High intraspecies allelic diversity in *Arabidopsis* NLR immune receptors is associated with distinct genomic and epigenomic features

Authors

Chandler A. Sutherland[1], Daniil M. Prigozhin[2], J. Grey Monroe[3], and Ksenia V. Krasileva[1]

[1] Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA 94720
[2] Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA 94720
[3] Department of Plant Sciences, University of California Davis, Davis, CA, USA 95616

C.A.S. ORCID: 0000-0001-5840-7661
D.M.P. ORCID: 0000-0003-2075-0231
J.G.M. ORCID: 0000-0002-4025-5572
K.V.K. ORCID: 0000-0002-1679-0700

Abstract

Plants rely on Nucleotide-binding, Leucine-rich repeat Receptors (NLRs) for pathogen recognition. Highly variable NLRs (hvNLRs) show remarkable intraspecies diversity, while their low variability paralogs (non-hvNLRs) are conserved between ecotypes. At a population level, hvNLRs provide new pathogen recognition specificities, but the association between allelic diversity and genomic and epigenomic features has not been established. Our investigation of NLRs in *Arabidopsis* Col-0 has revealed that hvNLRs show higher expression, less gene body cytosine methylation, and closer proximity to transposable elements than non-hvNLRs. hvNLRs show elevated synonymous and nonsynonymous nucleotide diversity and are in chromatin states associated with an increased probability of mutation. Diversifying selection maintains variability at a subset of codons of hvNLRs, while purifying selection maintains conservation at non-hvNLRs. How these features are established and maintained, and whether they contribute to the observed diversity of hvNLRs is key to understanding the evolution of plant innate immune receptors.

Introduction

Plants, lacking the adaptive immune systems of vertebrates, use germline-encoded innate immune receptors to defend against rapidly evolving pathogens. Despite their inability to create antibodies through somatic hypermutation and recombination, plants are protected against pathogens due to population-level receptor diversity. Nucleotide-binding, Leucine-rich repeat Receptors (NLRs) are the intracellular sensors of the plant immune system, detecting pathogen-secreted, disease-promoting effector proteins. After binding of a pathogen target to the LRR domain, NLRs initiate defense responses through oligomerization of the central

nucleotide-binding domain, leading to transcriptional reprogramming, hormone induction, and hypersensitive cell death response (Ngou, Ding and Jones, 2022). Plant NLRs are differentiated into three anciently diverged classes based on their N-terminal domains: Resistance To Powdery Mildew 8-NLR (RNL), Coiled-Coil-NLR (CNL), or Toll/Interleukin-1 Receptor-NLR (TNL) that are responsible for the downstream signaling.

NLRs exhibit remarkable levels of intraspecies allelic diversity (Van de Weyer *et al.*, 2019), due to both the genomic processes that generate variation and selection that promotes its maintenance (Karasov, Horton and Bergelson, 2014; Barragan and Weigel, 2021; Märkle, Saur and Stam, 2022). NLRs are organized into clusters more often than other genes, which can asymmetrically drive NLR expansion and diversification through unequal crossing over and gene conversion (Michelmore and Meyers, 1998; Lee and Chae, 2020) as well as accumulation of point mutations (Kuang *et al.*, 2004). Point mutations are a major source of within-species NLR diversity, but have been difficult to fully resolve through short-read sequencing approaches. The NLR gene family includes the most polymorphic loci and contains the highest frequency of major effect mutations in the *Arabidopsis* genome (Gan *et al.*, 2011). There is evidence for balancing selection maintaining polymorphisms and presence-absence variation at several NLR loci through frequency-dependent selection, spatial and temporal fluctuations in pathogen pressure, and heterozygote advantage (Thrall *et al.*, 2012; Karasov *et al.*, 2014; MacQueen *et al.*, 2019). Diversifying selection has also been observed at NLR loci as an excess of nonsynonymous to synonymous substitutions (Bakker *et al.*, 2006). The NLR gene and protein sequences within a species represent a snapshot of the ongoing interplay between mutation and selection, but disentangling their relative contributions remains challenging.

Mutation rates are unlikely to evolve on a gene by gene basis in response to selection given the barrier imposed by genetic drift (Lynch, 2010). However, selection on genic mutation rates is sufficiently strong when acting on mechanisms that couple mutation rate to expression states and epigenomic features, affecting the mutation rates of many genes simultaneously (Martincorena and Luscombe, 2013). The mutation rate of *Arabidopsis* is heterogeneous across the genome, consistent with expected effects of selection on mechanisms linking mutation rates to epigenomic features (Monroe *et al.*, 2022; Staunton, Peters and Seoighe, 2023). Several mechanisms have been described, including cytosine methylation which is positively correlated with mutation probability and known to increase the likelihood of spontaneous deamination (Cao *et al.*, 2011; Weng *et al.*, 2019) while H3K4me1, which is negatively correlated with mutation probability and a target of several DNA repair proteins (Quiroz *et al.*, 2022). Description of genomic features associated with diversity in NLRs will help to understand the role of mutation bias in NLR evolution.

Recent advances in enrichment-based long-read sequencing of NLRs (Jupe *et al.*, 2013) as well as long-read pan-genomes (Jiao and Schneeberger, 2020) allowed for re-examination of NLR variation within species (Barragan and Weigel, 2021). In *Arabidopsis* datasets, it has been

shown that NLRs are enriched in regions of synteny diversity and that NLR repertoires across species could not be easily anchored to a reference genome (Van de Weyer *et al.*, 2019). Phylogenetic analysis independent of reference-based assignment of pan-NLRomes from 62 *Arabidopsis thaliana* accessions (Van de Weyer *et al.*, 2019) and 54 *Brachypodium distachyon* (Gordon *et al.*, 2017) lines allowed for amino acid diversity quantification and delineation of highly variable NLRs (hvNLRs) from their low-variability paralogs (non-hvNLRs) (Prigozhin and Krasileva, 2021). At the population level, hvNLRs show rapid rates of diversification and are hypothesized to act as reservoirs of diversity for recognition of pathogen effectors. Comparison of hv and non-hvNLR gene sets allows for investigation of epigenomic, sequence, and regulatory features (hereafter genomic features) and signatures of selection associated with NLR diversification.

In this paper, we report that hvNLRs show a higher transcription level, less gene body CG methylation, and closer proximity to transposable elements (TEs) than non-hvNLRs. Elevated gene-wide nucleotide diversity, a higher likelihood of mutation, and diversifying selection at a subset of sites promote the high amino acid diversity of hvNLRs, while non-hvNLRs are subject to purifying selection. These findings will serve as a starting point for investigation of the mechanisms that promote and maintain diversity generation in a subset of plant immune receptors.

Results

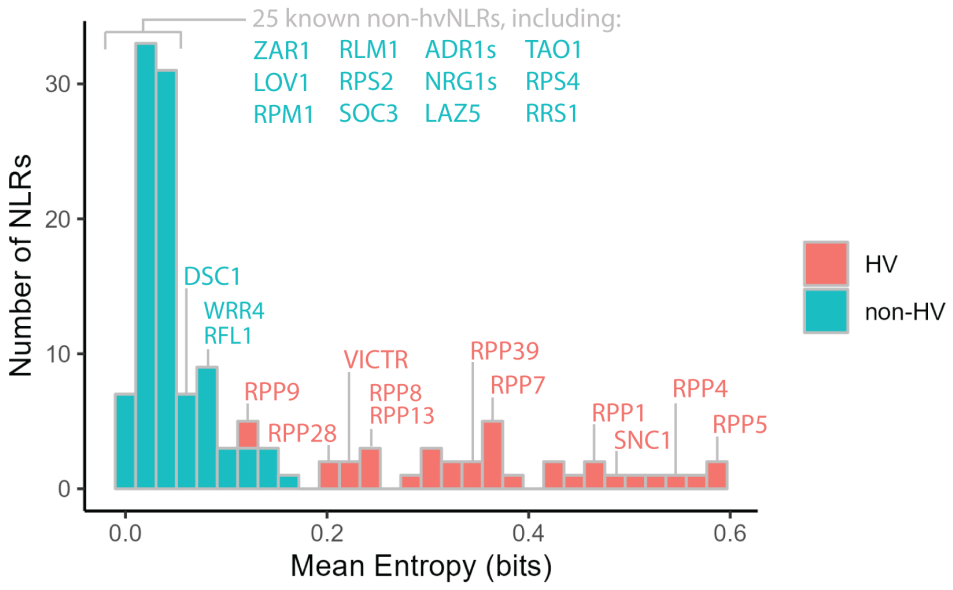


Figure 1: hvNLRs are defined by high amino acid diversity. Distribution of mean per gene Shannon entropy across the *Arabidopsis* NLRome in bits. Described NLRs are annotated.

Shannon entropy, a measure of variability derived from information theory, provides an unbiased metric of amino acid diversity of a protein within a population (Asti *et al.*, 2016; Wang

et al., 2017). Here, the Shannon entropy is the sum of the frequency of each amino acid times the logarithm of that frequency at each position in a protein sequence alignment, so sites with low variability have low entropy and highly diverse sites have high entropy. When applied to NLRs, this measure is predictive of highly variable effector binding sites (Prigozhin and Krasileva, 2021). Based on the bimodal distribution of Shannon entropy in NLRome, we defined hvNLRs as proteins with 10 or more amino acid positions with Shannon entropy greater than 1.5 bits (**Supplemental Fig. 1**) (Prigozhin and Krasileva, 2021). To examine the relationships between population level diversity and genomic features of a single accession, we plotted Shannon entropy by sequence in Col-0 (**Fig. 1**). As expected, there are functional hvNLRs and non-hvNLRs, with known direct recognition of effectors corresponding to hvNLRs and known indirect recognition to non-hvNLRs. hvNLRs also overlap with dangerous mix genes. Categorizing NLRs into low and high entropy groups allows for binary comparison of features and gene set enrichment analysis to compare NLRs to the rest of the genome.

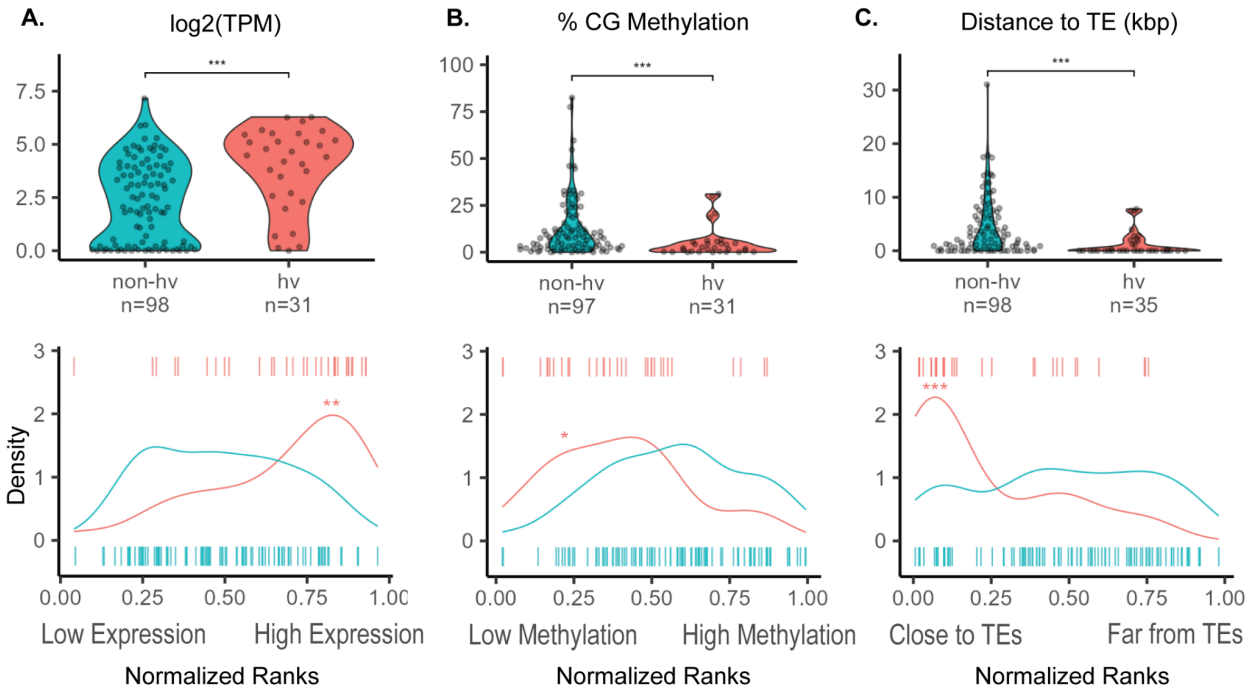


Figure 2: Expression, methylation, and proximity to transposable elements (TEs) distinguish hv and non-hv NLRs. **A:** average gene expression $\log_2$(Transcripts per Million (TPM)), **B:** average % CG methylation per gene, and **C:** distance to the nearest TE (kbp) with normalized mean percentile rank density plots of hv and non-hvNLRs.* indicates a p-value <0.05 and ≥ 0.01; ** indicates a p-value <.01 and ≥0.001; *** indicates a p-value <0.001.

To compare the expression and methylation status of hv and non-hvNLRs within an individual plant, we examined available paired whole genome bisulfite and RNA sequencing generated from the same rosette leaf (Williams *et al.*, 2022). We found that hvNLRs are expressed significantly higher than non-hvNLRs (**Fig. 2A**, unpaired Wilcoxon rank-sum test, p=7.9e-05). When we ranked all protein coding *Arabidopsis* genes based on their expression

level, we observed that hvNLRs are enriched in the most expressed genes in each leaf sample (singscore rank-based sample scoring, $p < 0.005$ for hvNLRs in each biological replicate) (Foroutan *et al.*, 2018).

In addition, hvNLRs have significantly lower gene body CG methylation than non-hvNLRs (**Fig. 2B**, unpaired Wilcoxon rank-sum test, p=4.3e-04), and hvNLRs are enriched in the CG hypomethylated genes across the genome (**Fig. 2B**, permutation test for difference in means, p= 0.003, n=10,000 replicates). Gene set analysis of methylation can be biased due to the uneven distribution of CG sites within each gene (Geeleher *et al.*, 2013). We repeated our permutation test to compare hvNLRs to a set of non-NLR genes with similar measured CG sites per gene to correct for this bias. Still, hvNLRs were significantly more hypomethylated than the rest of the genome ($p < 0.05$ each biological replicate, n=10,000). We noticed two hvNLRs, *RPP4* and *RPP7*, with higher CG methylation than the average for hvNLRs (**Fig. 2B**). Upon further inspection, we also found CHH and CHG context methylation within the gene bodies of *RPP4* and *RPP7* (**Supplemental Fig. 2**), which we rarely observed in other NLRs. Multi-context gene body methylation (CG, CHH, and CHG) is typically used to silence nearby or overlapping transposable elements (Quadrana *et al.*, 2016). This indicates that their elevated CG methylation is likely due to multi-context silencing related to a recent TE insertion.

We also found that hvNLRs are much more likely to be near TEs (Fig. 2C, unpaired Wilcoxon rank-sum test, p = 1.7e-06), and hvNLRs are enriched in the genes closest to TEs (permutation test for difference in medians, p=0, n=10,000 replicates). In Col-0, hvNLRs have a median TE distance of 0 kbp, meaning the TEs are within the UTR or intronic sequences, while non-hvNLRs have a median TE distance of 2.07 kbp. Highly variable status of NLRs is predictive of TEs within the genic sequence (Fisher's exact test, p=3.6e-05). It has been previously observed that TEs are associated with plant immune genes (Kawakatsu *et al.*, 2016), but this analysis suggests that the signal is driven by hvNLRs.

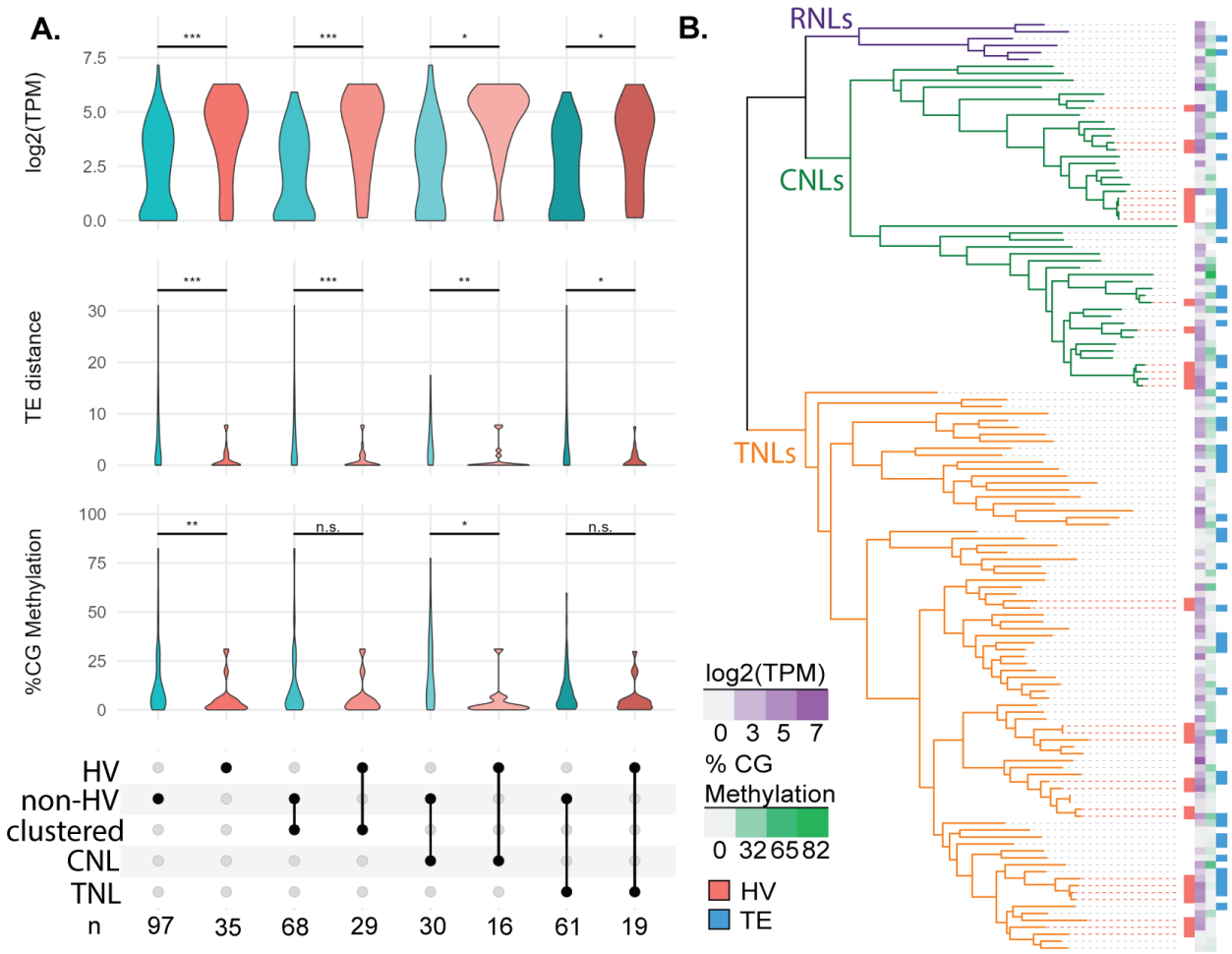


Figure 3: Cluster membership, NLR type, and phylogenetic distance do not account for genomic differences between hv and non-hvNLRs. **A:** Comparison of expression, distance to nearest TEs, and CG gene body methylation of hv and non-hvNLRs by cluster membership and N-term domain type. **B:** Features mapped onto a phylogeny of NLRs in *A. thaliana* Col-0. NLRs without $\log_2$(TPM) or %CG methylation data were determined to be unmappable (see methods).

NLRs are found in clusters more frequently than other genes (Lee and Chae, 2020). However, highly variable status of NLRs is not dependent on cluster membership (Fisher's exact test, p=0.18) and hv and non-hvNLRs maintain their distinct expression and TE-association patterns when comparing exclusively clustered hv and non-hvNLRs (**Fig. 3A**, unpaired Wilcoxon rank-sum tests, corrected for multiple hypothesis testing). Expression and TE distance patterns are also independent of the CNL and TNL N-terminal domain clades (**Fig. 3A**). CG methylation, however, is not significantly different between clustered hv and non-hvNLRs and between TNLs (**Fig. 3A**). CG methylation is the weakest association with hv status of the three examined features (**Fig 2B**), and further analysis with more accessions will reveal if cluster or hv status is more predictive of CG methylation. hvNLRs are distributed over the phylogeny of

NLRs, and maintain distinct genomic features despite close phylogenetic relationships with non-hvNLRs (**Fig. 3B**). Overall, we conclude expression and TE distance cannot be explained by cluster status, phylogenetic proximity, or NLR class.

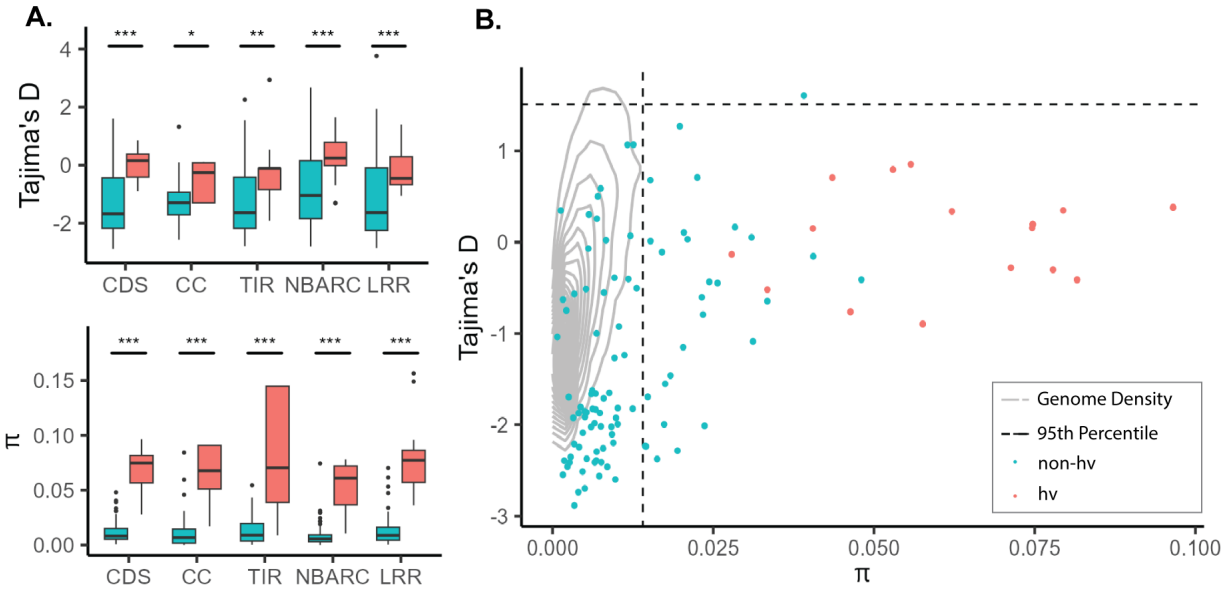


Figure 4: hvNLRs show higher Tajima's D and nucleotide diversity than non-hvNLRs. **A.** D and π calculated across the coding sequence (CDS), coiled-coil (CC), Toll/Interleukin-1 (TIR), nucleotide-binding (NBARC) and leucine rich repeat (LRR) domains. Within each box, horizontal black lines denote median values; boxes range from the 25th to 75th percentile of each group's distribution of values; whiskers extend no further than 1.5x the interquantile range of the hinge. Data beyond the end of the whiskers are outlying points and are plotted individually. **B.** CDS π vs. D. Gray lines represent the kernel density estimation of statistics computed on all coding sequences of *Arabidopsis*. Dashed lines represent the 95th percentile of the empirical distribution.

The high level of amino acid diversity in hvNLRs and associated difference in genomic features might be due to differences in mutational processes and/or selection. In order to investigate the contribution of balancing selection to the observed amino acid diversity at hvNLRs, we calculated Tajima's D (D) and nucleotide diversity per site (π) in each domain and across the gene body of hv and non-hvNLRs. hvNLRs have higher D than non-hvNLRs across the coding sequence and all individual domains (**Fig. 4A**; unpaired Wilcoxon rank-sum test, corrected for multiple comparisons). Reflecting their differences in amino acid diversity, hvNLRs have higher π than non-hvNLRs across all domains and the coding sequence (**Fig. 4A**; unpaired Wilcoxon rank-sum test, corrected for multiple comparisons). The difference in π and D between the two groups is not driven exclusively by variation in the LRR region, with the highest values reported for the hvNLR NBARC domains.

Due to the demographic history of *Arabidopsis*, the empirical distribution of summary statistics departs from the neutral model (Nordborg et al., 2005; Alonso-Blanco et al., 2016). We calculated the genome-wide values of D and π to test for selection, using whole genome SNP information from the accessions used to create the pan-NLRome (Alonso-Blanco *et al.*, 2016) (**Supplemental Fig. 3**). Both hv and non-hvNLRs have higher average π than the empirical distribution (**Fig. 4B;** permutation test for difference in means, p = 0; p=0, n=10,000 replicates), and there are significantly more NLRs in the top 5% of the empirical distribution than expected by chance (permutation test for number in the top 5%, p=0, n=10,000 replicates). This corroborates previously reported significantly high levels of nucleotide diversity of NLRs. (Bakker *et al.*, 2006; Van de Weyer *et al.*, 2019)

hvNLRs have a higher D and non-hvNLRs have a lower D than the genome average (**Fig. 4B**; permutation test for difference in means, p = 0.0009; p=0, n=10,000 replicates). There are no hvNLRs in either tail of the empirical distribution of D, which is not significantly different from the 0.43 expected by chance. There are, however, an excess of non-hvNLRs in the bottom 5% of the distribution of D (permutation test for number in the bottom 5%, p=0, n=10,000 replicates), indicating that purifying selection may be reducing diversity at non-hvNLRs. Defining individual genes under balancing selection to be the top 5% of the empirical distribution of π and D values (Bakker *et al.*, 2006; Gladieux *et al.*, 2022), we identified one non-hvNLR under balancing selection, *AT5G47260* **(Fig. 4B)**. However, one gene is not significantly different from the number of NLRs expected to be in the top 5% of both distributions by random chance.

To further investigate the nature of the high nucleotide diversity of NLRs, we compared nucleotide diversity at synonymous and non-synonymous sites (π_S;π_N). hvNLRs have greater π_S and π_N than non-hvNLRs (**Fig 5A**; unpaired wilcoxon rank sum test, p=5.6e-13, p=1.2e-15). However, the ratio of non-synonymous to synonymous nucleotide diversity (π_N/π_S), an intraspecies measurement of selection, is not significantly different between the two groups, indicating possible role of different mutational processes (**Fig 5B**; unpaired wilcoxon rank sum test, p=0.24). Average π_N/π_S is < 1 for both groups across the gene and in the LRR region, indicating purifying selection as an excess of synonymous polymorphisms relative to non-synonymous polymorphisms (**Fig. 5B**; **Supplementary Fig. 4**).

Since elevated π_N and π_S with no difference in π_N/π_S could be caused by an increase in mutation rate of hvNLRs, we compared the predicted SNVs and indels per base pair based on epigenomic states (mutation probability score) (Monroe *et al.*, 2022). The mutation probability is 35% higher for hvNLRs (**Fig 5C**; unpaired wilcoxon rank sum test, p=3.0e-05).

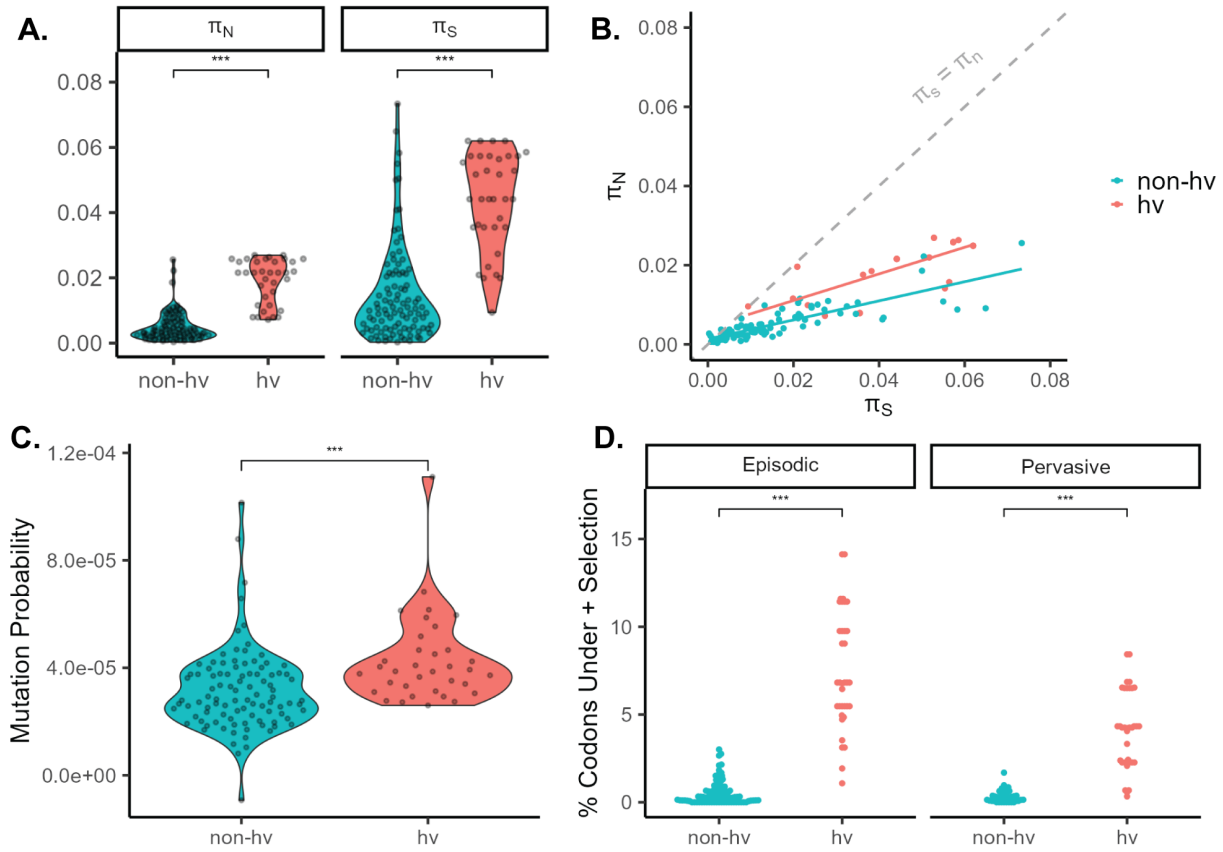


Figure 5: hvNLR nucleotide diversity is associated with a high likelihood of mutation and codons under diversifying selection. **A.** Average nonsynonymous pairwise nucleotide diversity per site (π_N) and average synonymous pairwise nucleotide diversity per site (π_S). **B.** π_S vs π_N of the coding sequence of NLRs with per group linear regressions. **C.** Mutation probability score of hv and non-hvNLRs. **D.** Percentage of codons under positive selection determined by MEME (episodic), and FEL (pervasive).

Gene-wide π_N/π_S is a conservative metric for testing positive selection because positive selection may only be acting at a few codon sites (Kosakovsky Pond and Frost, 2005). Therefore, we used maximum-likelihood based site models to test for positive, diversifying selection. Use of these dN/dS-based models on intraspecies data is problematic because the nucleotide differences do not represent substitutions fixed by selection, but rather polymorphisms segregating within a population (Kryazhimskiy and Plotkin, 2008). We mitigated this effect by restricting our analysis to internal branches of the protein phylogeny, which encompass at least one ancestral sequence that is visible to selection (Pond *et al.*, 2006; Avanzato *et al.*, 2019). hvNLRs have a higher proportion of codons under pervasive and episodic diversifying selection than non-hvNLRs, indicating that diversifying selection at a subset of sites is maintaining diversity at hvNLRs (**Fig. 5D,** unpaired wilcoxon rank sum test). Given the polymorphism data,

summary statistics, and mutational likelihood, hvNLR amino acid diversity appears to be driven by both a higher likelihood of mutation and positive, diversifying selection, while non-hvNLR conservation is maintained by purifying selection.

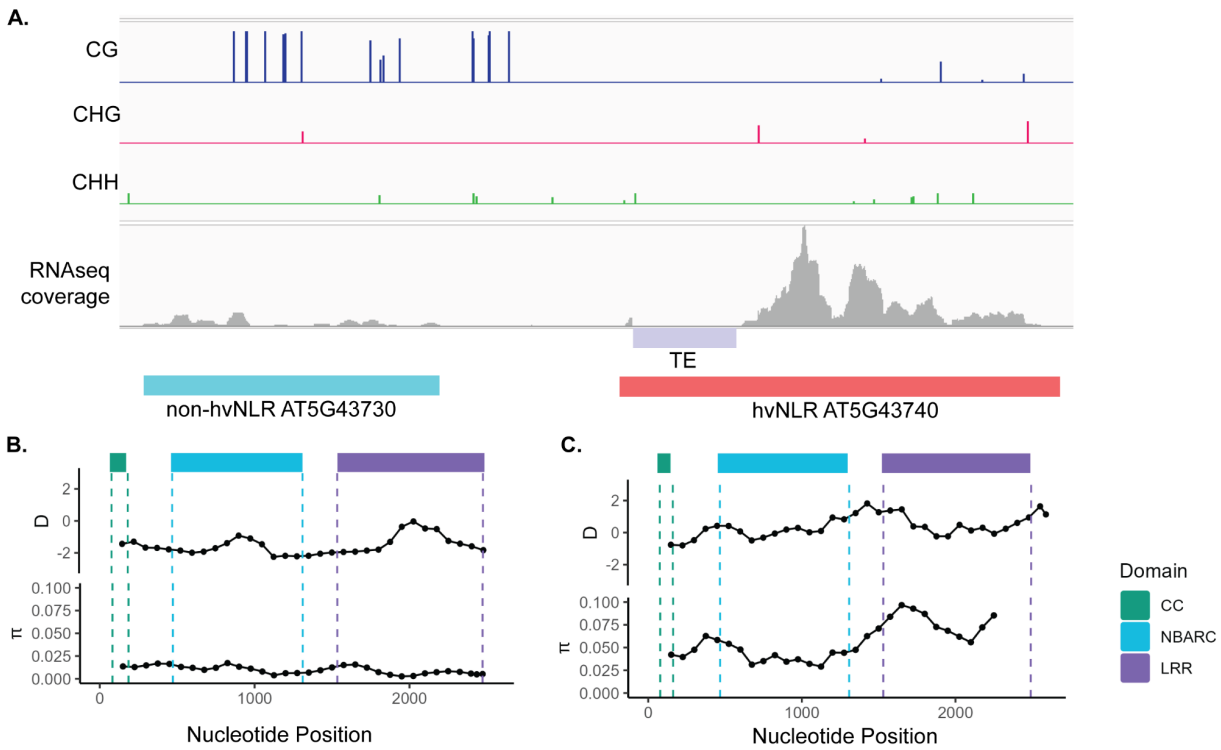


Figure 6: Neighboring NLRs retain distinct genomic and epigenomic features. **A.** Methylation, RNAseq coverage, and TE proximity of neighboring non-hvNLR *AT5G43730* and hvNLR *AT5G43740*. **B.** and **C.** Tajima's D and nucleotide diversity across the coding sequence of *AT5G43730* and *AT5G43740*. Statistics were calculated on 300bp windows with a step size of 75bp, and plotted at the nucleotide midpoint.

As described previously, hv and non-hvNLRs can co-exist as neighboring genes. We chose non-hvNLR *AT5G43730* and hvNLR *AT5G43740*, two CNLs of similar length 1.8kb apart, to examine the genomic features and signatures of selection of neighboring NLRs (**Fig. 6**). The hvNLR is highly expressed, hypomethylated, and has a TE within its 5' UTR sequence (**Fig. 6A**). The non-hvNLR shows signatures of purifying selection with a gene-wide Tajima's D value of -1.9, while the hvNLR has a gene-wide Tajima's D of -0.24 (**Fig. 6B, 6C**). The hvNLR has higher π, π_N, and π_S, but the two genes have similar π_N/π_S values (0.48 and 0.41) (**Fig. 6B, 6C**). Despite neighboring genomic positions, *AT5G43730* and *AT5G43740* show distinct genomic features and signatures of selection reflective of their species-level amino acid diversity (**Fig. 6B, 6C**). Therefore, we conclude that genomic features that distinguish hvNLR and non-hvNLRs are not driven by broader genome states, but may instead be related to function and evolutionary speed.

Discussion

The high allelic diversity of NLRs has long been appreciated, though the mechanisms that generate and maintain this diversity have remained difficult to disentangle. Taking advantage of Shannon entropy and available long read sequencing datasets, we can delineate rapidly and slowly diversifying NLRs and begin to investigate these mechanisms through gene set comparison. Our results show that rapidly evolving NLRs have distinct genomic features from their conserved paralogs and the rest of the genome. Specifically, we found that hvNLRs are more expressed, less methylated, and closer to TEs than non-hvNLRs. Interestingly, hvNLRs are enriched across the genome in highly expressed genes, hypomethylated genes, and genes closest to TEs, while non-hvNLRs are uniformly dispersed among other genes.

Since we observed distinct genomic features between hv and non-hvNLRs, we investigated the possibility of increased mutation rate in hvNLRs through examination of nucleotide diversity and mutation probability. Synonymous substitutions are under reduced selection compared to nonsynonymous substitutions because they do not alter the amino acid sequence, but are not invisible to selection due to codon bias, GC biased gene conversion, and RNA folding stability (Martincorena, Seshasayee and Luscombe, 2012; James, Castellano and Eyre-Walker, 2017; Wei, 2020). π_S is therefore an imperfect predictor of mutation rate, but an elevated mutation rate of hvNLRs could result in increased π_S and π_N relative to non-hvNLRs, but not influence the π_N/π_S ratio, as we report here (Bromham, Cowman and Lanfear, 2013). We also find that hvNLRs are maintained in chromatin states associated with a higher mutation probability per base pair relative to non-hvNLRs, leading to the hypothesis that locally high mutation rate at hvNLRs contributes to the observed amino acid diversity. However, high depth quantification of *de novo* mutations at NLRs before selection is required to evaluate this hypothesis.

The distinct genomic features between the two NLR groups may point to mechanisms of increased mutation rate. Transcription is a source of genomic instability through the exposure of vulnerable single-stranded DNA, which is countered by targeting DNA repair machinery to actively transcribed genes through the stalling RNA polymerase or histone marks associated with actively transcribed genes (Oztas *et al.*, 2018; Quiroz *et al.*, 2022). If the high transcription of hvNLRs is not accompanied by targeted DNA repair, this would result in an increased probability of mutation (Staunton, Peters and Seoighe, 2023). Methylated cytosines increase the likelihood of mutation by increasing the frequency of spontaneous deamination of cytosines (Xia, Han and Zhao, 2012; Weng *et al.*, 2019; Monroe *et al.*, 2022). However, in *Arabidopsis*, gene body CG methylation is found preferentially in the exons of conserved, constitutively transcribed housekeeping genes, and gene body CG methylation is associated with lower polymorphism than unmethylated genes across accessions (Gaut *et al.*, 2011; He *et al.*, 2022; Kenchanmane Raju *et al.*, 2023). The CG gene body methylation of non-hvNLRs may therefore be related to their low diversity through some unknown mechanism. TEs generate large effect

mutations (Quadrana *et al.*, 2019) and alter the methylation and expression landscape of surrounding genes. hvNLRs are closer to TEs and more likely to have them within their genic sequence than non-hvNLRs, and this likely contributes to hvNLR diversification.

Once generated, nucleotide diversity can be actively maintained by diversifying or balancing selection, or passively accumulate in the absence of selection. We do not observe any difference in diversifying selection between hv and non-hvNLRs using the π_N/π_S metric, but hvNLRs have a significantly higher proportion of codons under pervasive and episodic diversifying selection. While hvNLRs have higher Tajima's D values than the genome average and non-hvNLRs, they are not present in the tails of the genome-wide distribution. The 5th and 95th percentiles of the empirical distribution is a conservative cutoff, and it is possible for a locus to be under selection but not in a tail of the distribution if selection is weak. Therefore, balancing selection may play a role in promoting hvNLR diversity, but cannot be distinguished from evolution under relaxed selection using this criteria. non-hvNLRs, however, have a strong signature of purifying selection, which helps to explain their low amino acid diversity relative to hvNLRs.

Given the heterogeneous mutation rate across the *Arabidopsis* genome, it is tempting to speculate that the distinctive genomic features we observed in hvNLRs may be related to their allelic diversity. Alternatively, there might be a selection of specific features on non-hvNLRs to enhance DNA repair and inhibit other diversity-generation activities facilitating their maintenance. Our findings serve as a starting point for the investigation of the mechanisms that promote diversity generation in a subset of the plant immune receptors.

Materials and Methods

To examine the methylation and expression of NLRs, we used available matched bisulfite and RNA sequencing from split Col-0 leaves (Williams *et al.*, 2022). Reads were trimmed using Trim Galore! v0.6.6 with a Phred score cutoff of 20 and Illumina adapter sequences, with a maximum trimming error rate 0.1 (Babraham Bioinformatics). Using Bismark v0.23.0, reads were mapped to the Araport11 genome, PCR duplicates were removed, and percent methylation at each cytosine was determined using the methylation extraction function (Krueger and Andrews, 2011). Cytosines with at least 5 reads were used for analysis, and the symmetrical cytosines within CG base pairs were averaged (Williams *et al.*, 2022). The percent methylation of each CG site was averaged across each NLR gene, and across four biological replicates. Five hvNLR genes did not have sufficient coverage at any cytosines and were excluded from analysis *(AT1G58807*, *AT1G58848*, *AT1G59124*, *AT1G59218*, and *AT4G26090*).

RNA-seq reads from four matched leaf samples (explained above) were mapped to the Araport11 genome using STAR v2.7.10a and were counted using htseq-count v2.0.2 (Dobin *et al.*, 2013). Counts were converted to transcripts per million and averaged across four biological replicates, then log2 transformed for visualization. NLRs are repetitive and often similar, making them difficult to sequence with short reads. To determine if any NLRs were unmappable, RNAseq reads were simulated using Polyester v1.2.0 (Frazee *et al.*, 2015). Four NLRs were determined to be unmappable due to zero assigned

read counts and were excluded from expression analysis (*AT1G58807*, *AT1G58848*, *AT1G59124*, and *AT1G59218*). Single sample gene set enrichment of hvNLRs and non-hvNLRs was performed on each replicate using singscore (Foroutan *et al.*, 2018).

We determined distance to transposable elements based on the TE annotation file TAIR10_Transposable_Elements.txt and gene annotation file TAIR10_GFF3_genes.gff available from arabidopsis.org. The phylogenetic tree of all NLRs in Col-0 was generated as described previously (Prigozhin and Krasileva, 2021) with feature annotations using iTOL. The UpSet plot was generated using the R package ComplexUpset v1.3.3.

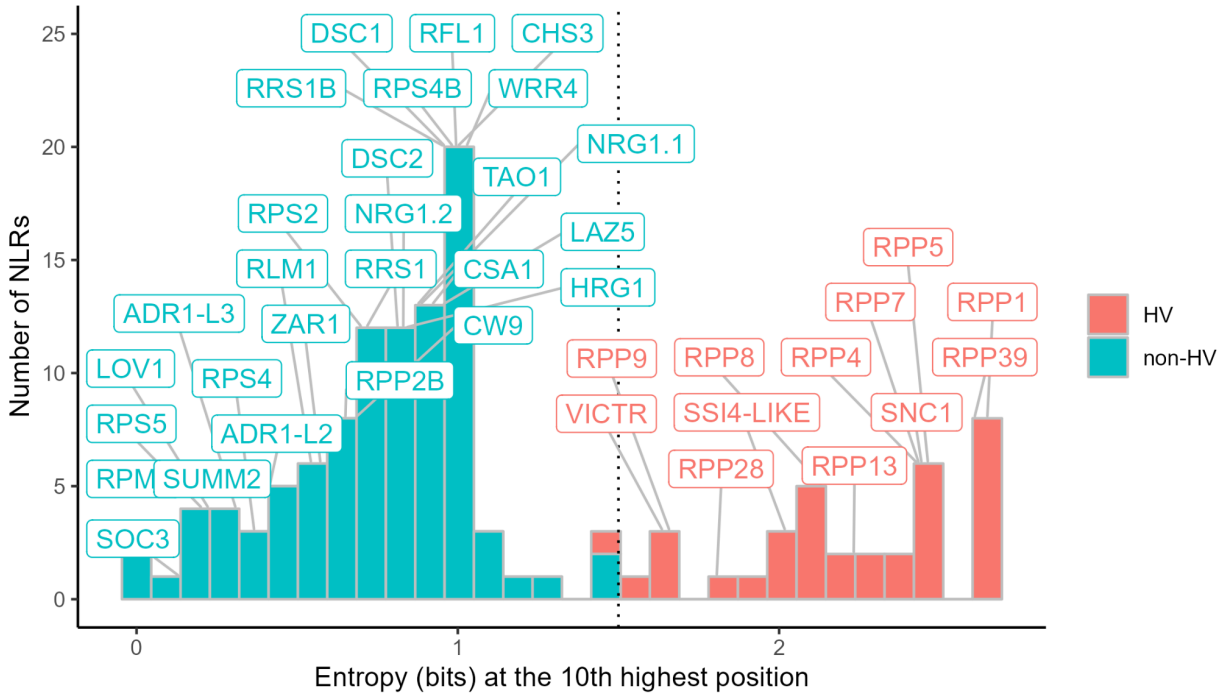
Protein alignments for each NLR were generated as described previously (Prigozhin and Krasileva, 2021) and converted to codon alignments using PAL2NAL v14 (Suyama, Torrents and Bork, 2006). The population genetics statistics of NLRs were calculated using EggLib v3.1.0 (Siol *et al.*, 2022). Domain specific statistics were calculated on subsets of codon alignments using majority vote across annotations. NB-ARC, TIR, and CC annotations were collected from previous work (Van de Weyer *et al.*, 2019), and LRR annotations were determined using LRRpredictor (Martin *et al.*, 2020). Sliding window analysis was performed with 300 base pair windows with a 75 base pair step. Sites under pervasive diversifying selection were identified using FEL (Kosakovsky Pond and Frost, 2005) and sites under episodic diversifying selection were identified using MEME (Murrell *et al.*, 2012) using the internal branches of the phylogeny. Empirical distributions of population genetics statistics of coding sequences were calculated from the all sites 1001 Genomes VCF subset to the accessions used to generate the NLRome long read dataset using vcftools v0.1.17 (Danecek *et al.*, 2011; Alonso-Blanco *et al.*, 2016; Van de Weyer *et al.*, 2019).

All the data generated in this study is hosted on the Zenodo Public Repository at 10.5281/zenodo.7527904. The processing pipelines and figure generation code are available on Github (https://github.com/chandlersutherland/nlr_features).

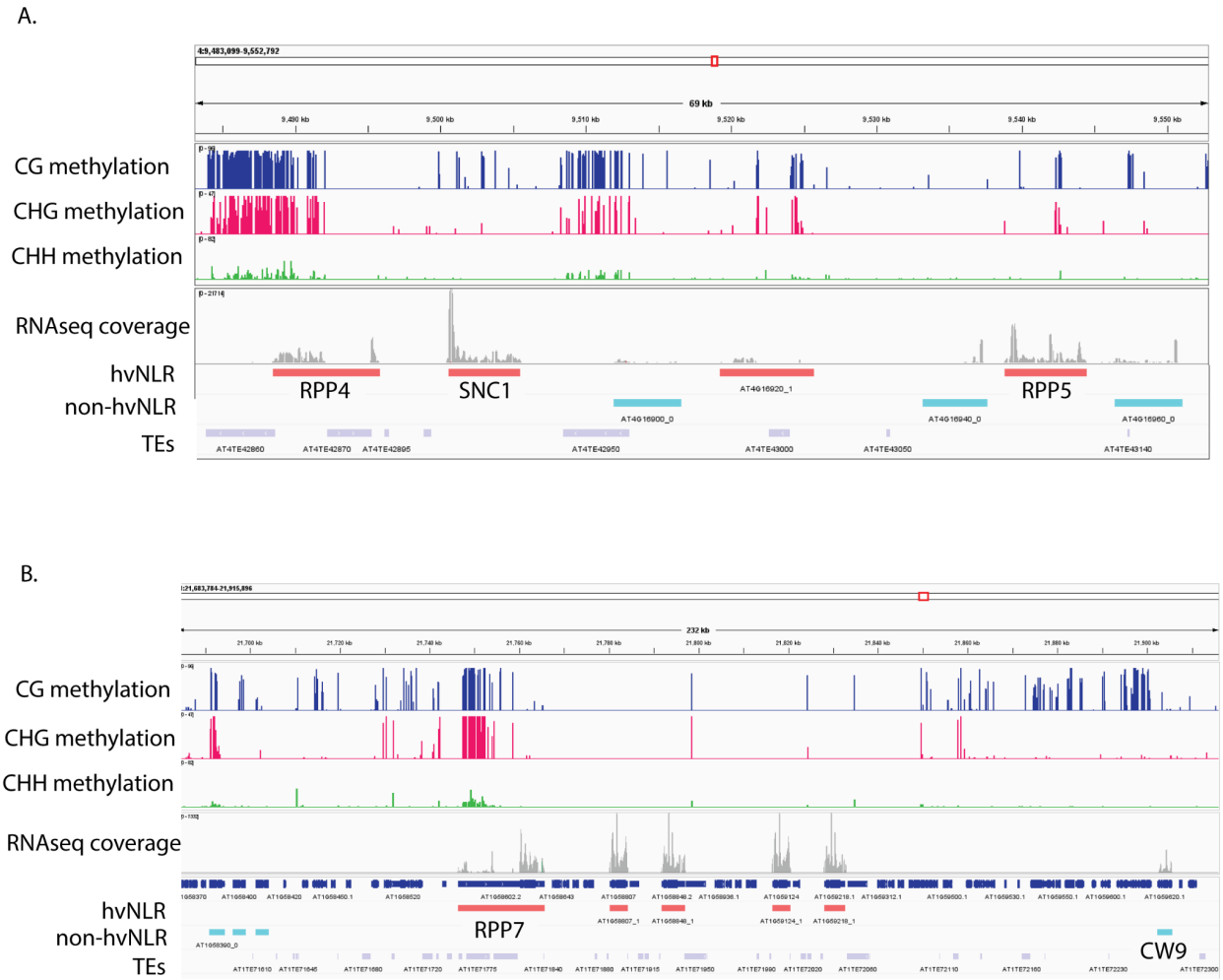
Acknowledgements

We are grateful to the Krasileva Lab for the critical reading of the manuscript. This research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer). Chandler A. Sutherland has been supported by the Grace Kase-Tsujimoto Graduate Fellowship. Ksenia V Krasileva is funded by NIH Director's Award (1DP2AT011967-01), Gordon and Betty Moore Inventor Fellowship (grant number: 8802) and the Innovative Genomics Institute.

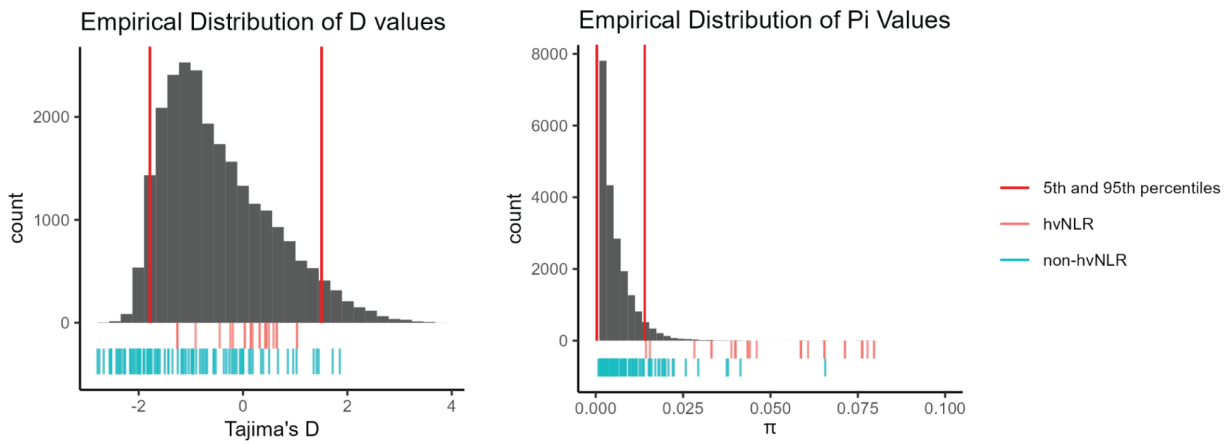
Supplemental Figures



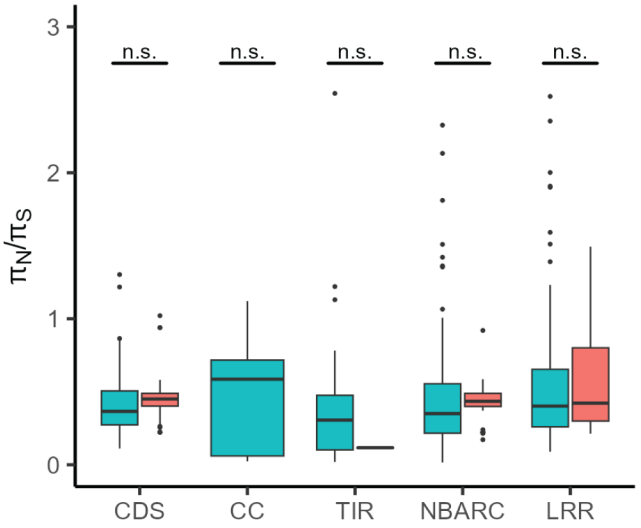
Supplemental Figure 1: Distribution of NLR Shannon entropy at the tenth highest amino acid position as shown as a histogram with 30 bins. Described NLRs are annotated. The designation of hvNLR is entropy of >1.5 bits at the tenth highest position, as shown by the dashed line.



Supplemental Figure 2: Integrative Genomics Viewer screenshot of Methylation, RNAseq coverage, and TE proximity of the **A.** *RPP4* and **B.** *RPP7* clusters.



Supplemental Figure 3: Empirical distribution of Tajima's D and π calculated on coding sequences of *Arabidopsis*. Position of hv and non-hvNLRs shown via rug plot, as well as the 5th and 95th percentiles of the distribution.



Supplemental Figure 4: π_N/π_S calculated per domain.

References

Alonso-Blanco, C. *et al.* (2016) '1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana', *Cell*, 166(2), pp. 481–491. Available at: https://doi.org/10.1016/j.cell.2016.05.063.

Asti, L. *et al.* (2016) 'Maximum-Entropy Models of Sequenced Immune Repertoires Predict Antigen-Antibody Affinity', *PLOS Computational Biology*, 12(4), p. e1004870. Available at: https://doi.org/10.1371/journal.pcbi.1004870.

Avanzato, V.A. *et al.* (2019) 'A structural basis for antibody-mediated neutralization of Nipah virus reveals a site of vulnerability at the fusion glycoprotein apex', *Proceedings of the National Academy of Sciences*, 116(50), pp. 25057–25067. Available at: https://doi.org/10.1073/pnas.1912503116.

Bakker, E.G. *et al.* (2006) 'A Genome-Wide Survey of *R* Gene Polymorphisms in *Arabidopsis*', *The Plant Cell*, 18(8), pp. 1803–1818. Available at: https://doi.org/10.1105/tpc.106.042614.

Barragan, A.C. and Weigel, D. (2021) 'Plant NLR diversity: the known unknowns of pan-NLRomes', *The Plant Cell*, 33(4), pp. 814–831. Available at: https://doi.org/10.1093/plcell/koaa002.

Bromham, L., Cowman, P.F. and Lanfear, R. (2013) 'Parasitic plants have increased rates of molecular evolution across all three genomes', *BMC Evolutionary Biology*, 13(1), p. 126. Available at: https://doi.org/10.1186/1471-2148-13-126.

Cao, J. *et al.* (2011) 'Whole-genome sequencing of multiple Arabidopsis thaliana populations', *Nature Genetics*, 43(10), pp. 956–963. Available at: https://doi.org/10.1038/ng.911.

Danecek, P. *et al.* (2011) 'The variant call format and VCFtools', *Bioinformatics*, 27(15), pp. 2156–2158. Available at: https://doi.org/10.1093/bioinformatics/btr330.

Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics (Oxford, England)*, 29(1), pp. 15–21. Available at: https://doi.org/10.1093/bioinformatics/bts635.

Foroutan, M. *et al.* (2018) 'Single sample scoring of molecular phenotypes', *BMC Bioinformatics*, 19(1), p. 404. Available at: https://doi.org/10.1186/s12859-018-2435-4.

Frazee, A.C. *et al.* (2015) 'Polyester: simulating RNA-seq datasets with differential transcript expression', *Bioinformatics*, 31(17), pp. 2778–2784. Available at: https://doi.org/10.1093/bioinformatics/btv272.

Gan, X. *et al.* (2011) 'Multiple reference genomes and transcriptomes for Arabidopsis thaliana', *Nature*, 477(7365), pp. 419–423. Available at: https://doi.org/10.1038/nature10414.

Gaut, B. *et al.* (2011) 'The Patterns and Causes of Variation in Plant Nucleotide Substitution Rates', *Annual Review of Ecology, Evolution, and Systematics*, 42(1), pp. 245–266. Available at: https://doi.org/10.1146/annurev-ecolsys-102710-145119.

Geeleher, P. *et al.* (2013) 'Gene-set analysis is severely biased when applied to genome-wide methylation data', *Bioinformatics*, 29(15), pp. 1851–1857. Available at: https://doi.org/10.1093/bioinformatics/btt311.

Gladieux, P. *et al.* (2022) 'Extensive immune receptor repertoire diversity in disease-resistant rice landraces'. bioRxiv, p. 2022.12.05.519081. Available at: https://doi.org/10.1101/2022.12.05.519081.

Gordon, S.P. *et al.* (2017) 'Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure', *Nature Communications*, 8(1), p. 2184. Available at: https://doi.org/10.1038/s41467-017-02292-8.

He, L. *et al.* (2022) ‘DNA methylation-free Arabidopsis reveals crucial roles of DNA methylation in regulating gene expression and development’, *Nature Communications*, 13(1), p. 1335. Available at: https://doi.org/10.1038/s41467-022-28940-2.

James, J., Castellano, D. and Eyre-Walker, A. (2017) ‘DNA sequence diversity and the efficiency of natural selection in animal mitochondrial DNA’, *Heredity*, 118(1), pp. 88–95. Available at: https://doi.org/10.1038/hdy.2016.108.

Jiao, W.-B. and Schneeberger, K. (2020) ‘Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics’, *Nature Communications*, 11(1), p. 989. Available at: https://doi.org/10.1038/s41467-020-14779-y.

Jupe, F. *et al.* (2013) ‘Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations’, *The Plant Journal: For Cell and Molecular Biology*, 76(3), pp. 530–544. Available at: https://doi.org/10.1111/tpj.12307.

Karasov, T.L. *et al.* (2014) ‘The long-term maintenance of a resistance polymorphism through diffuse interactions’, *Nature*, 512(7515), pp. 436–440. Available at: https://doi.org/10.1038/nature13439.

Karasov, T.L., Horton, M.W. and Bergelson, J. (2014) ‘Genomic variability as a driver of plant–pathogen coevolution?’, *Current Opinion in Plant Biology*, 18, pp. 24–30. Available at: https://doi.org/10.1016/j.pbi.2013.12.003.

Kawakatsu, T. *et al.* (2016) ‘Epigenomic diversity in a global collection of Arabidopsis thaliana accessions’, *Cell*, 166(2), pp. 492–505. Available at: https://doi.org/10.1016/j.cell.2016.06.044.

Kenchanmane Raju, S.K. *et al.* (2023) ‘Epigenomic divergence correlates with sequence polymorphism in Arabidopsis paralogs’, *New Phytologist*, n/a(n/a). Available at: https://doi.org/10.1111/nph.19227.

Kosakovsky Pond, S.L. and Frost, S.D.W. (2005) ‘Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection’, *Molecular Biology and Evolution*, 22(5), pp. 1208–1222. Available at: https://doi.org/10.1093/molbev/msi105.

Krueger, F. and Andrews, S.R. (2011) ‘Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications’, *Bioinformatics (Oxford, England)*, 27(11), pp. 1571–1572. Available at: https://doi.org/10.1093/bioinformatics/btr167.

Kryazhimskiy, S. and Plotkin, J.B. (2008) ‘The Population Genetics of dN/dS’, *PLOS Genetics*, 4(12), p. e1000304. Available at: https://doi.org/10.1371/journal.pgen.1000304.

Kuang, H. *et al.* (2004) ‘Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce’, *The Plant Cell*, 16(11), pp. 2870–2894. Available at: https://doi.org/10.1105/tpc.104.025502.

Lee, R.R.Q. and Chae, E. (2020) ‘Variation Patterns of NLR Clusters in Arabidopsis thaliana Genomes’, *Plant Communications*, 1(4), p. 100089. Available at: https://doi.org/10.1016/j.xplc.2020.100089.

Lynch, M. (2010) ‘Evolution of the mutation rate’, *Trends in genetics: TIG*, 26(8), pp. 345–352. Available at: https://doi.org/10.1016/j.tig.2010.05.003.

MacQueen, A. *et al.* (2019) ‘Population Genetics of the Highly Polymorphic RPP8 Gene Family’, *Genes*, 10(9), p. 691. Available at: https://doi.org/10.3390/genes10090691.

Märkle, H., Saur, I.M.L. and Stam, R. (2022) ‘Evolution of resistance (R) gene specificity’, *Essays in Biochemistry*, 66(5), pp. 551–560. Available at:

https://doi.org/10.1042/EBC20210077.
Martin, E.C. *et al.* (2020) 'LRRpredictor—A New LRR Motif Detection Method for Irregular Motifs of Plant NLR Proteins Using an Ensemble of Classifiers', *Genes*, 11(3), p. 286. Available at: https://doi.org/10.3390/genes11030286.
Martincorena, I. and Luscombe, N.M. (2013) 'Non-random mutation: The evolution of targeted hypermutation and hypomutation', *BioEssays*, 35(2), pp. 123–130. Available at: https://doi.org/10.1002/bies.201200150.
Martincorena, I., Seshasayee, A.S.N. and Luscombe, N.M. (2012) 'Evidence of non-random mutation rates suggests an evolutionary risk management strategy', *Nature*, 485(7396), pp. 95–98. Available at: https://doi.org/10.1038/nature10995.
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 Research*, 8(11), pp. 1113–1130. Available at: https://doi.org/10.1101/gr.8.11.1113.
Monroe, J.G. *et al.* (2022) 'Mutation bias reflects natural selection in Arabidopsis thaliana', *Nature*, pp. 1–5. Available at: https://doi.org/10.1038/s41586-021-04269-6.
Murrell, B. *et al.* (2012) 'Detecting Individual Sites Subject to Episodic Diversifying Selection', *PLOS Genetics*, 8(7), p. e1002764. Available at: https://doi.org/10.1371/journal.pgen.1002764.
Ngou, B.P.M., Ding, P. and Jones, J.D.G. (2022) 'Thirty years of resistance: Zig-zag through the plant immune system', *The Plant Cell*, pp. 1447–1478. Available at: https://doi.org/10.1093/plcell/koac041.
Nordborg, M. *et al.* (2005) 'The Pattern of Polymorphism in Arabidopsis thaliana', *PLOS Biology*, 3(7), p. e196. Available at: https://doi.org/10.1371/journal.pbio.0030196.
Oztas, O. *et al.* (2018) 'Genome-wide excision repair in Arabidopsis is coupled to transcription and reflects circadian gene expression patterns', *Nature Communications*, 9(1), p. 1503. Available at: https://doi.org/10.1038/s41467-018-03922-5.
Pond, S.L.K. *et al.* (2006) 'Adaptation to Different Human Populations by HIV-1 Revealed by Codon-Based Analyses', *PLOS Computational Biology*, 2(6), p. e62. Available at: https://doi.org/10.1371/journal.pcbi.0020062.
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', *The Plant Cell*, 33(4), pp. 998–1015. Available at: https://doi.org/10.1093/plcell/koab013.
Quadrana, L. *et al.* (2016) 'The Arabidopsis thaliana mobilome and its impact at the species level', *eLife*. Edited by D. Zilberman, 5, p. e15716. Available at: https://doi.org/10.7554/eLife.15716.
Quadrana, L. *et al.* (2019) 'Transposition favors the generation of large effect mutations that may facilitate rapid adaption', *Nature Communications*, 10(1), p. 3421. Available at: https://doi.org/10.1038/s41467-019-11385-5.
Quiroz, D. *et al.* (2022) 'The H3K4me1 histone mark recruits DNA repair to functionally constrained genomic regions in plants'. bioRxiv, p. 2022.05.28.493846. Available at: https://doi.org/10.1101/2022.05.28.493846.
Siol, M. *et al.* (2022) 'EggLib 3: A python package for population genetics and genomics', *Molecular Ecology Resources*, 22(8), pp. 3176–3187. Available at: https://doi.org/10.1111/1755-0998.13672.
Staunton, P.M., Peters, A.J. and Seoighe, C. (2023) 'Somatic mutations inferred from RNA-seq data highlight the contribution of replication timing to mutation rate variation in a model

plant', *Genetics*, p. iyad128. Available at: https://doi.org/10.1093/genetics/iyad128.
Suyama, M., Torrents, D. and Bork, P. (2006) 'PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments', *Nucleic Acids Research*, 34(Web Server), pp. W609–W612. Available at: https://doi.org/10.1093/nar/gkl315.
Thrall, P.H. *et al.* (2012) 'Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation', *Ecology Letters*, 15(5), pp. 425–435. Available at: https://doi.org/10.1111/j.1461-0248.2012.01749.x.
Van de Weyer, A.-L. *et al.* (2019) 'A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana', *Cell*, 178(5), pp. 1260-1272.e14. Available at: https://doi.org/10.1016/j.cell.2019.07.038.
Wang, L. *et al.* (2017) 'Entropy is a Simple Measure of the Antibody Profile and is an Indicator of Health Status: A Proof of Concept', *Scientific Reports*, 7(1), p. 18060. Available at: https://doi.org/10.1038/s41598-017-18469-6.
Wei, L. (2020) 'Selection On synonymous Mutations Revealed by 1135 Genomes of Arabidopsis thaliana', *Evolutionary Bioinformatics Online*, 16, p. 1176934320916794. Available at: https://doi.org/10.1177/1176934320916794.
Weng, M.-L. *et al.* (2019) 'Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis thaliana', *Genetics*, 211(2), pp. 703–714. Available at: https://doi.org/10.1534/genetics.118.301721.
Williams, B.P. *et al.* (2022) 'Somatic DNA demethylation generates tissue-specific methylation states and impacts flowering time', *The Plant Cell*, 34(4), pp. 1189–1206. Available at: https://doi.org/10.1093/plcell/koab319.
Xia, J., Han, L. and Zhao, Z. (2012) 'Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome', *BMC Genomics*, 13(8), p. S7. Available at: https://doi.org/10.1186/1471-2164-13-S8-S7.