing the same underlying gene(s). We used NLRannotator[89] and NLRtracker[90] to find NLR remnants or NLR genes missing from our annotated genes. We devised a decision tree (see Figure S11) to categorize genes according to the amount and source of consistent evidence. Where annotators agreed on the coding sequence and exon structure of a gene, we selected the most evidence-rich annotation (PASA > liftoff > Augustus). Where RNA-based annotation disagreed with *ab initio* and/or liftoff prediction, or where RNA evidence was missing, we manually curated NLR annotations.

The manual curation process involved examining the locus in both WebApollo[91] and igv-js,[92] and assessing which evidence track contained the valid gene model with the most canonical NLR domain structure supported by both RNA and homology evidence. Manual annotation was mostly needed for a small set of predictions where the Iso-seq evidence overlapped with multiple *ab initio* and liftoff annotations, or where there were multiple liftoff annotations for a single Augustus annotation and vice versa. Where we encountered valid open reading frames encoding NLRs with non-canonical domain structure, we used Iso-Seq evidence as the putative gene model. In the absence of Iso-Seq evidence, if putative annotations supported differing numbers of genes, we examined the domain structure of these ORFs to determine whether the "split" or "conjoined" putative annotation was likely more accurate, informed by the domain structure of homologous sequences from other accessions. For example, if liftoff of Araport11 predicted two gene models but the IsoSeq and the domain structure predicted a single NLR with a TE insertion, we selected the gene model predicted using the IsoSeq. We used a custom script to parse the manual annotation decisions and output a composite GFF. GFF

entries were adjusted for some truncated pseudogenes, where the extent of the mRNA boundaries was lengthened to encompass the full 3' UTR of the truncated gene. Finally, we combined the outputs for each accession, which we then sanitized for a variety of common minor problems using both AGAT[95] and custom tools.[93] Where multiple transcripts were predicted for a gene model, the transcript with the 5'-most start codon and then longest CDS was chosen as representative. Genes were labelled as pseudogenes if there was a liftoff annotation that did not have a valid ORF. Proto-pseudogenes/truncated genes were labelled as part of the manual NLR annotation process, usually where only one mutation interrupted the NLR ORF, often with Iso-seq evidence for a transcript with a long 3' UTR downstream of the mutation truncating the ORF.

Isoform diversity

We summarized isoform variation based on the ORF structure of each isoform (i.e. collapsing isoforms with the same open reading frame) using SQUANTI3.[96] We calculated the isoform diversity for each gene as Simpson's index of diversity (1-Simpson's D; 1=maximal diversity). As observed isoform diversity is limited by the number of reads sequenced, we also recalculated Simpson's index of diversity only for genes that had at least 10 Iso-seq reads. As we did not explicitly enrich for intact 5' caps, we also re-calculated isoform diversity disregarding isoform variation at the 5' end to account for potential transcript degradation.

Repeat annotation

We annotated satellite repeats such as telomeres, centromeres and rDNA clusters by homology to a library of consensus sequences from the reference accession Col-0[132] using RepeatMasker[97] (v 4.0.5) with the following parameters: -e ncbi -s -a -inv -xsmall -div 40 -no_is -nolow -cutoff 200 -norna. To annotate transposable elements (TEs) in each genome, we first used EDTA_raw.pl (v.1.9.7)[98] to annotate LTRs with LTRharvest (v1.5.10),[99] LTR_FINDER_parallel (v1.0)[100] and LTR_retriever,[133] Helitrons with HelitronScanner (v1.0),[101] and TIR elements with TIR-Learner (v1.23)[102] and MITE-Hunter (v1.0).[103] We then merged all 90 chromosomes from the 18 accessions, combined the raw output and proceeded with the rest of the EDTA pipeline, adding the current TE library from TAIR10. Originally 18 genomes were included in this dataset, but we could not generate IsoSeq expression evidence for at6137, so it was not included in any analysis of NLR diversity and evolution. The result was a combined TE library and GFF annotation of TEs common for all the accessions.

To refine these automated annotations and to detect previously unknown repeat families, we used additional steps: To mitigate rampant mis-annotation of tandem repeats as TEs, we intersected EDTA annotation with the independent satellite annotation of centromeres, telomeres and rDNA and removed EDTA-annotated TEs overlapping >20% with satellite repeats. We removed repeats not assigned to a known repeat family, as they were predominantly either artefacts of the joint analysis of all 18 genomes, or unidentified satellite repeats.

To increase the confidence of the *de novo* Helitron annotations, we considered whether a TE family had at least one intact member, and whether there was a Rep/Hel protein domain in at least one member (using RexDB Viridiplantae v3.0.[134] Finally, we ran EAhelitron[104] to reannotate Helitrons and report whether a given TE copy intersected with both EDTA and EAhelitron.

The EDTA pipeline did not assign a TE family to several TE instances because they were single copy, likely because we had run the raw EDTA module independently for each genome. To correct this, we associated elements with known TE families via BLAST matches following the 80/80/80 rule. We clustered the remaining TE instances that did not correspond to a known TE family following the 80/80/80 rule using CD-hit.[105] Clusters with at least two copies were assigned new TE family names and their corresponding TE models were incorporated into the TE library of the pangenome. Finally, we used TEsorter[106] to assign all LTR families in our TE library to known clades using RexDB Viridiplantae v3.0.[134] Finally, we estimated the age of intact LTRs by aligning their two LTR ends and calculating the pairwise distance between the two ends using the dist.dna function of the R package "ape"[107] with the model "K80" and translating it to millions of years using the mutation rate calculated in Ossowski et al.[135]

TE gene annotation

To distinguish *bona fide* TE genes from non-TE genes, we devised the following decision tree (Figure S13). We combined all TE and non-TE genes produced by the annotation (together referred to as putative genes from now on), and used Diamond v2[75] to compare their protein products against a curated repeat database of TE-related protein domains (RexDB Viridiplantae 3.0[134]). We overlapped all putative genes with a collapsed version of the TE annotation without Helitrons, merging all annotated TE copies, minus Helitrons, that overlap or are within 100 bp of each other. We classified putative genes as TE genes if they overlapped at least 20% with merged TEs and had a RexDB hit. In addition, we classified putative genes as TE genes if they did not have a hit against RexDB but overlapped at least 90% with merged TEs. We compared all putative genes with a RexDB hit with a merged Helitron annotation and classified those with more than 20% overlap as TE genes. We classified as "TE-like protein genes" all other putative genes with a RexDB hit but without overlap with the TE or Helitron annotations. All remaining genes without a RexDB hit and without overlap with merged TEs or Helitrons as protein coding genes.

Orthogrouping and phylogenetic inference

To identify NLR sequences that represent the same gene, we first extracted the protein sequences of all genes with a valid ORF (excluding pseudogenes but including truncated genes) using AGAT.[95] We also extracted the nucleotide sequences for all genes (including pseudogenes and pseudogenic regions). We clustered all protein sequences across the 17 accessions using

OrthoFinder2[108] as an initial first pass with default settings. This resulted in 35,697 orthogroups (OGs), of which 249 contained NLR proteins. We confirmed the global synteny of the 17 genomes using Genespace[109] (Figure S1c).

Many diversity metrics are heavily affected by how orthogrouping is performed, which can be quite arbitrary, particularly for NLRs with both recent and not so recent copy number variation. We attempted to account for the variation in NLRs by defining broad orthogroups (OGs) along with finer orthogroups (OG70s, Table S4). For OGs, we first conducted an all-by-all protein blast (BLASTP[110]) to ensure that sequences had not been mis-assigned by OrthoFinder and to determine whether any OGs needed to be merged. If only a few sequences in an OG had close hits ($\geq$ 85% amino acid sequence identity) to another OG, we moved these sequences to the other OG. If most sequences of an OG had a close hit with sequences from another OG, we merged the OGs. This procedure reduced the number of OGs from 249 to 204 and led to OG reassignment of 24 sequences. To produce final OGs representing sequences of nominally the same NLR gene, we clustered sequences within OGs using CD-hit[105] with a 70% sequence similarity threshold to produce OG70s. By definition, all members of an OG70 belong to the same OG. We aligned sequences in each OG70s using mafft –auto[136] calculated average pairwise distance using a custom script - p-distance_script_v3.py and Shannon entropy using Blindschleiche (blsl entropy[93]).

We assigned pseudogenes to OGs using tblastn (NCBI BLAST+[110]), aligning all NLR proteins to the nucleotide sequences of the pseudogene regions. A pseudogene was assigned to the orthogroup of the protein that best aligned to each pseudogenic region. We did not search against the entire protein complement as the dataset contains TE proteins, many of which exist inside the pseudogenes.

We constructed gene trees for orthogroups (OGs and OG70s) of interest (e.g. the orthogroup containing ADR2, Figure 5B). We aligned nucleotide sequences for the entire genes using MAFFT[136] and reconstructed a Maximum likelihood phylogeny using IQtree[113] and visualised the tree using ggtree.[114]

Pangenome graphs

We induced a graph of each of the five chromosomes from all assemblies with PGGB[115] using the parameters recommended for *A. thaliana* (-s 5000, -p 95, -n 18, -k 47). We visualised graphs and subgraphs with odgi viz.[116] We calculated local graph complexity (''node radius'') with a graph-walking metric: for each node in the graph counted the total number of unique nodes that could be visited with a specific number of steps without back-tracking and ignoring self-loops. We walked each assembly's path through the graph, transferring node-level statistics to assembly-specific genome coordinates. The functionalities are implemented in the novel tool 'raugraf'.[137]

Definition of NLR-dense neighborhoods

We defined NLR-dense gene neighborhoods (''NLR neighborhoods'') with a semi-automated pangenome graph traversal algorithm. For each chromosome, we find the position of an NLR in the pangenome graph. We then traverse the sequence graph outward in both directions until we find at least five graph nodes that together represent at least 100 bp of syntenic, single-copy sequence shared across all accessions (Figure 1B). These syntenic anchors from the left and right borders of a region with at least one NLR in at least one accession. We then define a NLR neighborhood in each accession by finding the corresponding genome coordinates of these syntenic anchor nodes with odgi. We repeat this for all NLRs in all accessions, skipping those already covered by a neighborhood, until all annotated NLRs are contained in neighborhoods. We removed two neighborhoods that contained only a single incomplete NLR fragment across the pangenome, in both cases a partial TIR sequence appears to be inserted into pericentromeric TE repeat clusters.

Genotyping with data from the 1001 Genomes project

To determine the species-wide frequency of mutations that disrupt the coding sequence of NLR genes, we applied the Acanthophis variant calling pipeline[94] to short reads from the 1001 Genomes project.[23] We removed low quality and adapter sequences from reads with AdapterRemoval[117] and mapped reads to the at9852 assembly with BWA MEM,[118] calling variants with freebayes.[120] We predicted variant consequences with bcftools csq,[74,119] using our final gene annotation of at9852. We calculated the total frequency of severe mutations (frameshifts, starts/stops gained or lost) as the sum of the allele frequency of all variants with a minor allele frequency above 1%, a variant quality above 1,000, and a total coverage within 5,000-50,000 (across 1,135 samples, equivalent to an average depth of approximately 5-50x).

ADDITIONAL RESOURCES

In addition to sequence deposition in the European Nucleotide Archive (see key resources table), we have created an online portal to collate additional data releases and browsers, available at https://dl20.coevolutionlab.org/datarelease/ and mirrored at https://github.com/coevolutionlab/teasdale-2025-public.