Machine learning empowers precise discovery of disease-resistance genes in plants

Zhenya Liu[1†], Xu Wang[1, 3†], Shuo Cao[4, 1†], Tingyue Lei[1], Zhuyifu Chen[1, 3], Mengyan Zhang[1, 3], Zhongqi Liu[6, 1], Jiacui Li[7, 1], Jianzhong Lu[1], Wenqi Ma[1], Bingxiong Su[6, 1], Yanling Peng[1], Yanshuai Xu[1], Xiaodong Xu[2], Wei Zhang[2], Cong Tan[8], Chengjie Chen[2], Yiwen Wang[5*], and Yongfeng Zhou[2,1*]

1 *National Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Buxin Road, Dapeng New District, Shenzhen, China*

2 *National Key Laboratory of Tropical Crop Breeding, Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Xueyuan Road, Longhua District, Haikou, China*

3 *School of Agriculture and Food Science, University College Dublin, Belfield, Dublin 4, Ireland*

4 *Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA*

5 *Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC, Australia*

6 *College of Plant Protection, Hunan Agricultural University, Hunan, China*

7 *College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, China*

8 *BGI Research, Shenzhen, China*

* *Corresponding authors: zhouyongfeng@caas.cn; yiwen.wang@unimelb.edu.au*

† *These authors contributed equally to this work.*

Short title: Mining plant resistance genes with R-Predictor

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/General-Instructions) is Yongfeng Zhou.

ABSTRACT

Identifying plant disease-resistance genes is essential for understanding the plant immune system and accelerating the breeding of disease-resistant crops. There is a pressing need for a method capable of accurately identifying plant disease-resistance genes on a genome-wide scale. In this study, we propose Evolutionary Scale Modeling for LRR (ESM-LRR), a deep protein language model designed to accurately predict LRR domains which are substantially variable structures in disease-resistance proteins. ESM-LRR achieved its highest F1 score of 0.80 on a test set using 90% identity as the matching threshold. Building upon ESM-LRR, we developed R-Predictor, a plant disease-resistance gene predictor to simultaneously annotate 15 diverse domain topologies, covering characterized resistance genes across the whole genome. R-Predictor integrates four modules, each employing superior methods that outperform existing methods (achieving F1 scores of 0.89 for RLKs and 0.88 for NLRs), demonstrating its high accuracy and practicality in annotating plant disease-resistance genes. R-Predictor integrated with gene expression profiles to identify candidate R genes associated with grape gray mold and downy mildew, outperforming existing methods and detecting dozens of candidate R genes. Overall, this study presents a novel approach to advancing our understanding of plant immunity and facilitating crop breeding for disease resistance.

Keywords: R genes, plant immunity, disease resistance, deep learning

INTRODUCTION

Plant diseases are responsible for an average loss of 20-30% in global crop yields (Savary et al., 2019). The plant immune system is conceptualized into a “zig-zag” model, which consists of pattern-triggered immunity (PTI), effector-triggered susceptibility (ETS), and effector-triggered immunity (ETI) (Jones and Dangl, 2006). The recognition of pathogen-, damage-, microbe-, or herbivore-associated molecular patterns (PAMPs, DAMPs, MAMPs, HAMPs) by pattern-recognition receptors (PRRs) on the plant cell surface initiates defense responses against invading pathogens, a process known as PTI (Zipfel, 2014). PRRs include receptor-like kinases (RLKs) and receptor-like proteins (RLPs). To counteract or suppress PTI, pathogens have evolved to secrete specific effectors, resulting in effector-triggered susceptibility (ETS) (Ngou et al., 2022). However, plants have also evolved intracellular nucleotide-binding leucine-rich repeat

receptors (NLRs) that can directly or indirectly recognize these effectors, thereby activating ETI (Sun et al., 2024). In addition to classical immune receptors, tandem kinase proteins (TKPs) are a specialized class of resistance proteins distinguished by their dual-kinase domain architecture (Sung and Coaker, 2025; Powell et al., 2026). Consequently, the precise identification of disease-resistance genes involved in the production of TKPs, RLKs, RLPs and NLRs is essential for the enhancement of plant immunity.

TKPs are identified based on the presence of two or more fused kinase domains (KDs) and the maximum number of KDs in confirmed multi-kinase proteins reached only five (Reveguk et al., 2025). Plant RLKs comprise a signal peptide, an ectodomain (ECD), a transmembrane domain and a protein kinase domain, whereas RLPs are essentially RLKs that lack the protein kinase domain. The leucine-rich repeat (LRR)-RLKs and LRR-RLPs are the most abundant subfamilies of RLKs and RLPs, respectively, with the LRR domain serving as their ECD (Escocard de Azevedo Manhães et al., 2021). The LRR domain is a widespread structural motif consisting of two or more LRR units, each comprising 20-30 amino acids, characterized by a sequence pattern rich in leucine (Enkhbayar et al., 2004). The LRR unit typically repeats in tandem, forming a curved solenoid structure that facilitates protein-protein interactions. Similar to RLKs and RLPs, plant NLRs are modular proteins generally composed of three domains: an N-terminal domain, a central NB-ARC domain and a C-terminal LRR domain. Plant NLRs can be divided into TIR-type NLRs (TNLs) and CC-type NLRs (CNLs), distinguished by the presence of either a TOLL/interleukin-1 receptor domain or a coiled-coil (CC) domain at their N-terminal domain respectively (Wang and Chai, 2020). However, some plant disease-resistance genes lack typical domain combinations or topologies, exhibiting replaced, missing or additional domains, such as DAR5 (Karamat et al., 2022), yet they still play crucial roles in plant immunity. Therefore, characterizing the domain topology of plant disease-resistance genes is essential for identifying potentially overlooked resistance genes and mechanisms.

High-quality disease-resistance gene annotations have been established for the model species *Arabidopsis thaliana* (*A. thaliana*) through collaborative community efforts and meticulous manual curation. However, with the rapid advancement of sequencing technologies, numerous reference genomes have been assembled both within and across species, creating an urgent demand for automated annotation methods for plant disease-resistance genes in crops and

their wild relatives. To date, several methods have been developed to accomplish this goal, including NLRtracker (Kourelis et al., 2021), DeepLRR (Liu et al., 2022), NLRexpress (Martin et al., 2022), RLKdb (Yin et al., 2024) and Resistify (Smith et al., 2024). Nevertheless, these methods exhibit various limitations. For instance, RLKdb is restricted to identifying RLKs, while NLRtracker, NLRexpress and Resistify are confined to identifying NLRs. Although DeepLRR can identify RLKs, RLPs and NLRs simultaneously, it is constrained to typical domain structures. To overcome these shortcomings, different methods have been combined to achieve complementary results. However, no comprehensive benchmarking has been conducted to identify the optimal combination of methods for annotating the same type of receptors.

Furthermore, the LRR domain, a critical domain of RLKs, RLPs and NLRs exhibits substantial sequence variability, which poses challenges for accurate identification using existing methods. To address this, we propose a novel approach, Evolutionary Scale Modeling for LRR (ESM-LRR), which leverages representation learning from the large-scale protein language model ESM-1v (Meier et al., 2021) to extract features from input protein sequences, and feeds these features into a lightweight machine learning algorithm, RandomForestRegressor (RFR), to accurately characterize variable LRR domains. To systematically annotate plant disease-resistance genes with diverse domain architectures, we developed a novel computational framework, the Plant Disease-Resistance Gene Predictor (R-Predictor). This pipeline builds upon our newly proposed ESM-LRR method for accurately identifying highly variable LRR domains, and integrates it with established tools for domain feature detection, including SignalP 6.0 for signal peptides (Teufel et al., 2022), TMHMM-2.0 for transmembrane domains (Krogh et al., 2001), and Paircoil2 for coiled-coil domains (McDonnell et al., 2006). R-Predictor is specifically designed to facilitate the discovery of plant disease-resistance genes, providing a unified and scalable solution for the annotation of TKPs, RLKs, RLPs, and NLRs across complex domain topologies. Unlike conventional tools that focus on individual receptor types or canonical structures, R-Predictor enables the identification of a broader and more diverse repertoire of candidate resistance genes. It largely expands the pool of high-confidence candidates, offering valuable targets for future functional characterization. By doing so, R-Predictor serves as a practical and powerful tool for accelerating research in plant immunity and supporting breeding strategies such as genomic selection and gene editing.

RESULTS

Overview of ESM-LRR for the prediction of LRR domains

The ESM-LRR method was developed to accurately predict LRR domains in protein sequences based on 1,885 curated proteins from Swiss-Prot (**Figure S1**, **Tables S1, S2**). It leverages representation learning, a component of the popular deep protein language model ESM-1v (Meier et al., 2021). This learning functions as a sequence feature extractor, encoding amino acid sequences into high-dimensional feature vectors. These features are subsequently used with optimal machine learning algorithms for the prediction of LRR domains (**Figure 1A**). In the PCA plot, each point represents a sequence and is colored by its corresponding LRR score (**Figure 2A**). A clear gradient pattern emerged, suggesting that the features encoded by ESM-1v effectively captured the continuous spectrum of LRR-like properties. In addition, PCA reduced the dimensionality of the feature matrix, thereby improving the efficiency of downstream machine learning analysis.

To select the most effective prediction model, three widely used machine learning algorithms applicable to small to medium scale studies, including KneighborsRegressor (KNR), Support Vector Regression (SVR) and RandomForestRegressor (RFR) were benchmarked. Among all methods, RFR achieved the best performance in two out of three evaluation metrics, with a Root Mean Squared Error (RMSE) of 0.48, a Mean Absolute Error (Shiri et al.) of 0.39, and a Spearman Correlation Coefficient (Spearman) of 0.76. In comparison, KNR produced RMSE = 0.55, MAE = 0.43, and Spearman = 0.68, while SVR achieved RMSE = 0.51, MAE = 0.37 and Spearman = 0.74 (**Figure 2B**, **Tables S3, S4**). As a result, the final ESM-LRR method was constructed by combining ESM-1v representation learning for protein sequence encoding with RFR for LRR domain prediction.

ESM-LRR outperforms existing methods in annotating LRR domains

To evaluate the performance of ESM-LRR in comparison to other methods, including predict-phytoLRR (Chen, 2021), DeepLRR (Liu et al., 2022) and NLRexpress (Martin et al., 2022) in identifying LRR domains, we selected an additional set of 300 LRR proteins (comprising 2,812 LRR units) from the Swiss-Prot database (**Tables S1, 2**). Six different matching thresholds (50%, 60%, 70%, 80%, 90% 100% identity) were established to define successful identification as true

positives. F1 score and misclassification rate were used as metrics to evaluate the performance of these methods. ESM-LRR consistently achieved the highest F1 scores across all thresholds, with improvements ranging from 1.2% to 9.1% over the second-best method (**Figure 2C**, **Table S5**). Compared to other methods, ESM-LRR achieved the lowest error rate in five out of the six thresholds, except for the 100% threshold. At this threshold, NLRexpress missed 31 proteins with LRR domains, while ESM-LRR missed 35, corresponding to misclassification rates of 10.3% and 11.7 %, respectively (**Figure 2D**). One of the 300 LRR proteins, Q9SLI6 (putative receptor-like protein 8, RLP8), was used as a case protein to demonstrate the predictive capability of ESM-LRR. The predicted scores for each protein sequence are shown in **Figure 2E**. Given that the flanking sequences of true LRR units often display considerable scores and present false positive peaks on the score curve, we developed a filtering script based on a fixed score threshold (>1.2) and the distance relationship between true LRR units, which can be easily specified by users (**Figure S2**). Q9SLI6 contains a large LRR domain, consisting of 26 LRR units. Among them, 23 were correctly identified by ESM-LRR (**Figure 2F**). To gain a deeper understanding of domain topologies in plant disease-resistance genes, we collected 489 experimentally validated genes encompassing various domain topologies (**Table S6**). Among these genes, approximately 15.1% (74/489) encode TKPs/RLKs/RLPs, and the remaining 84.9% (415/489) encode NLRs (**Figure 1B, Figure S3**).

Construction of the R-Predictor for the annotation of plant disease-resistance genes

In addition to LRR domains, other critical domains, such as signal peptides, transmembrane domains, and coiled-coil domains, are also challenging to discern. To identify signal peptides, multiple tools were benchmarked using 9,796 curated proteins from Swiss-Prot, with SignalP 6.0 showing the best performance in terms of F1 score and misclassification rate across all thresholds (**Figures 3A, D, Table S7**). For transmembrane domain identification, four tools were benchmarked using 79,862 proteins from Swiss-Prot, with TMHMM-2.0 showing the highest F1 scores overall and the lowest misclassification rates at high identity thresholds (**Figures 3B**, **E**, **Table S8**). For coiled-coil domain identification, various tools were benchmarked using 3,783 curated proteins from Swiss-Prot, with Paircoil2 achieving the best overall performance in terms of F1 scores and misclassification rates across all thresholds (**Figures 3C**, **F, Table S9**).

The final framework, R-Predictor, designed for the *de novo* annotation of various plant disease-resistance genes integrate four modules for data pre-processing and the identification of different types of proteins. Each module incorporates customized filtering scripts and the best-performing methods identified through benchmarking (**Figure 3G**). R-Predictor provides detailed information on the domain topologies of plant disease-resistance genes along with the chromosomal locations of these genes when a corresponding genome annotation file is provided.

Evaluation of R-Predictor across multiple crop species

To assess the performance of R-Predictor, we benchmarked it against existing tools across four well-annotated plant genomes. While RLKdb (Yin et al., 2024) focuses exclusively on RLK identification and annotated 285, 202, 186 and 137 RLKs in each species respectively (**Table S10**), R-Predictor identified substantially more RLKs (340, 233, 187 and 242), and additionally detected other immune-related receptor families including TKPs (76, 57, 48 and 73) and RLPs (79, 92, 63 and 136) (**Figure 4A, Table S11**). Importantly, in Araport11 (Cheng et al., 2017), the most comprehensively curated reference genome, R-Predictor achieved an F1 score of 0.89 for RLK annotation, exceeding RLKdb by 4.8% (**Table S12**). It correctly identified 16 additional true LRR-RLKs, and 26 of the correctly identified RLKs were uniquely detected by R-Predictor. For RLPs, R-Predictor achieved a less favorable F1 score of 0.67 (**Figure 4C**).

For NLR annotation, R-Predictor performed competitively or better than three widely used tools, which are NLRexpress (Martin et al., 2022), NLRtracker (Kourelis et al., 2021) and Resistify (Smith et al., 2024). It annotated 168, 369, 212 and 745 NLRs across *A. thaliana*, rice, tomato, and grapevine, respectively (**Figure 4B, Table S13**), matching or exceeding the top-performing tools in each case (NLRexpress: 118, 245, 113 and 470, respectively (**Table S14**); NLRtracker: 167, 349, 212 and 581 (**Table S15**); Resistify: 167, 373, 212 and 746, respectively (**Table S16**)). In Araport11, R-Predictor achieved the highest recall (0.84) and F1 score (0.88) across all methods, outperforming Resistify, the second-best performer, by 2.4% in recall and 1% in F1 score (**Figure 4D, Table S17**).

To further validate the predictions of R-Predictor, we performed a phylogenetic analysis of the 168 NLRs it identified in Araport11 (**Figure S4**). Remarkably, the majority of NLRs (159/168) were successfully classified into evolutionary branches consistent with their predicted types. The remaining 7 CNLs and 2 TNs clustered into incorrect clades, likely due to high

similarity in the NB-ARC domains between these two topologies. These results further support the overall reliability of the 422 NLRs identified by R-Predictor. Consistent phylogenetic validation across rice, tomato, and grapevine (**Figures S5-7**, **Tables S18-20**) further confirmed R-Predictor's robust cross-species performance and its potential as a reliable tool for comprehensive plant immune receptor annotation.

R-Predictor identify potential R genes to grape gray mold and downy mildew

To further assess the reliability of R genes identified by R-Predictor, transcriptomic data from grape gray mold and downy mildew were analyzed (Toffolatti et al., 2020; Su et al., 2023). Two transcriptomic groups comprising 34 samples at six infection stages were used to detect differentially expressed disease-resistance genes between resistant and susceptible grape varieties (**Figures S8A, C**). At 0, 72, and 120 hours post-inoculation (hpi) with gray mold (*Botrytis cinerea*), R-Predictor identified 29, 45, and 37 upregulated genes encoding TKPs, RLKs, and RLPs, respectively. These numbers represent a 5.8-, 3.2-, and 3.7-fold increase compared to the 5, 14, and 10 upregulated RLKs detected by RLKdb at the corresponding stages (Figure 5A, Table S21). Furthermore, R-Predictor identified 18 genes (spanning TKPs, RLKs, and RLPs) that were consistently upregulated at 0, 72, and 120 hpi, in contrast to only 4 such RLKs identified by RLKdb. Finally, Gene Ontology (GO) enrichment analysis was performed to investigate the functional roles of the 50 uniquely upregulated genes identified exclusively by R-Predictor (**Figure S8B**). In the Biological Process category, these genes were significantly enriched in GO terms regulation of innate immune response, and jasmonic acid and ethylene-dependent systemic resistance (**Figure 5C**). At 1, 2, and 3 days post-inoculation (dpi) with downy mildew (*Plasmopara viticola*), R-Predictor and Resistify identified an identical set of 64 upregulated NLRs, whereas NLRexpress and NLRtracker identified 46 and 50 upregulated NLRs, respectively, all of which were subsets of the 64 NLRs (**Figure S8D**). By integrating ESM-LRR and Paircoil2, R-Predictor achieved more accurate identification of CNLs. At 1, 2, and 3 dpi, R-Predictor identified 29, 28, and 21 upregulated CNLs, respectively, compared with 24, 23, and 17 identified by Resistify, representing increases of approximately 1.21-, 1.22-, and 1.24-fold (**Figure 5B, Table S22**). R-Predictor identified 36 NLRs that were consistently upregulated across at 1, 2, and 3 dpi (**Figure 5D**), including 14 CNLs, 7 NLs, 5 CNs, 4 Ns, 4 TNLs, and 2 TNs.

DISCUSSION

Plants are constantly exposed to a wide range of pathogens, including viruses, bacteria, fungi, oomycetes and parasites, which can cause severe disease outbreaks when the immune defences of plants are compromised. Understanding the molecular basis and evolutionary strategy of plant immune responses is therefore crucial for developing strategies to strengthen disease resistance (Guo et al., 2025; Teasdale et al., 2025). Plant immunity is broadly classified into two types: PTI, which is primarily mediated by TKPs, RLKs and RLPs, and ETI, which is mainly governed by NLRs. Recent advances in sequencing technologies have enabled exploration of wild germplasms and have revealed novel resistance traits that extend beyond what is captured in reference genomes. This growing volume of genomic data highlights the urgent need for robust and scalable tools to annotate plant disease-resistance genes across diverse species.

The LRR domain is a critical structural motif present in the majority of RLKs, RLPs and NLRs, which exhibits high sequence variability thus influences accurate annotation. LRR domains have diverse functions, involving both plant immunity and a wide range of biological processes. Current methods for annotating LRR domains include: predict-phytoLRR based on the position-specific scoring matrix (PSSM) algorithm, DeepLRR based on convolutional neural networks (CNN), and NLRexpress based on a combination of machine learning predictors. In contrast, our proposed method, ESM-LRR, leverages representation learning from the deep learning model ESM-1v to extract high-dimensional features from protein sequences. These features are then used in a machine learning model RandomForestRegressor to learn the association between sequence features and LRR scores assigned according to the distance and overlap with the true LRR units, enabling accurate prediction of LRR domains in sequences. In this context, RandomForestRegressor offers a fast, interpretable, and resource-efficient solution that performs robustly on moderate-sized datasets, especially when end-to-end training with transformer architectures or graph neural networks is unnecessary or impractical. Compared to predict-phytoLRR and NLRexpress, which rely on pre-defined sequence patterns such as “LxxLxLxxN” and “LxxLxL” for the recognition of LRR domains, both DeepLRR and ESM-LRR are more data driven. They learn discriminative sequence features directly from annotated LRR units, enabling model-based prediction rather than relying solely on predefined motifs.

Among the two, DeepLRR formulates the task as a binary classification problem, predicting whether a sequence is an LRR unit or not, whereas ESM-LRR employs regression models to predict continuous LRR scores which reflects the likelihood and proximity of sequences to true LRR units. These continuous scores implicitly capture more characteristic information including spatial distance and overlap with the true units, which may contribute to ESM-LRR's superior performance compared to the other method. However, due to computational limitations, the input of ESM-LRR is currently restricted to 23 amino acid residues in length, which aligns with the most common length of canonical LRR units (**Figure S9**). While this may limit the ability to explicitly detect non-canonical or longer LRR repeats, the use of sliding windows combined with continuous scoring allows ESM-LRR to effectively identify LRR-like regions, even when they extend beyond 23 amino acid residues. This strategy helps mitigate the limitations imposed by fixed input length.

Compared to existing methods for annotating plant disease-resistance genes, our proposed R-Predictor has two major improvements: 1) it simultaneously annotates 15 distinct types of domain topologies for a wide range of resistance genes on a genome-wide scale. R-Predictor can also identify the functional genes lacking typical domain topologies, such as DAR5 (Karamat et al., 2022), providing an opportunity to recover previously overlooked resistance resources. 2) R-Predictor integrates the most efficient methods for identifying specific functional domains, such as SignalP 6.0 for signal peptides, TMHMM-2.0 for transmembrane domains, Paircoil2 for coiled-coil domains, and ESM-LRR for LRR domains, thereby enhancing its ability to accurately classify resistance genes. For instance, R-Predictor demonstrates higher accuracy in identifying signal peptides, transmembrane domains and LRR domains compared to RLKdb, as shown particularly in Araport11, and potentially outperforms other methods designed for general annotation. Based on ESM-LRR and Paircoil2, R-Predictor can annotate more CNLs than other methods focused on NLRs identification. The primary goal of our tool is to provide a broad set of potential disease-resistance gene candidates for downstream analytical and functional studies. Therefore, experimental validation of predictions across multiple species falls outside the scope of this study. To assess the reliability of R-Predictor, we focused on the well-curated model species *A.thaliana*. In this context, phylogenetic consistency and domain-level evidence offer strong computational support for the accuracy and robustness of our predictions.

We developed R-Predictor, a framework that integrates the ESM-LRR method with established domain annotation tools to enable genome-wide annotation of resistance genes across 15 distinct domain topologies. These topologies were derived by deconstructing the proteins encoded by 489 experimentally validated plant disease-resistance genes, extensive including recently characterized resistance genes. R-Predictor not only outperformed existing methods in *A. thaliana* but also generalized well to crop genomes including rice, tomato and grapevine. Transcriptomic analyses of grapevine gray mold and downy mildew provided strong support for the reliability of R-Predictor predictions. In grape gray mold, resistance is predominantly mediated by PTI responses involving RLKs and RLPs (Rahman et al., 2024). R-Predictor identified a substantially higher number of DEGs that were consistently upregulated across all infection stages compared with RLKdb (18 versus 4), representing a 4.5-fold increase. Moreover, R-Predictor uniquely detected six additional classes of putative resistance genes that could not be identified by RLKdb, including LysM-RLK, TKP, S-TM-PK, LRR-RLP, LysM-RLP and S-LysM, thereby expanding the repertoire of PTI-associated resistance genes. In contrast, resistance to grape downy mildew is mainly conferred by ETI responses mediated by NLRs (Paineau et al., 2024). Although R-Predictor and Resistify identified the same number of upregulated DEGs with identical gene IDs, R-Predictor exhibited superior performance in NLR classification. Specifically, it identified a higher number of CNLs while assigning fewer genes to the NL category, resulting in a more refined and biologically informative annotation of NLRs.

Like most methods for annotating plant disease-resistance genes, R-Predictor still relies on traditional tools such as Pfam 36.0 (Mistry et al., 2021) to identify conserved protein kinase and NB-ARC domains. While these tools are well-established and widely used, they may limit the discovery of novel or non-canonical resistance genes due to their dependence on existing curated domain profiles. Emerging deep learning methods have shown improved sensitivity in domain annotation and offer promising avenues for expanding and refining detection capabilities. Future improvements may include hybrid frameworks, where tools like Pfam 36.0 are used for rapid initial screening, followed by deep learning models to refine domain boundaries or recover incomplete regions. Additionally, developing integrated, community-driven open-source platforms that dynamically update domain annotations based on new literature or experimental data could further enhance the annotation of plant disease-resistance genes.

MATERIALS AND METHODS

Datasets

We screened 1,885 curated proteins (**Table S1**, **S2**) containing LRR domains from the Swiss-Prot database (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/) to train, verify and test ESM-LRR. 1,585 proteins were used for the development of ESM-LRR, and 300 proteins were applied for comparison of the performance of ESM-LRR and other methods for predicting LRR domains. We extracted 23aa sequence fragments along both sides of the LRR unit of 1,585 LRR proteins, and a normal distribution density function (μ=0, σ=0.2) was employed to assign scores for these fragments, ultimately generating a set consisting of 128,855 fragments, of which 13,905 fragments are true LRR units (**Figure S1**). Referring to the division ratio of 80%:20%, 103,084 fragments are responsible for 5-fold cross-validation to obtain the optimal hyperparameter combination of ESM-LRR, and 25,711 fragments are used as a test set. In addition to this, we also generated a signal peptide dataset, a transmembrane domain dataset and a coiled-coil dataset, which supported benchmarking of different modules in the R-Predictor. Proteins in these datasets are also from the Swiss-Prot database. The signal peptide dataset consists of 9,796 proteins, including 9,796 signal peptides, the transmembrane domain dataset consists of 79,862 proteins, including 381,188 transmembrane domains, and the coiled-coil dataset consists of 3,783 proteins, of which 6,438 coiled-coils.

Training ESM-LRR

The training process of ESM-LRR was divided into the following steps, (I) sequence embeddings were extracted from 103,084 training fragments based on the deep protein language model ESM-1v (Meier et al., 2021) (version: esm1v_t33_650M_UR90S_1); (II) loaded sequence embeddings and converted 1280-dimensional sequence embeddings to 60 dimensions by PCA (Wold et al., 1987; Abdi and Williams, 2010); (III) 5-fold cross-validation and grid search to explore the optimal hyperparameter combination corresponding to KNeighborsRegressor (Alsayadi et al., 2022), SVR (Fan et al., 2016) and RandomForestRegressor (Rigatti, 2017); (IV) an independent test set was used to determine the final downstream model of ESM-LRR. Random.seed() was used to achieve repeatability of model training and StratifiedKFold() was used to ensure the category distribution of each slice is consistent with the overall data. The development of ESM-LRR was implemented on torch (version: 2.3.0) and scikit-learn (version:

1.5) frameworks on a cloud computer (GPU: GeForce RTX 3090 24GB and CPU: 12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz). In addition, we successfully ran the R-Predictor docker image on a personal computer (GPU: RTX 1660s (8GB), CPU: Gen Intel(R) Core (TM) i7-12700F @ 2.10 GHz). In short, R-Predictor does not require expensive computing resources.

Identifying the LRR domain by ESM-LRR

If ESM-LRR only loads the downstream model, it can only assign scores to sequence fragments to recognize whether they are potential LRR units, but cannot identify the complete LRR domain. To improve the practicality of ESM-LRR, we incorporated a filtering script based on a fixed threshold and the distance relationship between LRR units. This script will filter the result given by the downstream model through three steps in sequence. (I) Remove sequence fragments with scores less than a given threshold, for example, threshold is 1.2 in this study. Regarding the fixed threshold set in the filtering script, we found in cross-validation that a threshold of 1.2 provides the best trade-off between precision and recall. (II) Pick a sequence fragment with a higher score if the distance between the two fragments is less than 20 amino acid residues. (III) If the distance between a fragment and its adjacent fragments is less than or equal to 30 amino acid residues, the fragment is considered as a potential LRR unit.

Evaluation measures

We adopt 10 measures to evaluate the performance in this study, 4 for training ESM-LRR, 2 for benchmarking different modules of R-Predictor and 4 for prediction of disease-resistance genes. To facilitate the description, hereafter, y_i represents the actual value, $\hat{y}_i$ represents the predicted value, $\bar{y}$ represents the mean of actual values, d_i represents the difference between the ranking of i-th pair of actual value and predicted value, and n represents total number of samples.

$$R - Squared = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$Spearman = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

F1 score and misclassification rate are responsible for benchmarking different modules in R-Predictor. They are defined as follows, where TP means the number of actual values correctly identified, FP means the number of non-actual values incorrectly identified as actual values, TN means the number of non-actual values correctly identified and FN means the number of actual values incorrectly identified as non-actual values.

$$F1\ score = \frac{2TP}{2TP + FP + FN}$$

$$misclassification\ rate = \frac{FN}{TP + FP + TN + FN}$$

As predicting plant disease-resistance genes is a multi-classification task, $Accuracy$, $Precision_{Weighted}$, $Recall_{Weighted}$ and $F1\ score_{Weighted}$ are used to evaluate the performance of various methods. They are defined as follows, where t means the number of different labels and w_i means the proportion of the number of samples corresponding to the i-th label to the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision_{Weighted} = \sum_{i=1}^{t} w_i Precision_i = \sum_{i=1}^{t} w_i (\frac{TP}{TP + FP})_i$$

$$Recall_{Weighted} = \sum_{i=1}^{t} w_i Recall_i = \sum_{i=1}^{t} w_i (\frac{TP}{TP + FN})_i$$

$$F1\ score_{Weighted} = \sum_{i=1}^{t} w_i F1\ score_i = \sum_{i=1}^{t} w_i (\frac{2TP}{2TP + FP + FN})_i$$

Configuring the various programs for benchmarking

Predict-phytoLRR (Chen, 2021) (https://github.com/phytolrr/predict-phytolrr), DeepLRR (Liu et al., 2022) (version: 1.01 https://github.com/zhenyaliu77/DeepLRR) and NLRexpress (Martin et al., 2022) (https://github.com/eliza-m/NLRexpress) were installed to identify LRR domain from 300 LRR proteins. LipoP (Juncker et al., 2003) (https://services.healthtech.dtu.dk/services/LipoP-1.0), DeepSig (Savojardo et al., 2018) (https://deepsig.biocomp.unibo.it), Phobius (Käll et al., 2004) (https://phobius.sbc.su.se) and SignalP 6.0 (Teufel et al., 2022) (https://github.com/fteufel/signalp-6.0) were implemented to participate in the benchmarking of signal peptide based on 9,796 protein sequences. SCAMPI2 (Peters et al., 2016) (https://github.com/ElofssonLab/scampi2), HMMTOP (Tusnady and Simon, 2001) (http://www.enzim.hu/hmmtop), Phobius (https://phobius.sbc.su.se) and TMHMM-2.0 (Käll et al., 2004) (https://services.healthtech.dtu.dk/services/TMHMM-2.0) were used to identify the transmembrane domain from 79,862 proteins for benchmarking. In the benchmarking of coiled-coil, DeepCoil (Ludwiczak et al., 2019) (version: 2.0 https://github.com/labstructbioinf/DeepCoil), COILS (Lupas et al., 1991) (version: 2.2 https://mybiosoftware.com/coils-2-2-prediction-coiled-coil-regions-proteins.html), CoCoNat (Madeo et al., 2023) (https://github.com/BolognaBiocomp/coconat) and Paircoil2 (McDonnell et al., 2006) (https://cb.csail.mit.edu/cb/paircoil2) were employed to process 3,783 proteins. To present the performance of each program fairly, only default parameters are allowed in the benchmarking.

Construction of the R-Predictor

From the benchmarking results, we determined a series of methods that presented high precision, recall and F1 score. Integrating these methods, we have developed a pipeline called R-Predictor, which combines the best-performing methods and subsequent filtering scripts for de novo identification of different types disease-resistance gene. R-Predictor is deployed on the Linux system in a modular way, each method has its own virtual environment to facilitate the user to update and maintain the pipeline. The operation of R-Predictor can be summarized as the following steps. (I) Loading a protein file and check whether it conforms to the fasta format. (II) Running pfam_scan.pl to search the Pfam 36.0 database (Mistry et al., 2021) to identify protein sequences that have either KDs (PF00069, PF07714) or NB-ARC domain (PF00931) or neither.

A sequence containing between two and five KDs will be classified TKPs, and TMHMM-2.0 (Käll et al., 2004) was executed to identify protein sequences that contain both KDs and TM domain. (III) SignalP 6.0 (Teufel et al., 2022) was used to identify signal peptides in RLKs and RLPs, and TMHMM-2.0 was used to examine TM domains in RLPs. (IV) Implementing ESM-LRR to identify potential LRR-RLKs and LRR-RLPs, as well as to obtain protein sequences without the LRR domains for next step. (V) Running pfam_scan.pl again to identify LysM domain (PF01476), filtering out non-target domains and determining potential LysM-RLKs, LysM-RLPs and protein sequences that present the two domain topologies of S-TM-PK and S-LysM. (VI) The Pfam 36.0 and PROSITE databases (Sigrist et al., 2010; Mistry et al., 2021) were search to identify protein sequences containing TIR domain (PF01582, PF13676, PF18567, PS50104) or RPW8 domain (PF05659, PS51153) based on pfam_scan.pl and ps_scan.pl, respectively, and TNLs, TNs, RNLs and RNs were returned. Next, Paircoil2 (McDonnell et al., 2006) was applied to protein sequences that contain only NB-ARC domain or both NB-ARC domain and LRR domain to detect the presence of CC, and CNLs, CNs, NLs and Ns were returned. (VII) If the genome annotation file is provided, an embedded script will automatically return the gene ID and its chromosome position corresponding to the proteins.

Execution of current methods for identifying plant disease-resistance genes

RLKdb (https://biotec.njau.edu.cn/rlkdb) is a database containing ~220,000 RLKs from 300 plant genomes, from which we screened LRR-RLKs belonging to Araport11 (Cheng et al., 2017), IRGSP-1.0 (Kawahara et al., 2013), ITAG4.0 (Hosmani et al., 2019) and PN40024_T2T (Shi et al., 2023). NLRexpress (https://github.com/eliza-m/NLRexpress), NLRtracker (version: v1.0.3 https://github.com/slt666666/NLRtracker) and Resistify (https://github.com/SwiftSeal/resistify) were deployed on the Linux system (Ubuntu 20.04.4) to quickly and large-scale identify NLRs in Araport11, IRGSP-1.0, ITAG4.0 and PN40024_T2T. In-house developed Python scripts were used to collate and evaluate the results produced by these methods.

Phylogenetic tree

The sequences of NB-ARC domain in NLRs were delivery to MEGA (Kumar et al., 1994) (version: 11.0.13 https://www.megasoftware.net) for multiple sequence alignment （MUSCLE

) and phylogenetic tree construction (maximum likelihood), and the default parameters were used. iTOL (Letunic and Bork, 2021) (version: v6 https://itol.embl.de) was employed to add annotations to the phylogenetic tree, including modifying color and labeling.

Transcriptome analysis

Two independent transcriptome datasets PRJNA788159 and PRJEB24540 (Toffolatti et al., 2020; Su et al., 2023) were downloaded from the NCBI database. Grape gray mold transcriptomic dataset included two grape cultivars, Beta and Red Globe, sampled at three time points (0h, 72h and 120h post-infection), with three biological replicates at each time point (**Figure S8A**). Transcriptomic data of grape downy mildew infection were generated from two grape cultivars, Pinot noir and Mgaloblishvili, sampled at 1 dai, 2 dai and 3 dai. Two biological replicates were excluded during data quality control (**Figure S8C**). Quality control and adapter trimming of raw fastq data were carried out using Fastp with default settings (Chen et al., 2018). The processed reads were mapped to the PN_T2T reference genome using STAR (Dobin et al., 2013) (v1.5.2). PCA and differential expression analyses were conducted using DESeq2 (Love et al., 2014) (v1.36.0), enabling pairwise comparisons between resistant and susceptible grape samples at each time point. Differentially expressed genes (DEGs) were defined by $|\log_2 \text{FoldChange}| \geq 2$ and an adjusted P-value < 0.05.

Data availability

The raw data in this study have been deposited into Github:

https://github.com/zhouyflab/R-Predictor

Code availability

All scripts performed in this study are available on Github:

https://github.com/zhouyflab/R-Predictor

ACKNOWLEDGEMENTS

We thank members of the Zhou lab at AGIS for discussion and comments on the project and Prof. Detlef Weigel at the Max Planck Institute for Biology Tübingen for his valuable suggestions regarding Arabidopsis NLR proteins.

FUNDING

This work was supported by the Hainan Provincial Natural Science Foundation of China (No.326JCQN0984, No.326MS0293, No.326MS0294, No.326QN0753) and the Ningxia Key Research and Development Program (No.2025BBF02019).

ATHHOR CONTRIBUTIONS

Y.Z. and Y.W. designed the project. Y.Z. and Y.W. supervised the project. Z.L. designed and trained the ESM-LRR. Z.L., X.W., S.C. and Z.L. performed the benchmarks. Z.L. designed and constructed the R-Predictor. Z.L. performed the bioinformatic analyses. Z.L., X.W., T.L., Z.C., M.Z., J.L., W.M. and B.S. prepared the figures and Supplementary data. Z.L. wrote the draft. Y.Z., Y.W., C.C., C.T., W.Z., X.X., Y.X. and Y.P. revised the manuscript. All authors contributed to manuscript preparation and read, commented on, and approved the manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1 | Construction of ESM-LRR and comparison with other methods for annotating LRR domains. (A) ESM-LRR assigns continuous scores to sequence fragments, derived from a Gaussian probability distribution ($\mu=0$, $\sigma=0.2$), to quantify the likelihood that a given sequence represents a true LRR unit. The pretrained ESM-1v as a fixed feature extractor with default parameters to generate sequence representations and PCA was used to reduce dimension of sequence representations. Finally, after cross-validation to select the optimal parameter combination and evaluation on the test set, we chose RFR as the downstream machine learning model for ESM-LRR. **(B)** Domain topologies of 489 experimentally verified disease-resistance gene products.

Figure 2 | Comparison of ESM-LRR with other methods for annotating LRR domains. (**A**) Sequence feature matrix in the training set was visualized using PCA for dimensionality reduction. (**B**) Built-in machine learning models in ESM-LRR were evaluated on an independent test. (**C and D**) Performance comparison of ESM-LRR and other methods on 300 LRR proteins based on F1 score and misclassification rate. (**E**) Q9SLI6 protein was divided into overlapping sequences using a sliding window of 23 amino acids, and each sequence was predicted with an LRR score based on ESM-LRR. The red dashed line denotes a pre-set scoring threshold; fragments scoring above this line were considered potential LRR units. (**F**) The 3D structure of Q9SLI6, with the LRR units that were not successfully annotated by ESM-LRR highlighted in red, green and yellow.

Figure 3 | Benchmark various methods for different domains and construction of R-Predictor. The performance of different methods in annotating signal peptides (**A**, **D**), transmembrane domains (**B**, **E**), and coiled-coil domains (**C**, **F**) was evaluated using F1 score and misclassification rate. (**G**) R-Predictor consists of four modules. In the data module, protein sequences are classified and transmitted to three downstream modules. The RLK module, NLR module and RLP module are responsible for outputting 4, 8 and 3 types of plant disease-resistance genes, respectively.

Figure 4 | Comparison of R-Predictor with other methods for annotating plant disease-resistance genes in various plant species. (**A**) TKPs, RLKs (LRR-RLK, LysM-RLK and S-TM-PK) and RLPs (LRR-RLP, LysM-RLP and S-LysM) with different domain topologies annotated by R-Predictor and RLKdb. (**B**) NLRs (CNL, CN, TNL, TN, RNL, RN, NL and N) with different domain topologies annotated by R-Predictor and other methods. (**C**) Evaluation of the performance of R-Predictor and RLKdb to annotate RLKs and RLPs from Araprot11 based on precision, recall, and F1 score. (**D**) Evaluation of the performance of R-Predictor and other methods to annotate NLRs from Araprot11 based on precision, recall, and F1 score.

Figure 5 | Identification of upregulated disease-resistance genes associated with grape gray mold and downy mildew by R-Predictor. (**A**) At 0, 72, and 120 hpi with gray mold (*Botrytis cinerea*), upregulated TKPs, RLKs (LRR-RLK, LysM-RLK and S-TM-PK) and RLPs (LRR-RLP, LysM-RLP and S-LysM) identified by RLKdb and R-Predictor in transcriptomic analyses of gray mold infection in the grape cultivars Beta and Red Globe. Bars with dashed and solid outlines represent RLKdb and R-Predictor, respectively. (**B**) At 1, 2, and 3 dpi with downy mildew (*Plasmopara viticola*) upregulated NLRs (CNL, CN, TNL, TN, RNL, RN, NL and N) identified by NLRexpress, NLRtracker, Resistify and R-Predictor in transcriptomic analyses of downy mildew infection in the grape cultivars Pinot noir and Mgaloblishvili. Bars with dashed and solid outlines represent the three tools (from left to right: NLRexpress, NLRtracker, Resistify) and R-Predictor, respectively. (**C**) GO enrichment analysis of 50 uniquely upregulated genes encoding TKPs, RLKs and RLPs identified by R-Predictor (**D**) Heatmap of $\log_2$(FPKM+1) for 36 NLRs that were consistently upregulated across all infection stages.

REFERENCES

Abdi H, Williams LJ (2010) Principal component analysis. Wiley interdisciplinary reviews: computational statistics **2:** 433-459

Alsayadi HA, Abdelhamid AA, El-Kenawy E-SM, Ibrahim A, Eid MM (2022) Ensemble of Machine Learning Fusion Models for Breast Cancer Detection Based on the Regression Model. Fusion: Practice & Applications **9**

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34:** i884-i890

Chen T (2021) Identification and characterization of the LRR repeats in plant LRR-RLKs. BMC molecular and cell biology **22:** 1-16

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant Journal **89:** 789-804

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29:** 15-21

Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, Matsushima N (2004) Structural principles of leucine-rich repeat (LRR) proteins. Proteins: Structure, Function, and Bioinformatics **54:** 394-403

Escocard de Azevedo Manhães AM, Ortiz-Morea FA, He P, Shan L (2021) Plant plasma membrane-resident receptors: Surveillance for infections and coordination for growth and development. Journal of integrative plant biology **63:** 79-101

Fan G-F, Peng L-L, Hong W-C, Sun F (2016) Electric load forecasting by the SVR model with differential empirical mode decomposition and auto regression. Neurocomputing **173:** 958-970

Guo B-C, Zhang Y-R, Liu Z-G, Li X-C, Yu Z, Ping B-Y, Sun Y-Q, van den Burg H, Ma F-W, Zhao T (2025) Deciphering plant NLR genomic evolution: synteny-informed classification unveils insights into TNL gene loss. Molecular Biology and Evolution **42:** msaf015

Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, van Haarst J, Cordewener J, Sanchez-Perez G, Peters S (2019) An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. biorxiv**:** 767764

Jones JD, Dangl JL (2006) The plant immune system. nature **444:** 323-329

Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Science **12:** 1652-1662

Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. Journal of molecular biology **338:** 1027-1036

Karamat U, Yang R, Ren Y, Lu Y, Li N, Zhao J (2022) Comprehensive in silico characterization and expression pro-filing of DA1/DAR family genes in Brassica rapa. Genes **13:** 1577

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice **6:** 1-10

Kourelis J, Sakai T, Adachi H, Kamoun S (2021) RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family. PLoS Biology **19:** e3001124

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology **305:** 567-580

Kumar S, Tamura K, Nei M (1994) MEGA: molecular evolutionary genetics analysis software for microcomputers. Bioinformatics **10:** 189-191

Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and

annotation. Nucleic acids research **49:** W293-W296
Liu Z, Ren Z, Yan L, Li F (2022) DeepLRR: an online webserver for leucine-Rich-Repeat containing protein characterization based on deep learning. Plants **11:** 136
Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology **15:** 1-21
Ludwiczak J, Winski A, Szczepaniak K, Alva V, Dunin-Horkawicz S (2019) DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences. Bioinformatics **35:** 2790-2795
Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science **252:** 1162-1164
Madeo G, Savojardo C, Manfredi M, Martelli PL, Casadio R (2023) CoCoNat: a novel method based on deep learning for coiled-coil prediction. Bioinformatics **39:** btad495
Martin EC, Spiridon L, Goverse A, Petrescu A-J (2022) NLRexpress—A bundle of machine learning motif predictors—Reveals motif stability underlying plant Nod-like receptors diversity. Frontiers in Plant Science **13:** 975888
McDonnell AV, Jiang T, Keating AE, Berger B (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics **22:** 356-358
Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A (2021) Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in neural information processing systems **34:** 29287-29303
Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, Tosatto SC, Paladin L, Raj S, Richardson LJ (2021) Pfam: The protein families database in 2021. Nucleic acids research **49:** D412-D419
Ngou BPM, Ding P, Jones JD (2022) Thirty years of resistance: Zig-zag through the plant immune system. The Plant Cell **34:** 1447-1478
Paineau M, Minio A, Mestre P, Fabre F, Mazet ID, Couture C, Legeai F, Dumartinet T, Cantu D, Delmotte F (2024) Multiple deletions of candidate effector genes lead to the breakdown of partial grapevine resistance to downy mildew. New Phytologist **243:** 1490-1505
Peters C, Tsirigos KD, Shu N, Elofsson A (2016) Improved topology prediction using the terminal hydrophobic helices rule. Bioinformatics **32:** 1158-1162
Powell OR, Guzmán-Vega FJ, Yu DS, Wang YL, Lu P, Arold ST, Liu Z, Banfield MJ, Wulff BB, Chen R (2026) The emerging role of kinase fusion proteins in cereal immunity. Nature Genetics**:** 1-9
Rahman MU, Liu X, Wang X, Fan B (2024) Grapevine gray mold disease: Infection, defense and management. Horticulture Research **11:** uhae182
Reveguk T, Fatiukha A, Potapenko E, Reveguk I, Sela H, Klymiuk V, Li Y, Pozniak C, Wicker T, Coaker G (2025) Tandem kinase proteins across the plant kingdom. Nature Genetics **57:** 254-262
Rigatti SJ (2017) Random forest. Journal of Insurance Medicine **47:** 31-39
Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A (2019) The global burden of pathogens and pests on major food crops. Nature ecology & evolution **3:** 430-439
Savojardo C, Martelli PL, Fariselli P, Casadio R (2018) DeepSig: deep learning improves signal peptide detection in proteins. Bioinformatics **34:** 1690-1696
Shi X, Cao S, Wang X, Huang S, Wang Y, Liu Z, Liu W, Leng X, Peng Y, Wang N (2023) The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding. Horticulture Research **10:** uhad061
Shiri Y, Solouki M, Ebrahimie E, Emamjomeh A, Zahiri J (2020) Gibberellin causes wide transcriptional modifications in the early stage of grape cluster development. Genomics **112:** 820-830
Sigrist CJ, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic acids research **38:** D161-D166

Smith M, Jones JT, Hein I (2024) Resistify-A rapid and accurate annotation tool to identify NLRs and study their genomic organisation. bioRxiv**:** 2024.2002. 2014.580321

Su K, Zhao W, Lin H, Jiang C, Zhao Y, Guo Y (2023) Candidate gene discovery of Botrytis cinerea resistance in grapevine based on QTL mapping and RNA-seq. Frontiers in Plant Science **14:** 1127206

Sun Y, Liu F, Zeng M, Zhang X, Cui Y, Chen Z, Wang L, Xu Y, Wu J, Guo S (2024) The ETI-dependent receptor-like kinase 1 positively regulates effector-triggered immunity by stabilizing NLR-required for cell death 4 in Nicotiana benthamiana. New Phytologist **242:** 576-591

Sung Y-C, Coaker G (2025) A plant dual-kinase protein traps a pathogen effector to trigger immunity. Nature Genetics

Teasdale LC, Murray KD, Collenberg M, Contreras-Garrido A, Schlegel T, van Ess L, Jüttner J, Lanz C, Deusch O, Fitz J (2025) Pangenomic context reveals the extent of intraspecific plant NLR evolution. Cell Host & Microbe **33:** 1291-1305. e1299

Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. Nature biotechnology **40:** 1023-1025

Toffolatti SL, De Lorenzis G, Brilli M, Moser M, Shariati V, Tavakol E, Maddalena G, Passera A, Casati P, Pindo M (2020) Novel aspects on the interaction between grapevine and Plasmopara viticola: dual-RNA-Seq analysis highlights gene expression dynamics in the pathogen and the plant during the battle for infection. Genes **11:** 261

Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics **17:** 849-850

Wang J, Chai J (2020) Structural insights into the plant immune receptors PRRs and NLRs. Plant physiology **182:** 1566-1581

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometrics and intelligent laboratory systems **2:** 37-52

Yin Z, Liu J, Dou D (2024) RLKdb: A comprehensively curated database of plant receptor-like kinase families. Molecular Plant **17:** 513-515

Zipfel C (2014) Plant pattern-recognition receptors. Trends in immunology **35:** 345-351

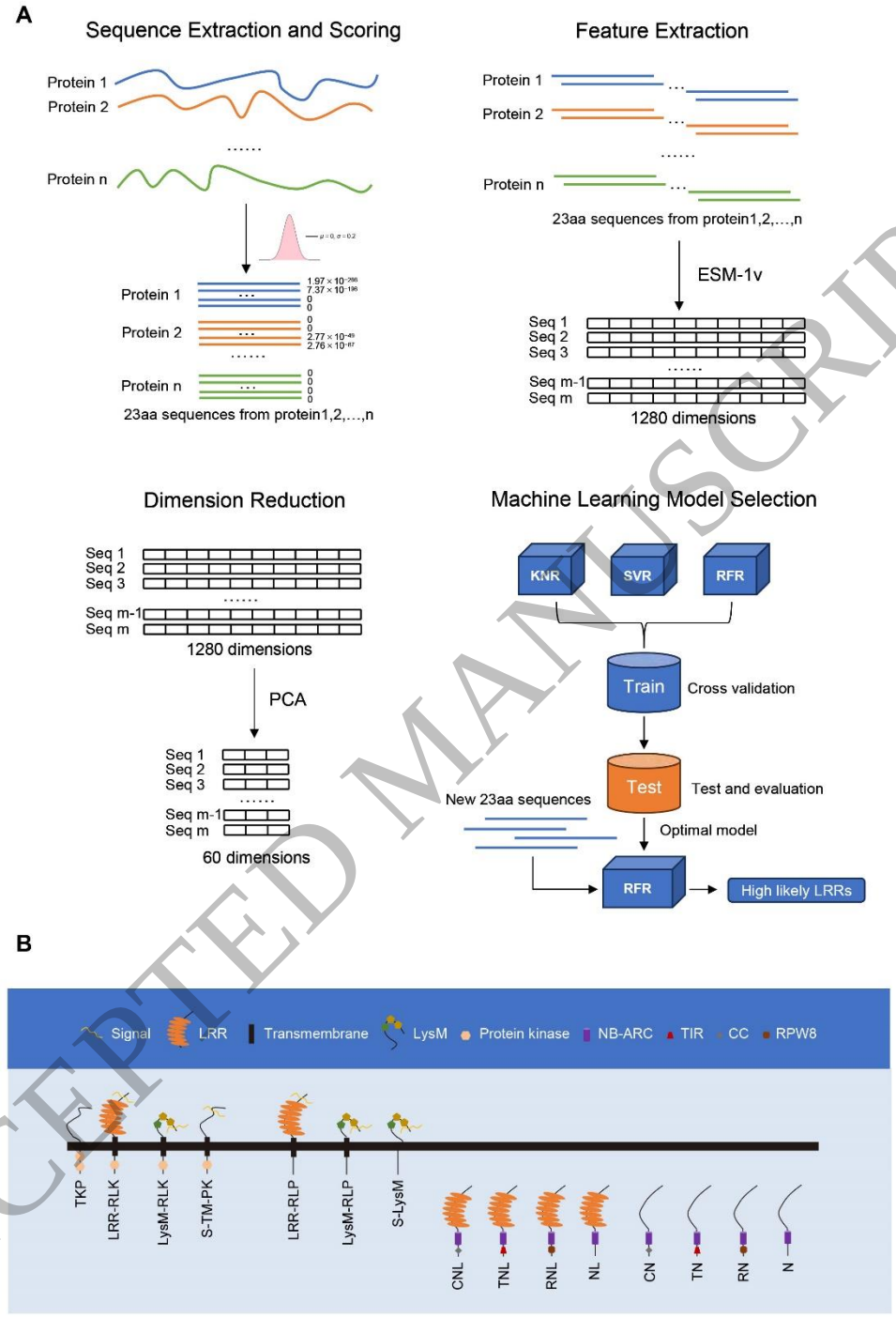

Figure 1
163x229 mm (x DPI)

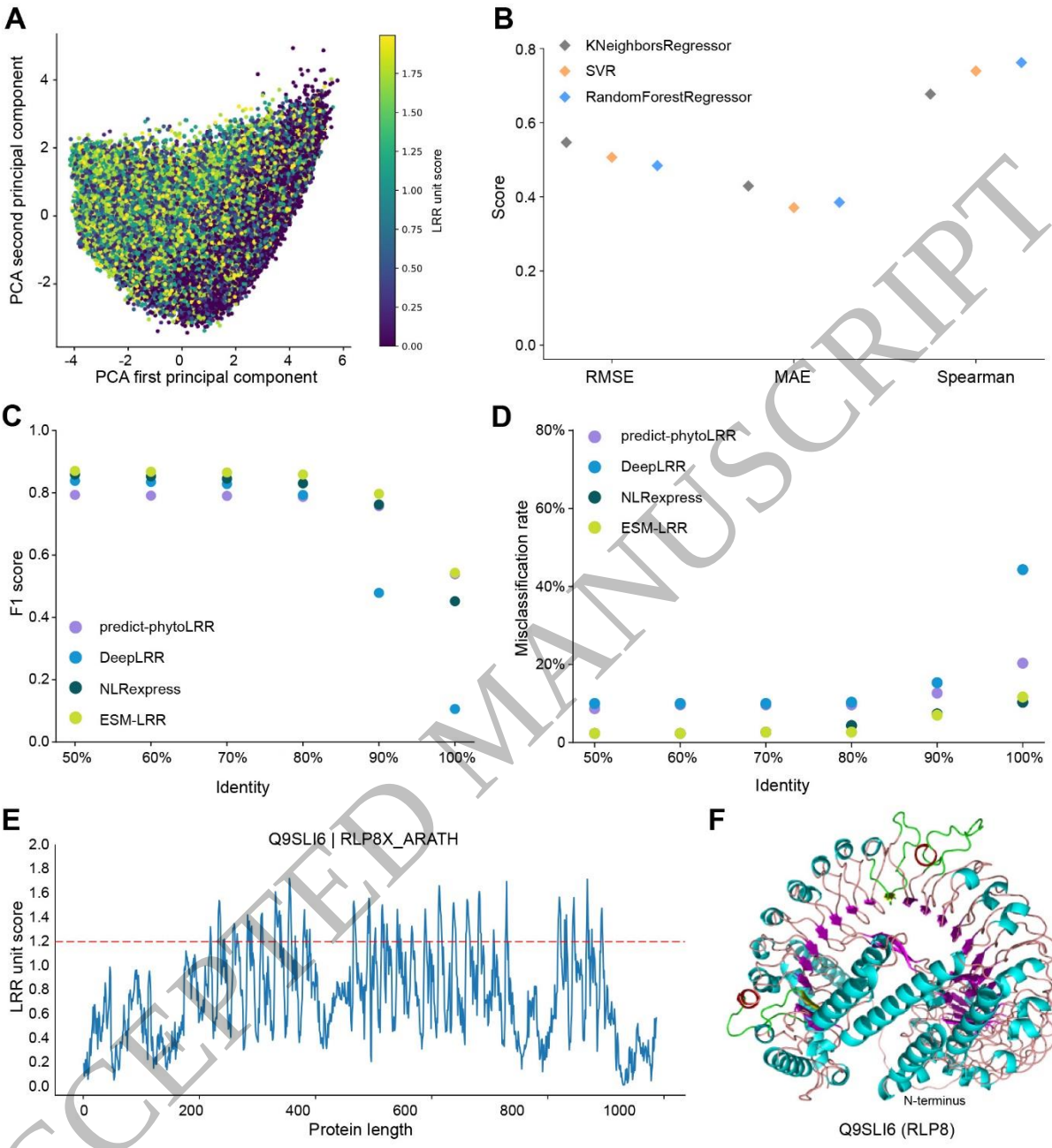

Figure 2
165x173 mm (x DPI)

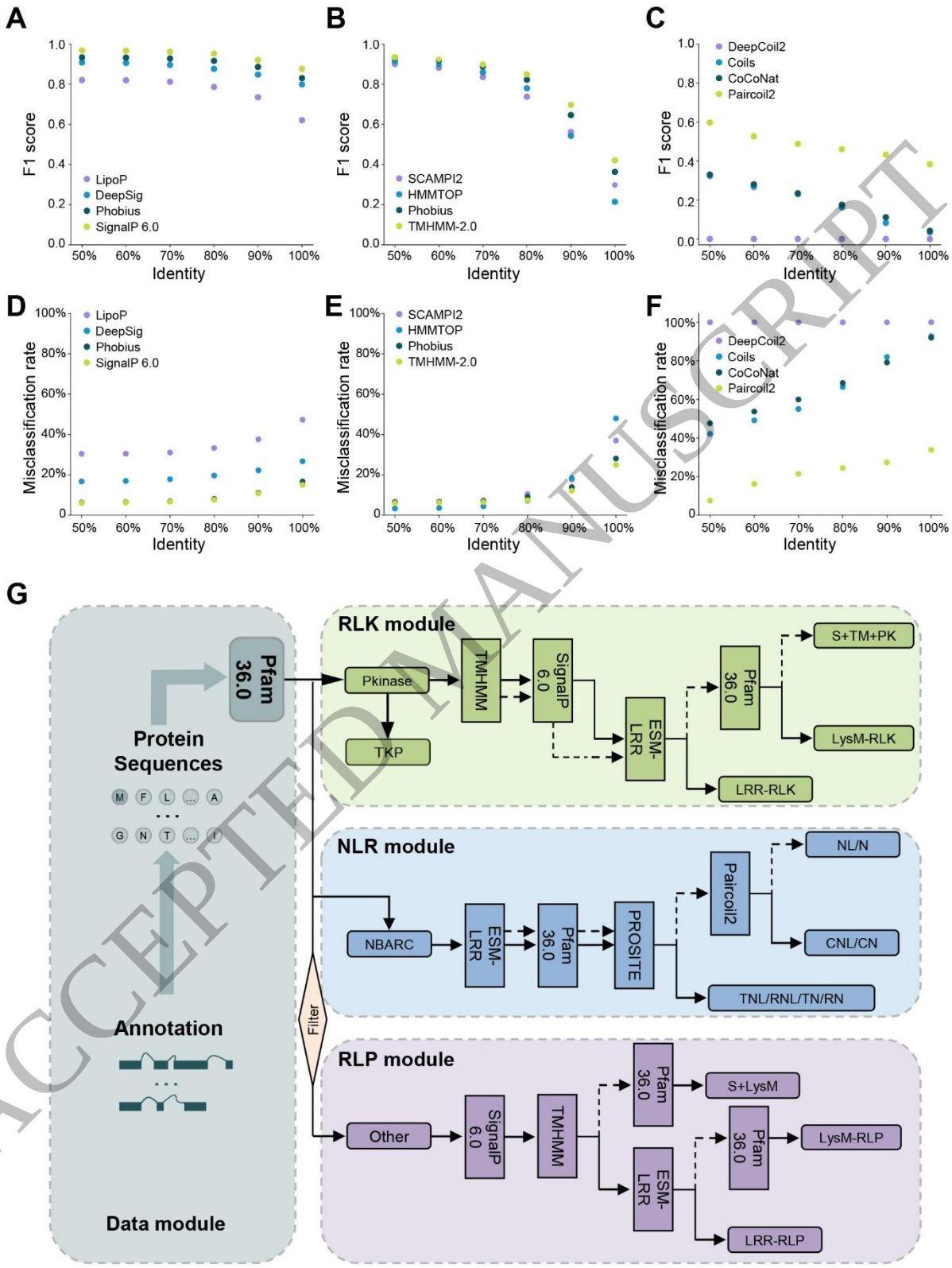

Figure 3
165x213 mm (x DPI)

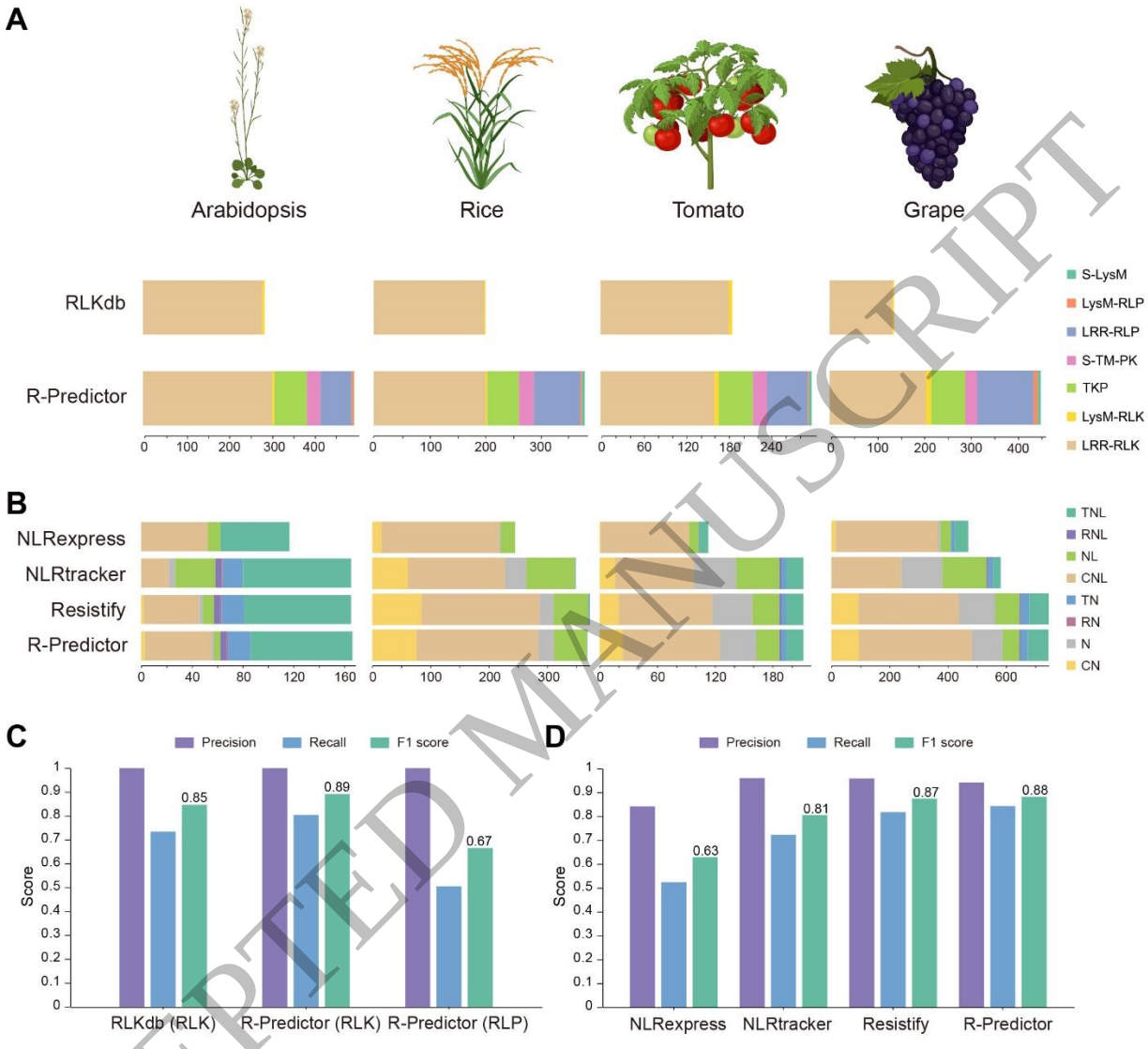

Figure 4
165x148 mm (x DPI)

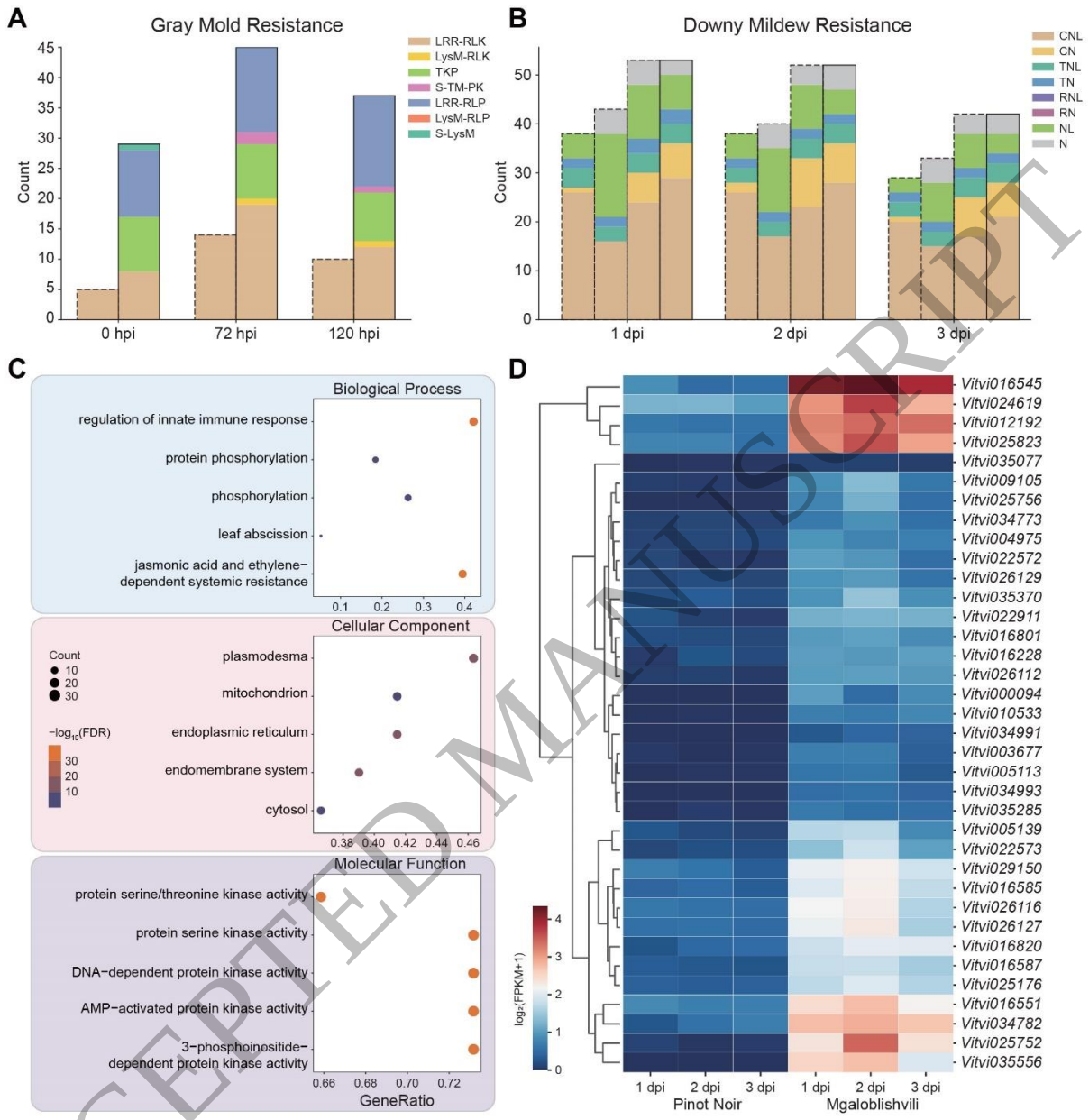

Figure 5
165x165 mm (x DPI)