Resistify - A rapid and accurate annotation tool to identify NLRs and study their genomic organisation

Moray Smith[1,2] – moray.smith@hutton.ac.uk
John T. Jones[1,2] – john.jones@hutton.ac.uk
Ingo Hein[1,3] – ingo.hein@hutton.ac.uk

[1] Cell and Molecular Sciences Department, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA
[2] School of Biology, University of St Andrews, St Andrews, KY16 9ST
[3] School of Life Sciences, The University of Dundee, Dundee

Abstract

Nucleotide-binding domain Leucine-rich Repeat (NLR) proteins are a key component of the plant innate immune system. In plant genomes NLRs exhibit considerable presence/absence variation and sequence diversity. Recent advances in sequencing technologies have made the generation of high-quality novel plant genome assemblies considerably more straightforward. Accurately identifying NLRs from these genomes is a prerequisite for improving our understanding of NLRs and identifying potential novel sources of disease resistance. Whilst several tools have been developed to predict NLRs, they are hampered by low accuracy, speed, and availability. Here, the NLR annotation tool Resistify is presented. Resistify is an easy-to-use, rapid, and accurate tool to identify and classify NLRs from protein sequences. It can correctly identify NLRs from a diverse range of species. Applying Resistify in combination with tools to identify transposable elements to a panel of *Solanaceae* genomes reveals a previously undescribed association between NLRs and Helitron transposable elements.

Availability: https://github.com/SwiftSeal/resistify

Introduction

The plant innate immune system is an essential part of plant disease resistance. A key component of this are the Nucleotide-binding domain Leucine-rich Repeat (NLR) proteins – an abundant and diverse family present in plant and animal genomes. NLRs act as intracellular immune receptors which elicit an immune response after the detection of pathogen derived factors or their activity (Biezen and Jones, 1998). Since their discovery, hundreds of functional NLRs have been cloned and many sequences have been identified (Kourelis and van der Hoorn, 2018). The abundance of NLRs and their sequence diversity shows considerable variation across the plant kingdom (Barragan and Weigel, 2021).

One driving force of NLR diversification is thought to be the proliferation of transposable elements (TEs). In *Arabidopsis*, sites of high TE insertion frequency have been previously identified to contain NLRs (Quadrana *et al.*, 2016), and NLRs exhibiting elevated intraspecific diversity are in close proximity to TEs (Sutherland *et al.*, 2023). Of the TEs, LTR-retrotransposons are the most prolific in plant genomes and are a major factor in genome expansion (Lee and Kim, 2014). Strikingly, the *Capsicum annuum* genome exhibits a marked expansion of NLRs nested within LTR-retrotransposons, and this effect is seen in other *Solanaceae* including tomato and potato (Seong *et al.*, 2020). Other transposable elements, such as the Helitrons, can also capture and transport genes across the plant genome (Lai *et al.*, 2005).

Canonical NLRs have a modular organisation. Central to the NLR structure is the highly conserved NB-ARC domain which regulates the activity of the protein. Downstream of this is the Leucine-Rich-Repeat (LRR) domain which has a role in ligand binding and auto-inhibition. Upstream of the NB-ARC domain is variable. Often, there is either a coiled-coil (CC), Resistance to Powdery Mildew 8-like (RPW8), or Toll/Interleukin-1 receptor/Resistance protein (TIR) domain. Together, this modular structure means that NLRs can be separated into different classifications: CNL, RNL, TNL, or NL if lacking a conserved N-terminal domain.

Recently, additional NLR associated signatures have been identified. For example, the post-LRR C-terminal jelly-roll/Ig-like domain (C-JID) (Martin *et al.*, 2020) may be present in TNLs and appears to bind with pathogen effectors, thereby contributing to the immune response. Similarly, other NLRs have integrated decoys to detect pathogen activities. In terms of initating cell death through the formation of a resistosome, CNLs rely on an N-terminal MADA motif, a hallmark of NRCs - NLRs which act as intermediaries in the NLR induced immune response or NLRs that function independently of NRCs (Adachi *et al.*, 2019).

Due to their conservation and modular structure, NLRs lend themselves to automated identification and classification. To date, several tools have been developed to achieve this including DRAGO2, NLGenomeSweeper, NLR-Annotator, RGAugury, RRGPredictor, and NLRtracker (table 1) (Li *et al.*, 2016; Osuna-Cruz *et al.*, 2018; Santana Silva and Micheli, 2020; Steuernagel *et al.*, 2020; Toda *et al.*, 2020; Kourelis *et al.*, 2021). The development of the RefPlantNLR database has allowed the benchmarking of these tools against the sequences of functionally characterised NLRs (Kourelis *et al.*, 2021). Performance varied between tools with NLRtracker being the most sensitive and accurate. All tools perform well at the identification of TIR, RPW8, and NB-ARC domains which are highly conserved, but are often less accurate in predicting CC domains which are more variable and frequently missed by InterProScan (Jones *et al.*, 2014). Additional to the currently available NLR classification tools is NLRexpress – a set of machine learning predictors for CC, TIR, NB-ARC, and LRR motifs (Martin *et al.*, 2022). Although NLRexpress does not directly identify or classify NLRs, it is well suited for rapidly and accurately screening large sets of sequences for NLR-associated motifs. One strength of NLRexpress is the CC-associated extended EDVID model which performs strongly in identifying this challenging motif.

Table 1: A summary of the currently available NLR annotation tools. Adapted from (Kourelis et al., 2021).

Tool	Input data	Method	Output	Distribution

DRAGO2	Protein, Transcript	Custom HMMs	Classification, Domains	Online only, API available
NLGenomeSweeper	Transcript, Genomic	InterProScan, MUSCLE, TransDecoder, BLAST, HMMER	Classification, Genome position, GFF annotation	GitHub, Manual dependency installation
NLR-Annotator	Transcript, Genomic	MEME motifs	Classification, Genome position, GFF annotation	GitHub
RGAugury	Transcript, Genomic	InterProScan	Classification, Genome position, GFF annotation	GitHub, Online or local webservice, Docker container
RRGPredictor	Protein, Transcript	InterProScan	Classification	GitHub, Manual dependency installation
NLRtracker	Protein, Transcript	InterProScan, HMMER, MEME motifs	Classification, NB-ARC sequence, Domains, GFF annotation	GitHub, Manual dependency installation
NLRexpress	Protein	HMMER, Scikit-learn	Motif position, Motif sequence	GitHub, Online or local,

	Conda environment provided

One drawback with the currently available tools is the reliance on InterProScan as the backend for domain annotation. InterProScan is designed as a comprehensive and generalised domain annotation tool. As a result, each input sequence is scanned against several databases each containing in total more than 180,000 protein signatures, the vast majority of which are not NLR associated. Compounding this, NLRs only represent a fraction of a plant proteome resulting in unnecessary searches against non-NLR sequences. In addition, domains common to NLRs are present in InterProScan databases with differing levels of curation which often results in overlapping or fragmented annotations. These must be parsed, particularly the LRR domain which is represented by multiple InterProScan signatures. The large databases that InterProScan relies upon must also be downloaded which provides an additional challenge for installation or automated deployment.

Software that is readily available with minimum manual configuration is a requirement for genome assembly and annotation projects. Projects may use an array of different tools, making manual installation and dependency resolution time-consuming. Dependency resolution solutions such as Conda, and automated deployment of tools in pipelines with Snakemake or Nextflow are becoming increasingly common (Di Tommaso *et al.*, 2017; Mölder *et al.*, 2021). Additionally, bioinformatics research is often restricted to shared High-Performance Clusters (HPCs) where users have limited privileges and where storage is at a premium. Tools that require root privileges, are distributed as web services, or require large databases are often not conducive for genomics projects.

Here, the new NLR annotation tool Resistify is presented that overcomes some of the constraints of currently available tools. We show that this tool accurately predicts NLR sequences from diverse plant using a range of well curated sources and can be used for pangenome analysis of NLRs in *Solanaceous* genomes. In addition, we use Resistify in

combination with the EDTA tool to investigate the genomic organisation of NLRs and their association with transposable elements.

Methods

Resistify is implemented in python as a command line executable. First, Resistify executes a hmmsearch of input protein sequences against a custom database of HMMs derived from curated Pfam entries (table 2). These models are used to identify CC, RPW8, TIR, and NB-ARC domains, as well as C-JID and MADA motifs. Domains of the same type are merged if they overlap or are within 100 amino acids of each other. Preliminary testing in development showed that this is necessary to overcome NB-ARC domain annotations which can become split. MADA motifs are restricted to the N-terminal of the sequence. Using this evidence, proteins are initially classified as belonging to either a CN, RN, TN, or N classification. Proteins which do not have any evidence of an NB-ARC domain are discarded at this stage.

Table 2: HMM models including in the initial hmmsearch stage of Resistify.

Domain	Source
NB-ARC	https://www.ebi.ac.uk/interpro/entry/pfam/PF00931/
TIR	https://www.ebi.ac.uk/interpro/entry/pfam/PF01582/
TIR	https://www.ebi.ac.uk/interpro/entry/pfam/PF13676/
RPW8	https://www.ebi.ac.uk/interpro/entry/pfam/PF05659/
CC	https://www.ebi.ac.uk/interpro/entry/pfam/PF18052/
C-JID	https://www.ebi.ac.uk/interpro/entry/pfam/PF20160/
MADA	https://cdn.elifesciences.org/articles/49956/elife-49956-supp2-v2.hmm

Following this, the filtered set of NB-ARC containing proteins is screened for NLR-associated motifs with the NLRexpress models. Sequences are then reclassified using the following logic. If a protein belongs to class N (i.e., does not have any evidence of an upstream CC, RPW8, or TIR domain) it is scanned for upstream TIR or CC motifs. As motif searches can be more promiscuous, restricting motif searches to this condition

prevents them from interfering with non-ambiguous NLR classification. The sequence is then screened for LxLxxL motifs to define the LRR domain. Following a previous definition, an LRR domain is annotated if four or more LxLxxL motifs are identified with inter-motif gaps of less than 75 amino acids [@martin_nlrscape_2023]. Gaps larger than 75 amino acids are predicted to be a break in the LRR domain and so the LRR annotation process is restarted from that position onwards. If less than four motifs exist across the whole sequence this process is skipped.

This combined evidence is then integrated into the domain annotation data. The domains are sorted by start position and a “domain string” is formed. For example, if the sorted domains took the order TIR, NB-ARC, LRR, C-JID, then the domain string would be TNLj. Alternatively, a canonical CNL with a MADA motif would take the form mCNL. The domain strings are searched for substrings CNL, RNL, TNL, or NL and classified accordingly. Sequences are further classified on the basis of whether they contain an N-terminal bound MADA motif or CJID domain. Additionally, a count of the unique occurrences of the conserved NB-ARC motifs is made as an indicator of NB-ARC integrity.

The primary output of Resistify is a table detailing the NLRs identified and the specific classification for each sequence. The complete domain string, classification, NB-ARC motif count, and MADA and CJID status are listed. A complete list of all annotations and NLRexpress motifs are given as separate tables. Additionally, all NB-ARC domain sequences are extracted and provided as a FASTA file.

Resistify is provided as a single executable available for download on GitHub, PyPi, and Conda. All databases and models are distributed with the executable so that manual setup is not required. All output files are placed in a nameable output directory for easy integration into automated workflows. Temporary files are also handled internally to reduce output clutter. Resistify requires scikit-learn v0.24.2 and hmmer v>=3.0.

RefPlantNLR benchmarking

Protein sequences of the RefPlantNLR database members were retrieved and used as input for Resistify with default settings (Kourelis *et al.*, 2021). Resistify classifications were compared directly in R v4.3.2 with tidyverse v2.0.0. Any sequence where the Resistify classification did not exactly match with the provided RefPlantNLR structure was taken as a potential misclassification.

Araport11 benchmarking

The latest release of Araport11 representative gene model protein sequences were downloaded from TAIR and used as input for Resistify with default settings(Lamesch *et al.*, 2012). A phylogenetic tree was built from the Resistify-extracted NB-ARC domain sequences with mafft v7.52.0 and fasttree v2.1.11 with default settings.

Pangenome pipeline

A Snakemake workflow was developed to predict genes and NLRs in a Solanum pangenome comprised of chromosome-scale genomes. Snakemake was executed in a mamba v1.4.2 environment with snakemake v7.32.3 and cookiecutter v1.7.3 (Mölder *et al.*, 2021). Genes were predicted *de-novo* from the genome alone using Helixer v0.3.2 with the land_plant_v0.3_a_0080.h5 model (Holst *et al.*, 2023). Protein sequences were extracted using AGAT v1.2.0 and used as input for Resistify (Dainat *et al.*, 2023). Ortholog analysis was carried out using OrthoFinder v2.5.5 using all predicted protein sequences (Emms and Kelly, 2019). Transposable elements were annotated with the latest GitHub release of EDTA (Ou *et al.*, 2019).

Gene models were translated to bed format using AGAT v1.2.0 and overlaps with intact transposable elements were identified using bedtools v2.31.1 with the command bedtools intersect using the options -f 0.9 -wo (Quinlan and Hall, 2010). To identify

202 previously characterised NLRs, protein sequences of a subset of *Solanum*-originating
203 NLRs in the RefPlantNLR database were queried against each genome with blastp
204 v2.15.0 (Camacho *et al.*, 2009). The full pipeline and all post-hoc analysis are available
205 at https://github.com/SwiftSeal/pangenomics/.
206

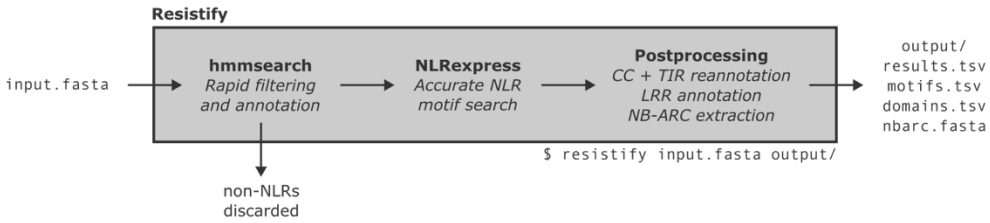
207 Results

208

209 Performance against RefPlantNLR

210

211 To evaluate the performance of Resistify (figure 1), it was applied against the
212 RefPlantNLR database - a curated set of 415 previously cloned NLRs from a diverse
213 range of species. NLRtracker was also applied to the RefPlantNLR database for
214 comparison.
215



216
217 **Figure 1: The Resistify program.**
218

219 In the default mode, only three RefPlantNLR entries were not identified by Resistify -
220 *AtNRG1.3* which carries only an LRR domain and *Pb1* and *RXL* which have significantly
221 truncated NB-ARC domain that are not listed in the RefPlantNLR database. However, if
222 required, these genes can be identified with Resistify using –ultra mode which skips the
223 initial filtering stage and reports and sequence which contains at least one NLRexpress
224 motif. Sequences are reported as an unmerged string of NLR motifs. Consequently,
225 *AtNRG1.3*, *Pb1*, and *RXL* are reported as NNLLLLLLL, CNNNLLLLLLLLLLL, and
226 CNNNLLLLLLLLL respectively.
227

228 The largest source of variation between RefPlantNLR and Resistify classification was 29
229 NLRs which had an NL structure according to RefPlantNLR but CNL according to
230 Resistify. All of these belonged to CNL-associated subclasses which are known to have
231 CC domains that are challenging to identify. This included 15 CC_{G10}-NLRs including
232 *Pvr4*, *Tsw*, *RPS2*, *RPS5*, *SUT1* and *SUMM2* which have previously been noted for their
233 lack of the CC-associated EDVID motif (Lee *et al.*, 2021). Others included six members
234 of the *Pm5* locus which despite not containing a CC domain in RefPlantNLR, have
235 previously been identified to contain CC-like domains (Xie *et al.*, 2020). This analysis
236 demonstrates that Resistify is highly sensitive at retrieving canonical NLRs and
237 accurately describing their structure.
238
239 As NLRtracker does not exclude NLRs without NB-ARC domains, *AtNRG1.3*, *Pb1* and
240 *RXL* were identified and classified as "CC-NLR or CCR-NLR or CCG10-NLR", "CC-NLR",
241 and "CC-NLR" respectively. NLRtracker relies on the NB-ARC associated RNBS-D motif
242 to classify NLRs as "CC-NLR or CCR-NLR or CCG10-NLR". As a result, 26 of the
243 conflicting CNLs identified by Resistify were classified as "CC-NLR or CCR-NLR or
244 CCG10-NLR" despite failing to identify a CC domain. The exception of this was for
245 *SpNBS-LRR*, *Rpp1-R1*, and *Rpp4C4* which were classified as "UNDETERMINED".
246 Reliance on RNBS-D also led to classification of *NtTPN1* as "CC-NLR or CCR-NLR or
247 CCG10-NLR". Whilst *NtTPN1* is a CNL ortholog, it lacks any upstream domain and is
248 structurally an NL. Unexpectedly, the TNL *DSC1* was misclassified as "CC-NLR" despite
249 having a domain structure of "(TIR)(NBARC)(LRR)(CJID)" according to NLRtracker.
250
251 In summary, Resistify performs well at identifying canonical NLRs from a diverse range
252 of species. In default mode, it does not assign genes as NLRs with extremely truncated
253 or entirely absent NB-ARC domains, unlike NLRtracker which reports any sequence with
254 NLR-associated domains. However, this can be replicated in Resistify with the –ultra
255 mode. Thus, Resistify's structure-based classification method is well suited for
256 correctly classifying NLRs, including members of challenging CNL subclasses.
257

Performance against the Araport11 proteome

To assess the performance of Resistify across a well characterised and annotated genome, the representative gene models of Araport11 were analysed (Cheng *et al.*, 2017). In total, Resistify identified 166 NLRs - the majority of which were TNLs and CNLs (figure 2a). Of the CNLs, 25% had a MADA motif, and 44.4% and 41.2% of NLs and TNLs had C-JID domain respectively. Partial NLRs either without an N-terminal or LRR domain were also identified.

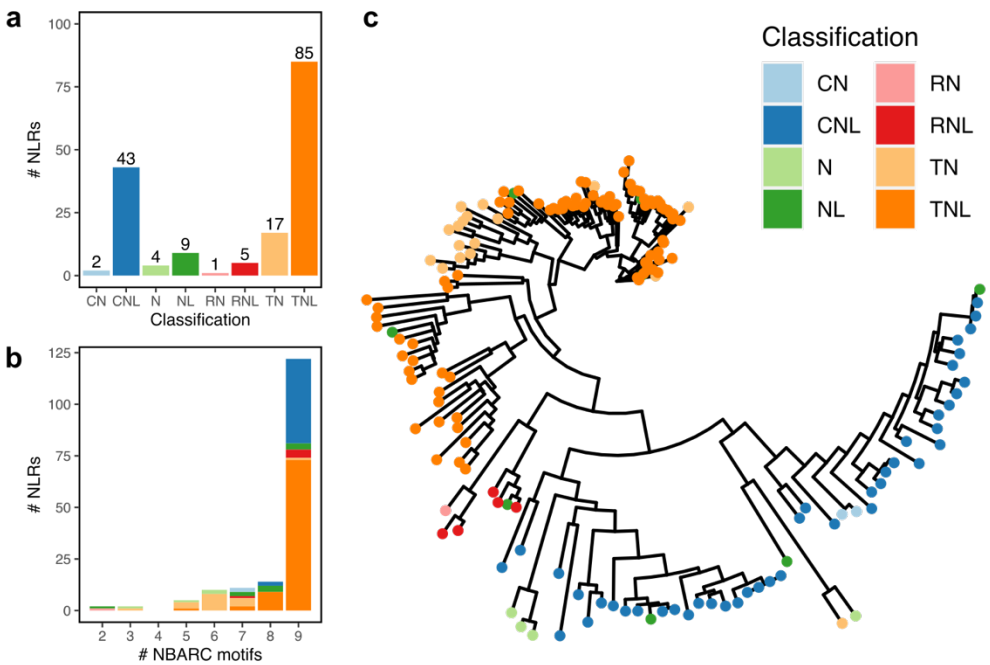


Figure 2: Resistify applied to Araport11.
a) The number of NLRs belonging to each identified classification. b) The number of NLRs grouped by number of identified unique NB-ARC associated motifs. c) A phylogenetic tree of NLRs based on the Resistify extracted NB-ARC domain sequences.

Manual inspection of the motifs within these sequences confirmed that they were not due to a failure to identify these elements. The majority of NLRs carried all nine conserved NB-ARC motifs - those with fewer increasingly belonged to partial NLR classifications (figure 2b). NLRs with as few as two of the conserved NB-ARC motifs were successfully identified. A phylogenetic tree of the Resistify-identified NB-ARC

domains independently validated the classifications and allowed the placement of ambiguous NL or N classified sequences into subclasses (figure 2c).

NLRtracker identified an additional 48 sequences, however these sequences did not contain any NB-ARC domain annotation according to either tool. Three sequences - AT4G19060.1, AT4G19060.1, and AT5G45440.1 - were not identified by NLRtracker. According to Resistify these contained a single NB-ARC domain each, with 5, 6, and 6 NB-ARC motifs respectively. Overall, Resistify performs well at identifying and classifying NLRs from whole proteomes and successfully retrieves NB-ARC domain annotations for phylogenetic analyses.

Application against an example workflow

To demonstrate how Resistify might be implemented to identify novel resistance genes, a pangenome experiment was performed. Eighteen contiguous *Solanum* genomes were downloaded and processed with a simple workflow which predicts genes de novo, identifies orthologues, and classifies NLRs with Resistify (Tang *et al.*, 2022; Li *et al.*, 2023). For gene annotations, the recently developed tool Helixer was selected for its near reference quality predictions and lack of requirement for repeat-masking which is known to result in false-negative NLR annotations (Bayer, Edwards and Batley, 2018; Holst *et al.*, 2023). The workflow also predicts transposable elements with EDTA. The *C. annuum* genome was included as an outgroup, but also because recent analysis has suggested a substantial expansion of NLRs in this species associated with LTR transposable elements (Kim *et al.*, 2017).

Predicted gene content varied between genomes from 29,223 in *S. habrochaites* to 61,015 in the larger *C. annuum* genome (table S1). Transposable element content ranged from 55.6% in *S. chmielewskii* to 76.8% in *C. annuum*. Whilst there was no significant difference in total TE content between tuber-bearing and non-tuber-bearing genomes ($p = 0.644$), there was a significant increase in the number of intact TEs reported by EDTA in the genomes of tuber-bearing species ($p = 0.008$).

In total, 8144 NLRs were identified across all genomes (figure 3). CNLs were the most abundant classification of NLR identified, ranging from 84 in *S. habrochaites* to 422 in *S. tuberosum* (group tuberosum RH10-15). There was a notable expansion of NLRs in tuber-bearing *Solanum* species. This is in agreement with the previous observation that potato-bearing *Solanum* species have an expansion of tuber expressed NLRs (Tang *et al.*, 2022). This effect does not correlate with the proportion of genome occupied by transposable elements of any classification (figure S1).

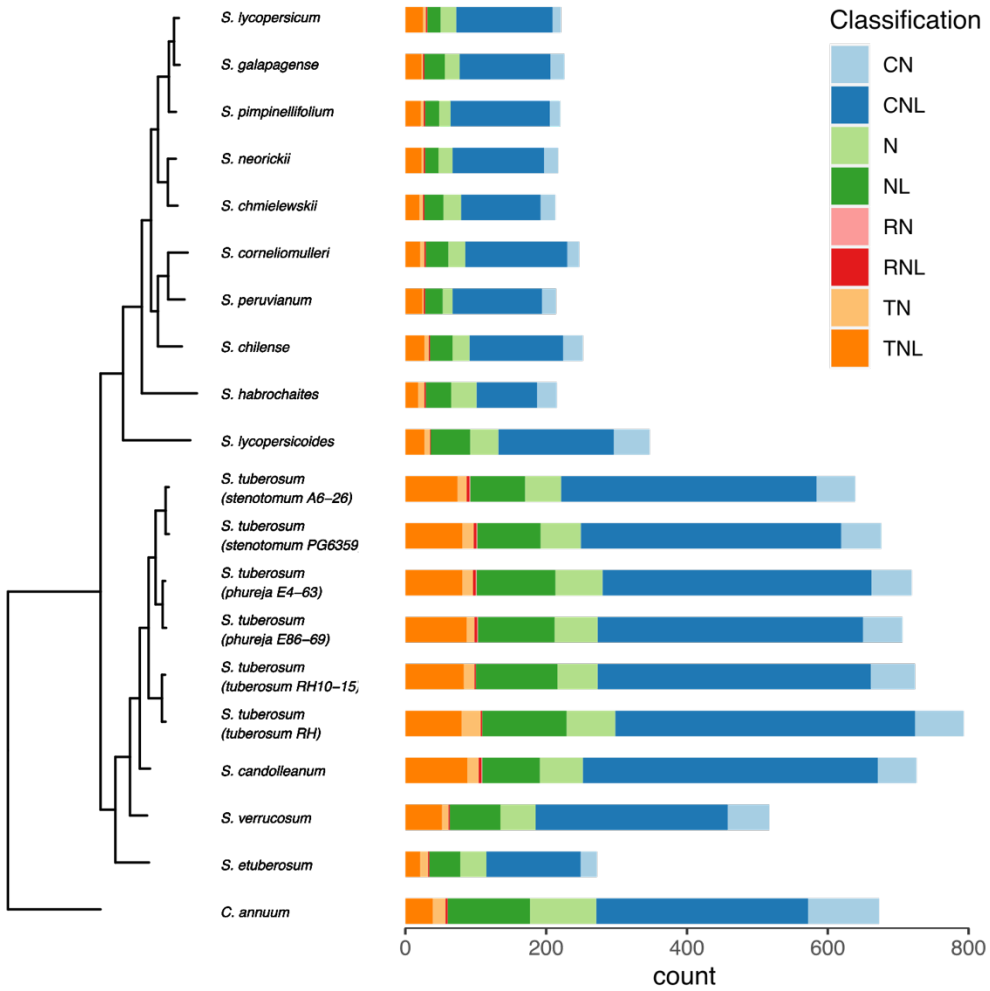


Figure 3: NLRs identified across the *Solanum* pangenome.

In total, 38,590 orthogroups were identified, of which 687 (1.8%) contained NLRs. The distribution of orthogroups showed a clear divide between core and shell/cloud orthogroups within the pangenome with the majority of NLR orthogroups existing within

the shell/cloud (fig 4a). This reflects the relatively large width of the pangenome, which captures NLR variation over a Genus level. Species specific NLRomes also exhibit an abundance of cloud orthographs, reflecting the high variability of NLRs in genomes (Van de Weyer *et al.*, 2019). Many (48.1%, n = 371) orthogroups were classified as containing N or NL NLRs (figure 4b). Whilst these classifications are less abundant (24.6%, n = 2081), they were often associated with smaller orthogroups contributing to this inflation.

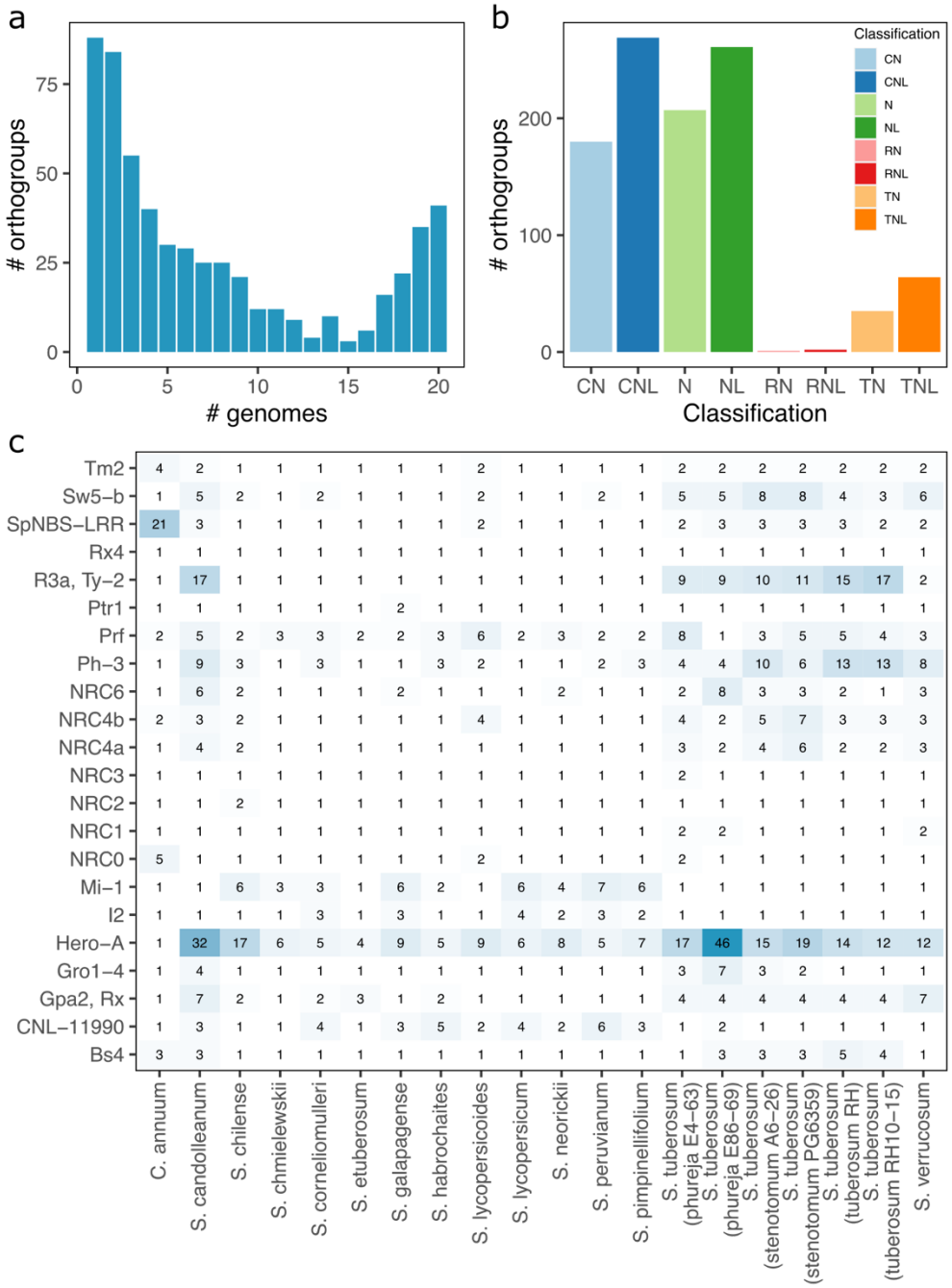


Figure 4: NLR-containing orthogroups.

a) The number of orthogroups shared between genomes. b) The number of orthogroups according to NLR classification. c) The number of orthologs of known *Solanaceae* NLRs identified in each genome.

The distribution of previously identified *Solanaceae* NLR orthologs across the pangenome was examined (figure 4c). In agreement with previous findings, members of the NRC group remained relatively stable across the pangenome except for NRC4a, NRC4b, and NRC6 which were expanded in the tuber-bearing genomes. By contrast, an expansion of *Mi-1* and *Hero* was evident in the non-tuber bearing and tuber bearing genomes respectively. At least one ortholog was identified in each genome for all genomes.

It has previously been reported that there is a vast expansion of NLRs in the *C. annum* genome due to retroduplication; NLRs nested within LTRs represented a large proportion (~13%) of NLRs within the genome, and this effect is also seen in tomato (8%) and potato (18%) (Kim *et al.*, 2017). To explore whether this effect could be linked to the expansion of NLRs within the tuberising members of *Solanum*, a similar analysis was repeated. Intact TEs were identified and considered to interact with NLRs if they covered >90% of the gene annotation.

Unexpectedly, the effect could not be replicated and across all genomes only five intact LTRs were identified as containing NLRs (figure 5a). None were identified in *C. annuum* but instead in the *S. tuberosum* group Phureja and clone RH. The putative retrotransposed NLRs within these all belonged to the same orthogroup (OG0000639) which was expanded in both Phureja and RH, but not *S. stenotomum*.

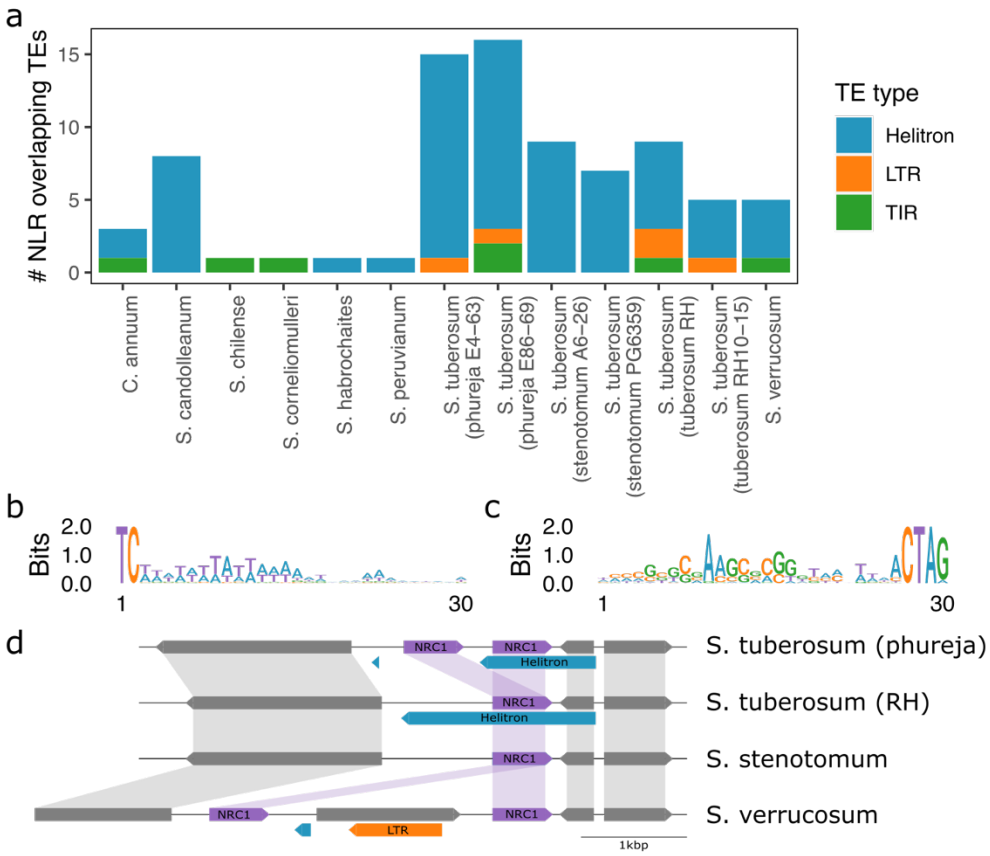


Figure 5: Helitrons associated with NLRs.

a) The number of intact TEs overlapping with NLRs across the pangenome. b) The 5' motif of NLR overlapping Helitrons. c) The 3' motif of NLR overlapping Helitrons. d) The *NRC1* locus and its association with Helitrons across four genomes. Helitron (blue) and LTR (orange) transposable elements are highlighted.

By contrast, a proliferation of Helitron-associated NLRs was identified in the tuber-bearing genomes. All predicted Helitrons carried the expected 5'-TC...CTRR-3' signature as well as a GC-rich region in the 3' terminal (figure 5b,c). An example of an NLR which may have undergone Helitron-associated duplication is *NRC1* (figure 5d) which is present as a single copy except in both *S. tuberosum* group Phureja genomes and *S. verrucosum*. Close examination of the *NRC1* locus reveals a duplication event in *S. tuberosum* group Phureja where one *NRC1* gene is nested entirely within a predicted Helitron. In *S. tuberosum* clone RH which only has one copy of *NRC1*, *NRC1* is similarly nested within a Helitron although this appears to be extended to a distant Helitron terminator past the site of duplication in group Phureja. In *S. stenotomum*, only one

NRC1 is present and no Helitron predicted. Interestingly, *S. verrucosum* has two copies of *NRC1* and whilst it does not have a nested Helitron copy, the leftmost *NRC1* copy does have a short predicted Helitron upstream of its start position in the same orientation as the other Helitron elements. In this case, the two copies of *NRC1* appear to have been further separated by an intact LTR insertion which has resulted in an additional gene prediction.

Discussion

Resistify is presented as a highly accurate and easy-to-use tool to aid the identification of novel resistance genes. Applying Resistify against the RefPlantNLR database demonstrates that it is effective at identifying NLRs from a diverse range of plant genomes. It is highly sensitive at classifying challenging CNL families which often evade other NLR classification tools due to the inclusion of the performant NLRexpress motif models. Members of the CC*G10*-NLR have previously been described as lacking the CC-associated EDVID motif and are often predicted as having an NL structure in the RefPlantNLR database (Lee *et al.*, 2021). By utilising NLRexpress' highly sensitive extended EDVID motif model, Resistify successfully classifies these elusive NLR clades, highlighting it as the most sensitive NLR classifier to date.

Resistify can be easily integrated into workflows and is scalable to large pangenome experiments. As high-quality genome assemblies become more common, providing accessible tools for the study of NLRs will be crucial for fully appreciating their diversity and identifying novel sources of resistance. The recent releases of chromosome level assemblies for potato and tomato pangenomes are a valuable resource for understanding the diversity of NLRs within *Solanum* (Tang *et al.*, 2022; Li *et al.*, 2023). An expansion of NLRs including essential families such as the NRCs is apparent in tuber-bearing species.

As Resistify relies on an initial NB-ARC domain search to reduce the search space prior to NLRexpress motif identification, NLRs with very truncated or entirely absent NB-ARC

408 domains can be missed. To resolve this, an additional mode is provided which does not
409 perform initial NB-ARC domain filtering which identifies sequences with any NLR
410 motifs. Unlike NLRtracker, Resistify does not search for noncanonical integrated
411 domains (Kroj *et al.*, 2016). As integrated domains are widespread and often critical to
412 NLR function, their identification and analysis are an important factor in studies of NLR
413 diversity (Barragan and Weigel, 2021). Pairing Resistify with a secondary sweep for these
414 domains with NLRtracker or InterProScan would permit this analysis whilst also
415 benefiting from a reduced search space and high-quality NLR classifications from
416 Resistify.
417
418 An unexpected finding of this study was the failure to replicate the previous observation
419 of an abundance of LTR associated NLRs across *Solanaceae* (Kim *et al.*, 2017).
420 Differences in gene and TE annotation and NLR classification likely contributed to this.
421 Although Helixer predicted ~70% more genes in *C. annuum* than were identified in the
422 previous study, this did not translate to an increase in NLRs identified. Instead, only 673
423 NLRs were identified here in comparison to the previous estimate of 835. Previously, a
424 tBLASTn search of the NB-ARC domain followed by ORF identification and BLASTP
425 searches against GenBank to classify NLRs was used to identify NLRs in the genome
426 (Seo *et al.*, 2016). This method may result in more false positives in comparison to
427 identifying NLRs directly from gene annotations. Here, EDTA was selected for identifying
428 TEs which has an improved sensitivity and selectivity for identifying intact LTRs in
429 comparison to LTRHarvest which was used previously. Although Helixer does not
430 require repeat masking which can result in false negative NLR annotations, annotations
431 used in training the model (which may have relied upon repeat masking) might
432 introduce TE-avoidant behaviour (Bayer, Edwards and Batley, 2018). However, the
433 identification of several LTR-associated NLRs indicates that this is unlikely to be the
434 source of the difference in the result.
435
436 The identification of Helitron-associated NLR expansion has not been previously
437 reported. Helitrons are challenging to identify due to their lack of structural elements
438 and as a result the pipeline used in this study is known to overestimate Helitron density
439 (Baril, Imrie and Hayward, 2022). However, a large part of this overestimation is likely

due to EDTA reannotating the genome with HelitronScanner predictions, which itself suffers from a high false positive rate, leading to a proliferation of fragmented Helitron annotations across the genome. For this study, only intact Helitrons which passed EDTAs stringent filter were considered. As a result, all predicted Helitrons had the required structural elements for activity. Further validation would be required to determine if Helitron association is a valid mechanism of NLR expansion. The Helitron/*NRC1* relationship highlighted here would be a good starting point for unpicking this mechanism.

Acknowledgements

The authors acknowledge the Research/Scientific Computing teams at The James Hutton Institute and NIAB for providing computational resources and technical support for the “UK’s Crop Diversity Bioinformatics HPC” (BBSRC grant BB/S019669/1), use of which has contributed to the results reported within this paper. The authors also acknowledge Peter Cock for his support with creating the conda distribution of Resistify.

Funding

This work was supported by the Rural & Environment Science & Analytical Services (RESAS) Division of the Scottish Government through the project JHI-B1-1, the Biotechnology and Biological Sciences Research Council (BBSRC) through awards BB/S015663/1 and BB/X009068/1. MS was supported through the East of Scotland Bioscience Doctoral Training Partnership (EASTBIO DTP), funded by the BBSRC award BB/T00875X/1.

References

Adachi, H. *et al.* (2019) 'An N-terminal motif in NLR immune receptors is functionally conserved across distantly related plant species', *eLife*. Edited by J.-M. Zhou, D. Weigel, and J.-M. Zhou, 8, p. e49956. Available at: https://doi.org/10.7554/eLife.49956.

Baril, T., Imrie, R.M. and Hayward, A. (2022) 'Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline'. bioRxiv, p. 2022.06.30.498289. Available at: https://doi.org/10.1101/2022.06.30.498289.

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.

Bayer, P.E., Edwards, D. and Batley, J. (2018) 'Bias in resistance gene prediction due to repeat masking', *Nature Plants*, 4(10), pp. 762–765. Available at: https://doi.org/10.1038/s41477-018-0264-0.

Biezen, E.A.V.D. and Jones, J.D.G. (1998) 'Plant disease-resistance proteins and the gene-for-gene concept', *Trends in Biochemical Sciences*, 23(12), pp. 454–456. Available at: https://doi.org/10.1016/S0968-0004(98)01311-5.

Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. Available at: https://doi.org/10.1186/1471-2105-10-421.

Cheng, C.-Y. *et al.* (2017) 'Araport11: a complete reannotation of the Arabidopsis thaliana reference genome', *The Plant Journal*, 89(4), pp. 789–804. Available at: https://doi.org/10.1111/tpj.13415.

Dainat, J. *et al.* (2023) 'NBISweden/AGAT: AGAT-v1.2.0'. Zenodo. Available at: https://doi.org/10.5281/zenodo.8178877.

Di Tommaso, P. *et al.* (2017) 'Nextflow enables reproducible computational workflows', *Nature Biotechnology*, 35(4), pp. 316–319. Available at: https://doi.org/10.1038/nbt.3820.

Emms, D.M. and Kelly, S. (2019) 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome Biology*, 20(1), pp. 1–14. Available at: https://doi.org/10.1186/s13059-019-1832-y.

Holst, F. *et al.* (2023) 'Helixer–de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model'. bioRxiv, p. 2023.02.06.527280. Available at: https://doi.org/10.1101/2023.02.06.527280.

Jones, P. *et al.* (2014) 'InterProScan 5: genome-scale protein function classification', *Bioinformatics*, 30(9), pp. 1236–1240. Available at: https://doi.org/10.1093/bioinformatics/btu031.

Kim, S. *et al.* (2017) 'New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication', *Genome Biology*, 18(1), p. 210. Available at: https://doi.org/10.1186/s13059-017-1341-9.

Kourelis, J. *et al.* (2021) 'RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family', *PLOS Biology*, 19(10), p. e3001124. Available at: https://doi.org/10.1371/journal.pbio.3001124.

Kourelis, J. and van der Hoorn, R.A.L. (2018) 'Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function', *The Plant Cell*, 30(2), pp. 285–299. Available at: https://doi.org/10.1105/tpc.17.00579.

Kroj, T. *et al.* (2016) 'Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread', *New Phytologist*, 210(2), pp. 618–626. Available at: https://doi.org/10.1111/nph.13869.

Lai, J. *et al.* (2005) 'Gene movement by Helitron transposons contributes to the haplotype variability of maize', *Proceedings of the National Academy of Sciences*, 102(25), pp. 9068–9073. Available at: https://doi.org/10.1073/pnas.0502923102.

Lamesch, P. *et al.* (2012) 'The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools', *Nucleic Acids Research*, 40(D1), pp. D1202–D1210. Available at: https://doi.org/10.1093/nar/gkr1090.

Lee, H. *et al.* (2021) 'Genome-wide functional analysis of hot pepper immune receptors reveals an autonomous NLR clade in seed plants', *The New Phytologist*, 229(1), pp. 532–547. Available at: https://doi.org/10.1111/nph.16878.

Lee, S.-I. and Kim, N.-S. (2014) 'Transposable Elements and Genome Size Variations in Plants', *Genomics & Informatics*, 12(3), pp. 87–97. Available at: https://doi.org/10.5808/GI.2014.12.3.87.

Li, N. *et al.* (2023) 'Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species', *Nature Genetics*, 55(5), pp. 852–860. Available at: https://doi.org/10.1038/s41588-023-01340-y.

Li, P. *et al.* (2016) 'RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants', *BMC Genomics*, 17(1), p. 852. Available at: https://doi.org/10.1186/s12864-016-3197-x.

Martin, E.C. *et al.* (2022) 'NLRexpress—A bundle of machine learning motif predictors—Reveals motif stability underlying plant Nod-like receptors diversity', *Frontiers in Plant Science*, 13. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2022.975888 (Accessed: 23 October 2023).

Martin, R. *et al.* (2020) 'Structure of the activated ROQ1 resistosome directly recognizing the pathogen effector XopQ', *Science*, 370(6521), p. eabd9993. Available at: https://doi.org/10.1126/science.abd9993.

Mölder, F. *et al.* (2021) 'Sustainable data analysis with Snakemake'. F1000Research. Available at: https://doi.org/10.12688/f1000research.29032.2.

Osuna-Cruz, C.M. *et al.* (2018) 'PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes', *Nucleic Acids Research*, 46(D1), pp. D1197–D1201. Available at: https://doi.org/10.1093/nar/gkx1119.

Ou, S. *et al.* (2019) 'Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline', *Genome Biology*, 20(1), p. 275. Available at: https://doi.org/10.1186/s13059-019-1905-y.

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.

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. Available at: https://doi.org/10.1093/bioinformatics/btq033.

Santana Silva, R.J. and Micheli, F. (2020) 'RRGPredictor, a set-theory-based tool for predicting pathogen-associated molecular pattern receptors (PRRs) and resistance (R) proteins from plants', *Genomics*, 112(3), pp. 2666–2676. Available at: https://doi.org/10.1016/j.ygeno.2020.03.001.

Seo, E. *et al.* (2016) 'Genome-Wide Comparative Analyses Reveal the Dynamic Evolution of Nucleotide-Binding Leucine-Rich Repeat Gene Family among Solanaceae Plants', *Frontiers in Plant Science*, 7. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2016.01205 (Accessed: 20 November 2023).

Seong, K. *et al.* (2020) 'Evolution of NLR resistance genes with noncanonical N-terminal domains in wild tomato species', *New Phytologist*, 227(5), pp. 1530–1543. Available at: https://doi.org/10.1111/nph.16628.

Steuernagel, B. *et al.* (2020) 'The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire1 [OPEN]', *Plant Physiology*, 183(2), pp. 468–482. Available at: https://doi.org/10.1104/pp.19.01273.

Sutherland, C.A. *et al.* (2023) 'High intraspecies allelic diversity in Arabidopsis NLR immune receptors is associated with distinct genomic and epigenomic features'. bioRxiv, p. 2023.01.12.523861. Available at: https://doi.org/10.1101/2023.01.12.523861.

Tang, D. *et al.* (2022) 'Genome evolution and diversity of wild and cultivated potatoes', *Nature*, 606(7914), pp. 535–541. Available at: https://doi.org/10.1038/s41586-022-04822-x.

Toda, N. *et al.* (2020) 'NLGenomeSweeper: A Tool for Genome-Wide NBS-LRR Resistance Gene Identification', *Genes*, 11(3), p. 333. Available at: https://doi.org/10.3390/genes11030333.

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.

Xie, J. *et al.* (2020) 'A rare single nucleotide variant in Pm5e confers powdery mildew resistance in common wheat', *New Phytologist*, 228(3), pp. 1011–1026. Available at: https://doi.org/10.1111/nph.16762.