JOSÉ CLEYDSON FERREIRA DA SILVA

DEVELOPMENT OF A NEW MACHINE LEARNING-DERIVED METHOD FOR HIGH-THROUGHPUT PREDICTION OF PLANT RECEPTOR-LIKE PROTEINS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

Orientadora: Elizabeth Pacheco Batista Fontes

Coorientadores: Fabyano Fonseca e Silva
Sabrina de Azevedo Silveira
Sergio Herminio Bromenschenkel

VIÇOSA - MINAS GERAIS
2020

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S586d
2020

Silva, José Cleydson Ferreira da, 1985-
Development of a new machine learning-derived method for high-throughput prediction of plant receptor-like proteins / José Cleydson Ferreira da Silva. – Viçosa, MG, 2020.
135 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Inclui apêndices.
Orientador: Elizabeth Pacheco Batista Fontes.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Receptores de células. 2. Aprendizado do computador. 3. Proteínas. 4. Proteínas quinases. I. Universidade Federal de Viçosa. Departamento de Bioquímica e Biologia Molecular. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 572.696

JOSÉ CLEYDSON FERREIRA DA SILVA

DEVELOPMENT OF A NEW MACHINE LEARNING-DERIVED METHOD FOR HIGH-THROUGHPUT PREDICTION OF PLANT RECEPTOR-LIKE PROTEINS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 28 de fevereiro de 2020.

Assentimento:

José Cleydson Ferreira da Silva
Autor

Elizabeth Pacheco Batista Fontes
Orientadora

Para minha esposa Vivian e meus filhos Luísa e Pedro...

ACKNOWLEDGMENTS

A Deus pela a saúde, família e amigos.

Em especial à minha esposa Vívian e meus filhos Luísa e Pedro pelo incentivo, compreensão e encorajamento durante todo este período.

Aos meus pais por ensinar-me com sabedoria valores que levarei por toda minha vida.

A toda família da minha esposa Vívian, pela amizade e acolhimento.

Aos meus irmãos, pela confiança e amizade.

À Universidade Federal de Viçosa, pela oportunidade de realização deste curso.

À Prof.ª Elizabeth Fontes pela orientação, incentivo, oportunidades e ensinamentos.

Ao Prof. Fabiano Silva pela coorientação e discussões e críticas valiosas sobre a tese.

A todos os amigos do Laboratório de Biologia Molecular de Plantas pelo convívio e amizade.

Aos meus amigos Otávio Brustolini e Pedro Vidigal pela troca de experiência e pela amizade.

Aos Professores Sérgio H. Bromenschenkel e Sabrina pelo apoio e pela coorientação neste trabalho.

Em especial ao meu amigo/irmão e sua esposa Vivian, pela amizade e companheirismo.

A todos que ajudaram e contribuíram direta ou indiretamente para a realização deste trabalho.

Agradeço aos órgãos de financiamento de pesquisas FAPEMIG, CAPES e CNPQ.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

ABSTRACT

SILVA, José Cleydson Ferreira, D.Sc., Universidade Federal de Viçosa, February, 2020. **Development of a new machine learning-derived method for high-throughput prediction of plant receptor-like proteins.** Adviser: Elizabeth Pacheco Batista Fontes. Co-advisers: Fabyano Fonseca e Silva, Sabrina de Azevedo Silveira and Sergio Herminio Bromenschenkel.

Machine learning (ML) is a field of artificial intelligence that has rapidly emerged in plant molecular biology, thus allowing the exploitation of massive data. The main challenges are to analyze massive datasets and extract new knowledge of cellular systems. Here, we just presented a systematic review to disentangle ML approaches is relevant for plant scientists (Chapter 1). We presented the main steps for ML development, including data selection, features extraction, training algorithms and evaluation of classification/prediction models, indicating role ML algorithm in the post-genomic era. Additionally, based on the systematic review we also developed a framework machine learning method for cell surface receptors prediction (Chapter 2). Two classes of cell surface receptors designated receptor-like protein kinase (RLK) and receptor-like protein (RLPs) are essential for perceiving and processing external and internal signals in plants and animal. Both are involved in plant development and pathogen responses and share a similar extracellular domain, capable of initial sensing environmental signal. However, RLPs have short divergent C-terminal regions not associated with conserved kinase domain characteristic of RLKs. The absence of C-terminal phylogenetic relationships between RLK and RLPs precludes the use of sequence comparison algorithms for high-throughput predictions of the RLP family. Thus, we developed the first RLP predictor in plants designated RLPredictiOme. The RLPredictiOme was implemented based on machine learning models associated with Bayesian inference. The ML models were developed in three stages to distinguish RLPs from noRLPs, RLPs from RLKs and classify new subfamilies of RLPs in plants. The evaluation of the models resulted in a high accuracy, precision, sensitivity, and specificity and relatively high probability ranging from 0.79 to 0.99 for RLPs predictions. In addition, a complete validate the of RLPredictiOme was performed with LRR-RLPs of previously characterized Arabidopsis RLPs, Arabidopsis and rice and more than 90% of known RLPs were correctly predicted. In addition to predicting previously characterized RLPs, RLPredictiOme uncovered new RLP subfamilies in the Arabidopsis genome. These include a probable lipid transfer (PLT)-RLP, plastocyanin-like-RLP, ring finger-RLP, glycosyl-hydrolase-RLP, and glycerophosphoryl diester phosphodiesterase (GDPDL)-RLP subfamilies, yet to be characterized. In comparison with the only Arabidopsis

GDPDL-RLK, molecular evolution studies confirmed that the ectodomain of GDPDL-RLPs from Arabidopsis might have undergone purifying selection with a predominance of synonymous substitutions. Expression analyses revealed that predicted GDPGL-RLPs display a basal level of expression and respond to developmental and biotic signals. The results of these biological assays substantiate the notion that the members of this subfamily have maintained functional domains during evolution and may play relevant roles in development and plant defense. Therefore, RLPredictiOme can provide new insights into the functional role of surface receptors and their relationships with different biological processes.

Keywords: Machine learning. Receptor-like protein. RLPredictiOme.

RESUMO

SILVA, José Cleydson Ferreira, D.Sc., Universidade Federal de Viçosa, fevereiro de 2020. **Desenvolvimento de um novo método derivado de aprendizado de máquina para previsão da família de receptores proteínas-like**. Orientadora: Elizabeth Pacheco Batista Fontes. Coorientadores: Fabyano Fonseca e Silva, Sabrina de Azevedo Silveira e Sergio Herminio Bromenschenkel.

O *Machine learning* (ML) é um campo de inteligência artificial que emergiu rapidamente na biologia molecular de plantas, permitindo assim a exploração de grandes conjuntos de dados. Os principais desafios são analisar esses dados e extrair novos conhecimentos de sistemas celulares. Nesta investigação, considera-se que uma revisão sistemática para desembaraçar as abordagens de ML seria relevante para os cientistas de plantas (capítulo 1). As principais etapas do desenvolvimento do ML, como seleção de dados, extração de recursos, algoritmos de treinamento e avaliação de modelos de classificação/previsão. Além disso, discute-se o papel do ML na era pós-genômica. Além disso, com base na revisão sistemática, foi desenvolvido um método de aprendizado de máquina para classificação de receptores de superfície celular (Capítulo 2). Duas classes de receptores da superfície celular, designados receptores do tipo cinase (RLK) e receptores do tipo proteína (RLPs), são essenciais para perceber e processar sinais externos e internos em plantas e animais. Ambos estão envolvidos no desenvolvimento da planta e na resposta a patógenos e compartilham um domínio extracelular similar, capaz de detectar o sinal ambiental externo do evento de detecção inicial. No entanto, as RLPs têm regiões C-terminais curtas e divergentes, não associadas com característica de domínio das quinases conservadas das RLKs. A ausência de relações filogenéticas C-terminais entre RLK e RLPs impede o uso de algoritmos de comparação de sequência para previsões de alto rendimento da família RLP. Assim, foi desenvolvido o primeiro preditor de RLP em plantas designadas como RLPredictiOme. O RLPredictiOme foi implementado com base em modelos de aprendizado de máquina em conjunto com a inferência bayesiana. Os modelos de ML incorporam em três estágios para distinguir RLPs de noRLPs, RLPs de RLKs e classificar novas subfamílias de RLPs em plantas. Os resultados da avaliação dos modelos mostram alta precisão, sensibilidade, especificidade e probabilidade relativamente alta variando de 0,79 a 0,99 para RLPs. Além disso, uma validação completa do RLPredictiOme foi realizada com LRR-RLPs de Arabidopsis e arroz previamente caracterizados e mais de 90% dos RLPs conhecidos foram previstos corretamente. Além de prever RLPs previamente caracterizados, o RLPredictiOme descobriu novas subfamílias de RLP no genoma de Arabidopsis. Isso inclui probable lipid transfer (PLT)- RLP, plastocyanin-

like-RLP, ring finger-RLP, glycosyl-hydrolase-RLP, e glycerophosphoryl diester phosphodiesterase (GDPDL)-RLP, ainda a serem caracterizadas. Em comparação com o único GDPDL-RLK, os estudos de evolução molecular confirmaram que o ectodomínio dos GDPDL-RLPs pode ter sido submetido a uma seleção purificadora com predominância de substituições sinônimas. As análises de expressão revelaram que o GDPGL-RLPs previstos exibem um nível basal de expressão e responde a sinais de desenvolvimento e sinais bióticos. Os resultados desses ensaios biológicos substanciam a noção de que os membros desta subfamília mantiveram domínios funcionais durante a evolução e podem desempenhar papéis relevantes no desenvolvimento e na defesa das plantas. Portanto, o RLPredictiOme pode fornecer novas idéias sobre o papel funcional dos receptores de superfície e suas relações com diferentes processos biológicos.

Palavras-chave: Aprendizagem de máquina. Receptores proteínas-like. RLPredictiOme.

SUMMARY

GENERAL INTRODUCTION

The advances of high-throughput sequencing technologies have allowed the complete sequencing of many plant genomes generating a considerable amount of data, which in turn requires sophisticated computational tools for analyses. Machine learning (ML) provides theoretical and technical bases for data mining and knowledge extraction from these large databases. ML is a field of artificial intelligence, which has been applied largely in bioinformatics and, to a lesser extent, in several other areas of knowledge (Larranaga et al., 2016). In molecular biology, ML has rapidly emerged as a powerful tool not only for the analyses of massive datasets but also for the interpretation of genome-wide studies generating new knowledge (Ma et al., 2014). A variety of ML approaches and training algorithms have already been employed in plant molecular biology to access new data and gain new insights into fundamental biological processes. Examples of ML-based studies include analyses of plant promoters, prediction of transcription factor target genes, prediction of microRNAs, global analysis of gene expression, and emerging studies in plant immunity (Shahmuradov et al., 2005, Cui et al., 2014, Ancilo et al., 2007, Xuan et al., 2013, Pal et al., 2016, Kushwaha et al., 2016, Silva et al., 2017). Nevertheless, in crucial areas of plant molecular biology, the ML algorithms have been underemployed by plant scientists. Signaling and transduction pathways, which are essential for plants to process developmental programs and to cope with biotic and abiotic stresses, represent major fields of cell biology that could benefit enormously from ML approaches.

Plants have developed sophisticated mechanisms for the perception of external signals due to the constant exposure to different environmental conditions of biotic and abiotic stresses. Likewise, tissue differentiation and developmental programs are triggered by endogenous signals via cell surface receptor recognition. The receptor-like kinases (RLK) and receptor-like proteins (RLP) are the two largest groups of cell surface receptors that play an essential role in both plant immunity and development (Tang et al., 2017, He et al., 2018). The structural organization of RLKs is composed of a signal peptide, an extracellular sensor domain (ectodomain), a transmembrane segment, and an intracellular kinase domain (Walker, 1994). The RLPs proteins, in turn, are membrane receptor proteins that share structural

similarity with RLKs but have lost the cytoplasmic kinase domain, which became a short cytosolic region. Although they do not have a kinase domain, RLPs function as true signaling receptors, which are capable of perceiving specific stimuli via the ectodomain but require dimerization or multimerization with RLKs and/or cytoplasmic kinases for transducing and relay the signal intracellularly (Jamieson et al., 2014).

Due to its essential role in plant development and defense, the RLK family has been extensively investigated and characterized in several different plant species (Sakamoto et al. 2012). They comprise a large protein family with more than 400 members identified in Arabidopsis, 647, 1418 and 1131 members in tomato, soybean, and rice, respectively (Sakamoto et al., 2012, Liu et al., 2015, Shiu et al., 2004). Functional studies have been conducted for several members of this family (Gomez-Gomez et al., 2000, Zipfel et al., 2006, Miya et al., 2007, Zorzatto et al., 2015; Teixeira et al., 2019). In contrast, few members of the RLP subfamily have been identified, and their biological functions are still poorly understood, in spite of their conceptual relevance in development and plant defense. In fact, genome-wide studies that could exploit the complexity of RLPs have been restricted to the LRR-RLP subfamily (Jamieson et al., 2018). Therefore, a characterization of the plant RLP subfamily has not been conducted and hence functional studies are lacking.

As an advantage in phylogenetic studies, RLKs harbor a conserved kinase domain that represented the bases for applying sequence comparison algorithms (Walker, 1994). These phylogenetic analyses clustered the kinase domains of RLKs together with similar motif-containing ectodomains as the bases for classification into 15 subfamilies of RLKs, which were designated according to the type of ectodomains, including LRR-RLKs, LysM-RLK, Malectin-RLK, etc. It is reasonable to suppose that, during evolution, RLPs have lost the kinase domain but kept similar ectodomains to RLKs because the experimentally identified RLPs contain at least five types of RLK ectodomains. The lack of the conserved kinase domain precludes the use of sequence comparison algorithms for phylogenetic studies of the RLP family. An alternative approach could be the use of ML-derived methods for the prediction of RLPs and classification into subfamilies. The objective of the present investigation was two-fold. In Chapter I, a review to disentangle ML approaches and their applications in plant molecular biology is discussed. In Chapter II, the major goal was to develop a machine learning-derived method for high-throughput prediction of plant RLPs.

In first chapter, the main ML approaches and methods for extracting attributes from amino acid sequences, attribute evaluation, and the main ML algorithms applied in plant molecular biology were revised. In the second chapter, we propose the RLPredictiOme, a machine learning-derived method associated with Bayesian inference, capable of performing high-throughput predictions of RLP subfamilies. To delimit the RLP subfamilies, the method used, as structural models, the ectodomains for the entire family of RLKLs. In addition to incorporating into the ML algorithms attributes related to amino acid composition and chemical nature as mono, di, and tripeptide, additional filters were applied related to the structural organization of RLPs, as the presence of signal peptide, a single transmembrane segment and the absence of a kinase domain at the C-terminus.

Reference

CUI, SONG et al. An improved systematic approach to predicting transcription factor target genes using support vector machine. **PLoS One**, 9.4. 2014.

GÓMEZ-GÓMEZ, LOURDES AND THOMAS BOLLER. FLS2: an LRR receptor–like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. **Molecular cell**, 5.6, pp. 1003–1011. 2000.

HE, YUNXIA et al. Plant cell surface receptor-mediated signaling–A common theme amid diversity. **J Cell Sci**, 131.2, jcs209353. 2018.

JAMIESON, PIERCE A, LIBO SHAN, AND PING HE. Plant cell surface molecular cypher: receptor-like proteins and their roles in immunity and development. **Plant science**, 274, pp. 242–251. 2018.

KUSHWAHA, SANDEEP K et al. NBSPred: a support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction. **Bioinformatics**, 32.8, pp. 1223–1225. 2016.

LARRANAGA, PEDRO et al. Machine learning in bioinformatics. **Briefings in bioinformatics**, 7.1, pp. 86–112. 2006.

LIU, JINYI et al. Soybean kinome: functional classification and gene expression patterns. **Journal of experimental botany**, 66.7, pp. 1919–1934. 2015.

MA, CHUANG, HAO HELEN ZHANG, AND XIANGFENG WANG. Machine learning for big data analytics in plants. **Trends in plant science**, 19.12, pp. 798–808. 2014.

MIYA, AYAKO et al. CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. **Proceedings of the National Academy of Sciences**, 104.49, pp. 19613–19618. 2007.

PAL, TARUN, VARUN JAISWAL, AND RAJINDER S CHAUHAN. DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. **Computers in biology and medicine,** 78, pp. 42–48. 2016.

SAKAMOTO, TETSU et al. The tomato RLK superfamily: phylogeny and functional predictions about the role of the LRRII-RLK subfamily in antiviral defense. **BMC plant biology**, 12.1, p. 229. 2012.

SHAHMURADOV, ILHAM A, VIKTOR V SOLOVYEV, AND AJ GAMMERMAN. Plant promoter prediction with confidence estimation. **Nucleic acids research**, 33.3, pp. 1069–1076. 2005.

SHIU, SHIN-HAN AND ANTHONY B BLEECKER. Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in Arabidopsis. **Plant physiology**, 132.2, pp. 530–543. 2003.

SILVA, JOSE CLEYDSON F et al., Geminivirus data warehouse: a database enriched with machine learning approaches. **BMC bioinformatics**, 18.1, p. 240. 2017.

TANG, DINGZHONG, GUOXUN WANG, AND JIAN-MIN ZHOU. Receptor kinases in plant pathogen interactions: more than pattern recognition. **The Plant Cell,** 29.4, pp. 618–637. 2017.

TEIXEIRA, RUAN M et al. Virus perception at the cell surface: revisiting the roles of receptor-like kinases as viral pattern recognition receptors. **Molecular plant pathology**, 20.9, pp. 1196–1202. 2019.

WALKER, JOHN C. Structure and function of the receptor-like protein kinases of higher plants. **Plant molecular biology**, 26.5, pp. 1599–1609. 1994.

XUAN, PING et al. MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs. **PLoS One**, 6.11. 2011.

ZIPFEL, CYRIL et al. Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts Agrobacterium-mediated transformation. **Cell**, 125.4, pp. 749–760. 2006.

ZORZATTO, CRISTIANE et al. NIK1-mediated translation suppression functions as a plant antiviral immunity mechanism. **Nature**, 520.7549, pp. 679–682. 2015.

CHAPTER I

Machine learning approaches and their current application in plant molecular biology: A systematic review

This paper was published in July 2019 in the Plant science journal, volume 284, pages 37-47.

Silva, J.C.F., Teixeira, R.M., Silva, F.F., Brommonschenkel, S.H. and Fontes, E.P., 2019. Machine learning approaches and their current application in plant molecular biology: A systematic review. Plant Science, 284, pp.37-47.

Machine learning approaches and their current application in plant molecular biology: a systematic review

Jose Cleydson F. Silva[a,b], Ruan M. Teixeira[a,b], Fabyano F. Silva[c], Sergio H. Brommonschenkel[a,d], Elizabeth P. B. Fontes[a,b]

[a]National Institute of Science and Technology in Plant-Pest Interactions, Bioagro, Universidade Federal de Viçosa, Av. PH Rolfs s/n, Centro, Viçosa, MG, 36570-000, Brazil.
[b]Departament of Biochemistry and Molecular Biology/Bioagro, Universidade Federal de Viçosa, Viçosa, MG, Brazil.
[c]Department of animal science, Universidade Federal de Viçosa, Viçosa, MG, Brazil.
[d]Departament of Phytopathology/ Bioagro, Universidade Federal de Viçosa, Viçosa, MG, Brazil

*Correspondence: bbfontes@ufv.br

Highlights

- Source databases for plant molecular data
- Main machine learning concepts and tools
- Machine learning in plant molecular biology

Abstract. Machine learning (ML) is a field of artificial intelligence that has rapidly emerged in molecular biology, thus allowing the exploitation of Big Data concepts in plant genomics. In this context, the main challenges are given in terms of how to analyze massive datasets and extract new knowledge in all levels of cellular systems research. In summary, ML techniques allow complex interactions to be inferred in several biological systems. Despite its potential,

ML has been underused due to complex computational algorithms and definition terms. Therefore, a systematic review to disentangle ML approaches is relevant for plant scientists and has been considered in this study. We presented the main steps for ML development (from data selection to evaluation of classification/prediction models) with a respective discussion approaching functional genomics mainly in terms of pathogen effector genes in plant immunity. Additionally, we also considered how to access public source databases under an ML framework towards advancing plant molecular biology and introduced novel powerful tools, such as deep learning.

Abbreviations: AUC, area under curve; FDR, false discovery rate; MCC, Matthews correlation coefficient; ML, Machine learning; MLP, multilayer perceptron; NB, naive Bayes; RF, random forest; SVM, support vector machine; SMO, sequential minimal optimization; DL, deep learning; CNN, convolutional neural network; DNN, deep neural network; RNN, recurrent neural network; TFBS, transcriptional factor binding site.

Keywords: Big data, computational intelligence, deep learning, gene expression, plant immunity

1. Introduction

Molecular biology focuses mainly on the study of interactions between different cellular systems and how these interactions are regulated. Molecular biology studies combine sophisticated techniques in genetics, biochemistry, and biotechnology to answer complex questions [1]. Advances in high-throughput sequencing technologies have enabled the rapid increase of public databases and have introduced genomics into the Big Data and postgenomic eras [2,3]. Therefore, many significant challenges have emerged, including how to analyze massive datasets and how to mine and extract new information on molecular biology. In this context, machine learning (ML) provides the technical basis for data mining to extract information from these larger raw databases.

ML is considered to be an application of artificial intelligence (AI), which is defined as a field of computer science that gives machines the ability to learn through training datasets [4]. ML is classified as supervised or unsupervised learning. In supervised learning, the data are labeled in classes, whereas in unsupervised learning, the data are not labeled. The majority of ML algorithms use supervised learning, thus requiring sophisticated computational strategies that can be partitioned into different stages. The development of supervised algorithms has always been a process whose quality depends on the accomplishment of fundamental steps, which include the following: (i) the choice of data sources, which is extremely important and requires more accurate databases and the minimum of data redundancy; (ii) the extraction of features from the selected data; (iii) the evaluation and selection of main attributes to be investigated; (iv) the choice of algorithms; (v) the creation of prediction/classification models; and (iv) the prediction/classification performance evaluation.

ML has been widely applied in several areas of knowledge, including medicine [5,6], meteorology [7], robotics [8] climatology [9] and bioinformatics [10,11]. Specifically, in plant studies, ML has been suggested for high-throughput data analysis in all levels of studies, such as genomics, transcriptomics, proteomics, and metabolomics. ML has also been applied for functional protein classification, with an emphasis on ribosomal proteins in plants [12], and for classification of genes and genera in the *Geminiviridae* family, which is a virus family that infects a broad range of cultivated and noncultivated plants [13]. In addition, ML has been

used for image processing to assess salt stress tolerance in wheat [14], the classification of grapevine varieties [15] and the detection of bacterial infection in melon plants [16].

ML has also been applied to the annotation of a large number of sequence elements, including transcription start sites, promoters, enhancers and splicing regions [17], thus being judged by scientists as a promising high-performance system. In biology, many traditional methods that are used in genomic studies are based primarily on statistics, which brings the need to reframe and rethink new procedures for extremely nonlinear systems [18]. In this context, ML has been shown to be a suitable mechanism for inferring nonlinear relations in biological systems due to the large volume and variety of data from different categories that should be included in a full dataset analysis. Some studies have shown that ML techniques, such as the artificial neural network, outperformed the use of traditional statistical methods in plant research [19, 20, 21].

Despite the aforementioned potential of ML, it has still been underexploited by plant scientists because of its complex biological systems that demand sophisticated computational algorithms. In this review, we present the main steps for the development of ML tools in molecular biology, which extend from the choice of the data source, the presentation of the primary plant databases and the data preprocessing stages for the extraction and evaluation of attributes for the composition of the training dataset. Furthermore, we present the main training algorithms that were used in the classification tools, including techniques for model evaluation and metrics for performance measurement. We also present ML applications in the prediction of micro-RNAs, a plant promoter analysis, the prediction of transcription factor-targeted genes and protein classification, a global analysis of gene expression and the functional analysis of protein interactions for the identification of plant resistance genes and pathogen effector genes, thus highlighting how these technologies facilitate basic and applied research in plant molecular biology. In summary, we expect that this systematic review will enable plant scientists with minimal programming skills to develop and apply ML to access new results and gain new research insights in molecular biology.

2. Source databases for plant molecular data

Public databases are essential sources of information for plant molecular biology. In this context, reliable data sources, cured databases, and more secure databases are necessary aspects to be considered when selecting the data to be used for training under an ML framework. Some relevant databases are described below.

Plant genomics data can be found in Phytozome v12.1.6 (http://phytozome.jgi.doe.gov/pz/portal.html), which hosts 93 assembled and annotated genomes, from 82 Viridiplantae species [22]. The Arabidopsis Information Resource (TAIR) (http://arabidopsis.org) provides a curated database of the available molecular biology data for plant Arabidopsis thaliana [23]. The SOL Genomics Network (SGN; http://sgn.cornell.edu) have extensive data for species of the Solanaceae family [24]. Data for legume species are available in the database Legume Information System (LIS) at http://legumeinfo.org, which contains all sequenced legume species and provides a set of gene families to allow traversal studies among orthologous and paralogous sequences across legume species [25]. Other platforms and databases such as GenBank (https://www.ncbi.nlm.nih.gov/genbank/), EMBL (https://www.ebi.ac.uk/), and DDBJ (https://www.ddbj.nig.ac.jp) have many different molecular datasets.

The plant disease resistance databases are essential data resources for obtaining datasets of plant disease resistance genes. Sanseverino [26] developed a database that represents the first bioinformatic resource providing a comprehensive overview of resistance genes (R-genes) in plants. The Plant Resistance Genes database (PRGdb; http://prgdb.org) is a database of resistance genes (R-genes) in plant genomes, which confers disease resistance against pathogens [27]. In addition, the SolRgene database, which is dedicated to Solanum species, provides data on the disease resistance genes in the tuber-bearing Solanum species [28].

Global gene expression (RNA-Seq), which is chromatin immunoprecipitation for the identification of protein binding sites (Chip-seq) and other types of sequencing-derived data can be obtained from the Gene Expression Omnibus-GEO database [29]. However, the understanding of the cellular function requires knowledge of all functional interactions between the expressed proteins; hence, machine learning models and specific databases of

protein-protein interaction, such as Stringdb (https://string-db.org/) and Plant Metabolic Network (www.plantcyc.org), are of particular relevance.

3. Main machine learning concepts and tools

3.1 Feature extraction methods

Classification of subcellular localization and the protein function by using the amino acid primary sequence is closely correlated with the protein biological functions and is a common approach used by ML prediction tools. The classification of protein function requires the reduction of the amino acid alphabet in the process of feature extraction [30]. Feature extraction is undeniably a useful process, not just for data exploration and understanding, but also for improving prediction accuracy and reducing computational requirements.

Physicochemical features have frequently been used in several ML prediction studies. There are different types of features that may numerically represent primary sequences of peptides and proteins, such as (i) amino acid composition, (ii) dipeptide composition (DPC), (iii) and tripeptide composition (TPC), which are proposed by Bhasin and Raghava [31], (iv) normalized Moreau–Broto autocorrelation (NMBroto) [32], (v) Moran autocorrelation (AAindex) [33], (vi) Geary autocorrelation [34], (vii) sequence-order-coupling number (SOCNumber) [35] (vii) pseudoamino acid composition (PAAC) [36], (viii) composition, transition and distribution of the various structural [37], (ix) physicochemical properties and (x) pseudo amino acid composition (PseAAC) descriptors [38, 39]. The PAAC features have been used to compose the datasets of the prediction model for disease resistance genes in plants [40]. Recently, new attributes have been suggested using the chemical properties of amino acid side chains (CPAASC) [12]. In general, these attributes allow for inferring the correlation of chemical properties of proteins with their functions. A summary of attribute extraction metrics is shown in Table 1.

Table 1. List of the most used descriptors in the literature.

Groups	Descriptors	Reference
AAindex	AAindex (AAINDEX)	Kawashima et al. [33]
Amino acid side chains	Amino acid side chains	Carvalho et al. [12]
Amino acid side chains and nucleotide proportion	Amino acid side chains and proportions of adenine (A), thymine (T), cytosine (C), and guanine (G)	Silva et al. [13]
Amino acid composition	Amino acid composition (AAC)	Chou et al. [36]
	Enhanced amino acid composition (EAAC)	Saravanan and Gautham. [129]
	The composition of k-spaced amino acid pairs (CKSAAP)	Chen et al. [130]
	Dipeptide composition (DPC)	Bhasin and Raghava [31]
	Dipeptide deviation from the expected mean (DDE)	Saravanan and Gautham [131]
	Tripeptide composition (TPC)	Bhasin and Raghava [31]
Autocorrelation	Moran (Moran)	Horne [32]
	Geary (Geary)	Sokal and Thomson [34]
	Normalized Moreau-Broto (NMBroto)	Horne [32]
BLOSUM62	BLOSUM62 matrix	Lee et al. [132]
Binary	Binary (BINARY)	Chen et al. [133]
Conjoint triad	Conjoint triad (CTriad)	Shen et al. [134]
	Conjoint k-spaced triad (KSCTriad)	Chen et al. [130]
C/T/D	Composition (CTDC)	Cai [37]
	Transition (CTDT)	Dubchak et al. [135]
	Distribution (CTDD)	Dubchak et al. [135]
Grouped amino acid composition	Grouped amino acid composition (GAAC)	Lee et al. [136]
	Enhanced grouped amino acid composition (GEAAC)	
	The composition of k-spaced amino acid group pairs (CKSAAGP)	-
	Grouped dipeptide composition (GDPC)	-
	Grouped tripeptide composition (GTPC)	-
K-nearest neighbor	K-nearest neighbor for proteins (KNNprotein)	Needleman and Wunsch [137]
	K-nearest neighbor for peptide (KNNpeptide)	-
Predicted secondary structure	Secondary structure elements content (SSEC)	Jones [138]
	Secondary structure elements binary (SSEB)	
Predicted protein disorder	Disorder (Disorder)	Obradovic et al. [139]
	Disorder content (DisorderC)	-
	Disorder binary (DicorderB)	-
Predicted accessible surface area	Accessible surface area (ASA)	Faraggi et al. [140]
Predicted main-chain torsional angles	Torsional angles (TA)	Faraggi et al. [140]
Pseudo-amino acid composition	Pseudo-amino acid composition (PAAC)	Chou [36]
	Amphiphilic PAAC (APAAC)	Chou [36]
PSSM	Position-specific scoring matrix (PSSM) profile	Radivojac et al. [141]
Quasi-sequence-order	Sequence-order-coupling number (SOCNumber)	Schneider and Wrede [35]
	Quasi-sequence-order descriptors (QSOrder)	-
Pseudo K-tuple reduced amino acids composition	PseKRAAC	Sandberg et al. [142]
Z-scale	Z-scale (ZSCALE)	

Computational tools for attribute extraction have been developed for automating and facilitating the use of these methodologies. The tool propy is appropriate for extracting protein and peptide features [41]. Another versatile tool is the iFeature web tool (and its version as command line), which is capable of calculating and extracting 53 different types of features descriptors [42]. The R software package protr (R Core Team, 2017) for feature extraction also presents a web interface. The software for feature extraction by reducing the alphabet of 20 amino acids, which was developed by Carvalho et al. [12], can be found using the link (http://geminivirus.org:8080/Rama/).

3.2 Feature evaluation and selection methods

In ML, attributes selection allows for the identification of those features that have the least variation. The benefits of performing feature selection before data modeling are as follows: (i) overfitting rate reduction that implies lesser data redundancy, with the chance to make decisions based on noise and misleading data; and (ii) training time reduction because of a lower amount of data means a faster algorithm.

In summary, having irrelevant features in data can decrease the predictive accuracy of the classification models, because they have little variance within the data set and have equal (or very similar) values when compared to two or more classes. To remove irrelevant features, selection criteria are required. Many methods, strategies, and algorithms for feature selection are described in the literature. Frequently, they are classified as wrapper, filters and embedded methods [43, 44].

Wrapper methods are based on a predictive model to score features. For each new subset of the features, a model of training is created by using the ML algorithms. Filter methods are generally used as a step of dataset preprocessing, thus providing quick calculation, since it allows for the identification of attributes that make the ML algorithms more efficient than the predictive classification models. Embedded methods are a set of approaches that perform attribute selection as part of the model building process. Table 2 displays several algorithms for feature selection.

Algorithms mentioned in Table 2 can be implemented by using different computational tools. For the Java programming language, the Weka library may be an excellent choice [45]. The

Scikit library has several algorithms in the Python language (http://scikit-learn.org). In R, there are the caret (https://cran.r-project.org/web/packages/caret/), Boruta [46], rmcfs and mlr (https://github.com/mlr-org/mlr) packages.

Table 2. List of evaluation methods for machine learning.

Approach	Method	Classifier	Evaluation Method	Reference
Embedded	Lasso	elastic net	-	Zou and Hastie [143]
Embedded	Regularized trees	Random Forest	-	Deng and Runger [144]
Embedded	Max-relevancy, Min-redundancy	mRMR		Peng et al. [145]
Wrapper	Subset selection algorithm	-	Independent	Kohavi and John [43]
Wrapper	Genetic Algorithm	SVM	Accuracy	Peng et al. [146]
Wrapper	Iterated Local Search	Regression	Probability	Hans et al. [147]
Filter	Information Gain (IG)	-	Independent	Dembo et al. [148]
Filter	Feature Selection using Feature Similarity	-	r^2	Phuong et al. [149]
Filter	Principal component analysis (PCA)	-	Independent	Wold et al. [150]
Filter	Chi-squared test (χ^2 test)	-	Independent	Cochran [151]
Filter	Relief algorithm	-	Independent	Kira and Rendell [152]

3.3 Machine learning algorithms

The ML algorithms are supervised, unsupervised and semisupervised. Supervised algorithms generate prevision models from previously labeled data of the training set, whereas unsupervised algorithms receive unlabeled data for training, and a prediction model is created by relating and evaluating the structure of the input data. Semisupervised algorithms are techniques that use the advantage of supervised and unsupervised algorithms, in which the training set is composed of labeled and unlabeled data. Here, we will describe supervised algorithms, which are popular in computation applied to molecular biology.

3.3.1 Artificial neural network

The MLP (multilayer perceptron) algorithm is a neural network class that is used for prediction in nonlinear systems [47]. The MLP consists of three layers (input, middle and output) of artificial neurons. The attributes define the number of neurons in the input layer. The middle layer contains different layers with different amounts of neurons, which connect the information that is provided by this layer with that of the posterior layer (feedforward connections), which in this case, is the output layer. The connections are based on weights given by values that were defined in the training process so that the output values will be as

similar as possible to the values that were obtained from the training model. Network fitting is conducted by means of the backpropagation algorithm, which estimates the weights through the connections that are performed in the opposite direction of the subsequent layer (Figure 1A).

3.3.2 Naive Bayes

The NB algorithm is a probabilistic classifier that is based on the Bayes' theorem. To construct a probabilistic model, NB assumes that there is independence between the variables and the conditional probability is calculated for each instance according to the assumed classes [48, 49] (Figure 1B).

3.3.3 LogitBoost

LogitBoost is a type of boosting method that implements an additive logistic regression that is similar to the AdaBoost algorithm [50]. In summary, LogitBoost maximizes the probability of a class occurring through direct log-likelihood optimization by using the Newton-Raphson algorithm, whereas AdaBoost optimizes the exponential cost function (Figure 1C).

3.3.4 J48 in decision trees

Decision trees are methods of supervised machine learning that are widely used in classification and regression tasks. The structure of a decision tree is composed of a root node, internal nodes, and leaf nodes. The internal nodes correspond to the values of the attributes, and each leaf node of the tree contains the probability distribution and the label of the class. To reduce the possibility of overfitting, the J48 algorithm performs a series of procedures that involve mainly pruning and adjustments of missing values [51] (Figure 1D).

3.3.5 Random Forest

RF is a classification algorithm that is based on the ensemble learning method. This algorithm creates multiple trees so that the classification consensus of trees is given by a voting process that determines the classification of new instances. Several advantages of the RF algorithm, including the ability to handle noise, prevent overfitting and the ability to manage a large number of features, have been listed [52] (Figure 1E).

3.3.6 Support vector machine (SVM)

SVM is a classifying and regression method that is capable of separating different classes of data. To find a separation line (hyperplane) between data, SVM maximizes the distance between the closest points in relation to each class. The distance between the hyperplane and the first point of each class is defined as the margin. Thus, SVM operates classifications that maximize the margin. Each classification step is performed through a predefined kernel function, which can be linear, polynomial, Gaussian or sigmoidal. There are different implementations of the SVM method. The sequential minimal optimization (SMO) algorithm is one of the fastest and most straightforward to be computationally implemented, because it performs several optimizations to reduce the execution time. [53, 54] (Figure 1F).

3.4 Model validation techniques

To evaluate the performance of a classification model under an ML framework, it is necessary to use cross-validation-based techniques, such as k-fold, leave-and-out and holdout methods, which are widely employed to evaluate classification and prediction models [55, 56].

The k-fold method consists of splitting the complete set of data into k independent subsets of the same size. Each subset is used for testing, and the remaining k-1 are used for model training. This process is performed k times until all subsets are tested. The leave-one-out method is a particular kind of k-fold, with k being equal to the total number of the data. Despite performing a broad investigation into the performance of tested models, leave-one-out has a high computational cost and is indicated for circumstances that involve small datasets. The holdout method consists of splitting the whole set of data into two subsets, one for the training set and another for the test set. The dataset can be separated into equal quantities or not. After the training model, the test set is applied, and the predicted error is measured. This approach is indicated when a large amount of data is available.

In general, the holdout method is widely accepted. The use of this technique is advisable mainly when the models are fitted to obtain a higher efficiency of the classification. One or more test sets can also be applied in the model validation process. The holdout method has been used with the test set of one species of plants, while the training set was from different species (intraspecies). Furthermore, 3-fold, leave-one-out and holdout methods can be simultaneously used to evaluate the classification and predictability of models [12, 13].

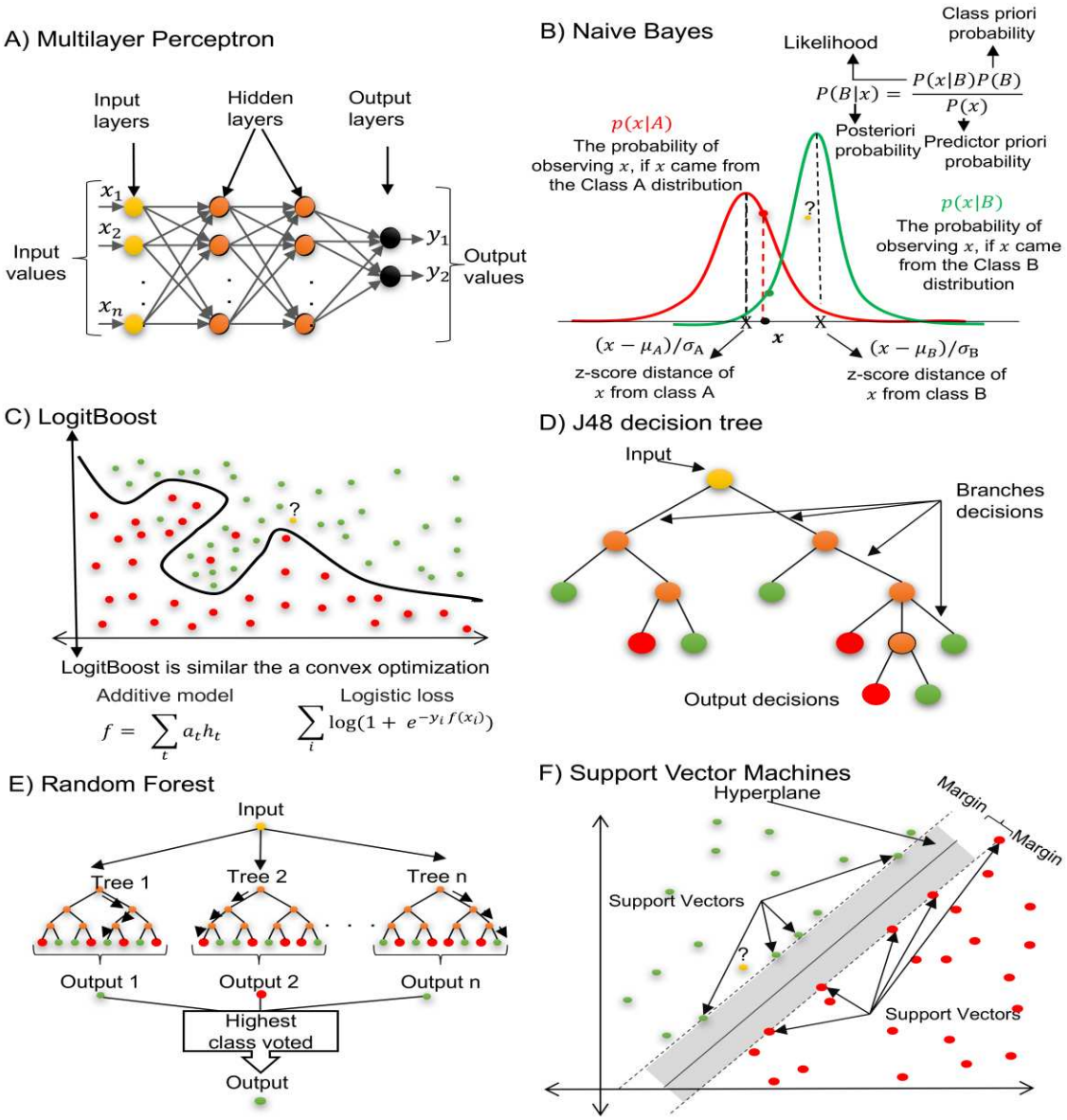

Fig. 1. Machine learning algorithm. A) An artificial neural network. The yellow balls are the neurons that define the input layer. The orange circles are the neurons that define the intermediate or hidden layers. The black circles are neurons that define the output layers. B) A probabilistic decision process of separation of classes A (density in red) and B (density in green). The red and green dots classify a data instance of the class. The yellow dot shows a new instance to be classified. C) Decision tree. The yellow circle represents the input data. The orange circle denotes an intermediate decision criterion. The red circle displays incorrect classification. The green circle confers correct classification. D) Convex optimization (classification by approximation methods). Red points show instances of negative classes, while the green points indicate the positive classes. The yellow dot designates a new instance of data to be classified. E) The ensemble methods. The yellow dot represents the input data. Following several different random decisions, the trees are built. The orange circle shows intermediate decision criteria. The red circle represents incorrect classification. The green circle denotes correct classification. F) Support vectors machine. The green dots represent classes with positive examples and red, negative examples. The yellow dots indicate a new instance to be classified.

3.5 Performance measure metrics

There are many metrics by which to measure the performance of a classifier or predictor ML method. Different branches of science have preferences for using specific metrics that consider different goals. The main metrics for the performance evaluation of classification models are sensitivity (Recall), specificity, precision, the false discovery rate (FDR), accuracy (ACC), the F1 score, the Matthews correlation coefficient (MCC) and the area under an ROC curve (AUC). These metrics are calculated based on the confusion matrix, which is also known as the contingency matrix. The confusion matrix has two rows and two columns that report the number of true positives, false positives and false negatives. Table 3 shows the most used evaluation metric in ML [57].

These evaluation metrics have been used in many areas of large-scale data analysis, including genomics [17], transcriptomics [58], systems biology [59], the prediction of localization and the function of proteins [60]. The specificity, sensitivity, ROC curve, and MCC metrics were widely used for the evaluation of the prediction model of disease resistance gene families that encode resistance proteins (R-protein) against plant pathogens [61]. The MCC is an excellent choice for evaluating prediction models for which the data set has unbalanced classes and multiclass datasets [62, 63]. The ROC curve metrics have been used to evaluate classification models of the transcriptome analysis via a comparison of gene coexpression networks in Arabidopsis [64].

4. Deep learning

Deep learning (DL) is a subfield of machine learning that has rapidly emerged in recent years due to its potential to handle large-scale datasets such as images, texts, sound, and omics [65]. DL has been widely applied for protein classification, protein–protein interactions prediction, protein structure and functional prediction, gene expression regulation and genomics [66-71]. Additionally, it has also been applied to biomedical image and signal detection analysis for the diagnosis of diseases [72,73]. In plant science, DL was successfully used for the evaluation of biotic and abiotic stresses, genetic breeding of large crops and the diagnosis of plant diseases based on leaf images [74-79].

In summary, DL is composed of several artificial neural networks with multiple layers of nonlinear functions [79,80]; thus, it can be seen as an evolution of conventional neural

networks. In bioinformatics, DL has been distinguished by the ability to extract characteristics automatically, whereas traditional ML depends on specialized feature extraction algorithms (see topic 3.1). Among the more well-known DL tools in bioinformatics are deep neural networks (DNNs), recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [80,81]. These tools are implemented in several software, as shown in Table 4.

Table 3. Metrics to measure the performance of machine learning classifiers.

Evaluation metrics	Description	Calculations
True Positives (TP)	These are the correctly predicted positive values	-
True Negatives (TN)	These are the correctly predicted negative values	-
False Positives (FP)	When actual class is no and predicted class is yes.	Type I error rate; False alarm rate
False Negatives (FN)	When actual class is yes but predicted class is no.	Type II error rate
Accuracy (ACC)	The number of correct predictions made by the model over all kinds of predictions made.	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$
False discovery rate (FDR)	It is a method of calculating the rate of type I errors.	$FDR=\frac{FP}{(FP+TP)}$
F1 score	It is a harmonic mean of the Specificity and Sensitivity	$F1=\frac{(2TP)}{(2TP+FP+FN)}$
Precision	It is defined as the number of TP over the number of TP plus the number of FP.	$\frac{TP}{(TP+FP)}$
Matthews correlation coefficient (MCC)	It considers the classifications of true and false positives and negatives. It is a balanced measure that is used even if classes are unbalanced.	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
ROC curve	ROC curves typically feature TP rate on the Y-axis, and FP rate on the X-axis. Then, it is possible to observe the area under the curve, a larger area under the curve (AUC) is usually better.	-
Sensitivity (recall)	Sensitivity is the proportion of TP that is correctly identified by the test.	$\frac{TP}{(TP+FN)}$
Specificity	Specificity is the proportion of TN that is correctly identified by the test.	$\frac{TN}{(TN+FP)}$

4.1 Deep neural networks

DNNs consist of an input layer, several hidden layers and an output layer, which is similar to the multilayer perceptron neural network structure. Once the input data is supplied to the DNNs, the output values are computed sequentially along the network layers that are based on nonlinear functions, such as hyperbolic, sigmoid, tangent or rectified linear unit (ReLU) [83,84]. DNNs have been applied to the data analysis of high-dimensional biological data, because they are efficient in terms of computational demands, thus encouraging studies on system biology and bioinformatics.

Table 4. The main deep learning software used for plant molecular biology.

Software	Operational system	Program language	CUDA support	License	Reference
BigDL	Apache Spark	Scala		Apache 2.0	www.bigdl-project.github.io
Caffe	Linux,macOS,Windows	C++	x	BSD	www.caffe.berkeleyvision.org
Chainer	Linux, macOS	Python	x	BSD	www.docs.chainer.org
Deeplearning4j	Linux,macOS,Windows, Android	C++, Java	x	Apache 2.0	www.deeplearning4j.org
Keras	Linux,macOS,Windows	Python	x	MIT license	www.keras.io
MATLAB	Linux,macOS,Windows	C, C++, Java, MATLAB	x	Proprietary	www.mathworks.com/solutions/deep-learning.html
Microsoft Cognitive Toolkit	Linux,macOS,Windows	C++, Python	x	MIT license	www.microsoft.com/en-us/cognitive-toolkit
PyTorch	Linux,macOS,Windows	Python, C, Cuda	x	BSD	www.pytorch.org
Apache SINGA	Linux,macOS,Windows	C++	x	Apache 2.0	www.singa.incubator.apache.org
TensorFlow	Llinux,macOS,Windows, Android	Python, C++, Cuda	x	Apache 2.0	www.tensorflow.org
Theano	Linux,macOS,Windows, Android	Python	x	BSD	www.deeplearning.net/software/theano
Torch	Linux,macOS,Windows,Android, iOS	C, Lua	x	BSD	www.torch.ch

4.2 Recurrent neural networks

RNNs are useful for the processing of data sequencing, such as speech, text analysis, and data of time series (e.g., plant development and global gene expression). This tool includes a feedback loop, where the output is fed back to the network to change the result of the next step, which repeats for each subsequent step. RNNs release models that change over time and generate the precise classes of the considered time period [83,85].

4.3 Convolutional neural networks

CNN has been widely used for image classification; however, some applications have been targeted with DNA sequences for predicting protein binding sites. CNN is sequentially described as follows: (i) Initially, a dataset (for example an image) is provided for a feature map to be constructed by using the convolution mathematical method. The convolution layer performs filters and organizes the data into resource maps. (ii) Nonlinear functions are used to perform grouping, being the ReLU (rectified linear unit), which is one of the most used functions. The process of generating convolutional layers and pooling can be repeated one or more times, and this may depend on the CNN parameter setting. (iii) After multiple convolutional layers, the neural network is trained through fully connected layers. The highly

connected layers become a feature list that will be introduced into a deep neural network as an input layer and into a classifying list as an output layer [83,86].

5. Machine learning in plant molecular biology

5.1 Prediction of pre-microRNAs and mature MicroRNAs

MicroRNAs (miRNAs) are small, noncoding RNA molecules with important regulatory roles in gene expression. In plants, mature miRNAs are mostly generated from the cleavage of hairpin-structured miRNA precursors (pre-miRNAs) derived from long primary transcripts (pri-miRNAs). The small RNA (sRNA) mediates gene silencing and affects gene regulation and antiviral defense [87]. sRNA-mediated transcriptional gene silencing (TGS) and posttranscriptional gene silencing (PTGS) have been implicated in regulating host defense against pathogens. The impact of TGS and PTGS in plant innate immunity have also been associated with a conserved role for miRNAs and secondary siRNAs in regulating NB-LRR and LRR innate immune receptor gene expression and pathogen resistance [88].

Frequently, the identification of miRNAs uses rule-based approaches to identify mature miRNAs based on several criteria that are derived from analyses of known miRNAs. However, ML has been recently applied to the identification of mature miRNAs by using structural features that are extracted from miRNA sequences and/or miRNA duplexes. The identification of miRNAs in plants through ML approaches has used different methodologies and algorithms. Most of these ML-based miRNA predictors have been proposed to identify animal miRNAs, and few ML approaches have been adapted to predict plant miRNAs [89]. The PlantMiRNAPred tools perform classification of miRNAs based on an SVM algorithm for predicting plant pre-miRNAs [90]. Other tools, including miRPara [91] and MiRduplexSVM [92], use the SVM algorithm to train the classification models. The mirLocator implements the RF algorithm and has shown high accuracy and rapidity in miRNAs prediction by using features that were extracted from miRNA duplexes [89]. These tools are efficient in discovering miRNAs, with particular interest in plant system immunity.

5.2 Analysis of plant promoters

The analysis of promoters has become a robust tool for developing disease-resistant or abiotic stress-tolerant plant varieties [93]. Studies of the structural composition of the plant gene promoter architecture are still very complex due to the lack of dense datasets and methods of analysis that bring large-scale statistical significance [94].

Recent studies that apply a machine learning approach to identify binding sites in promoters of plant genes showed both distinct and conserved features with the animal promoter model. The breakthrough in sequencing technologies promoted a significant increase in the number of sequenced plant genomes, thereby bringing the need for the development of computational methods that are able to detect plant promoters, thus overcoming the current difficulties in distinguishing non-promoters from promoters [95]. The complex and specific architecture of promoters for specific genes makes it challenging to develop computational strategies for promotor recognition; therefore, the computational recognition of promoters in eukaryotic DNA has been primarily based on specific binding sites and composition of base pairs in general [96]. The first tool specific to plant promoters was TSSP-TCM, which used the SVM algorithm to predict and identify approximately 35 of 40 TATA-promoters and 21 of 25 TATA-less promoters [97]. However, recent studies have shown that TATA and initiator elements in plants are not general characteristics. In general, coding DNA sequences (CDS) have been used as negatives in training, in addition to sequences that encode other functional RNAs [95].

Another algorithm, called PromMachine, uses the frequency of tetramers for promoter prediction, in which the training set does not use the presence of TATA boxes or initiators as ground criteria [98]. This program performs satisfactorily; however, it has limitations in locating the position of the transcription start site when the TATA box is not present. This limitation is a substantial disadvantage of PromMachine, since only approximately 19% of the genes of rice and 29% of Arabidopsis genes contain the TATA box in critical promoters [98]. Nevertheless, it achieves high sensitivity and specificity values of 0.86 and 0.90 for plant promoters.

The PromPredict program uses DNA degree of stability features of upstream and downstream regions relative to the transcriptional start sites that were determined

experimentally in Arabidopsis and rice genomes and performed promoter prediction in plants with high precision [100, 101]. Other computational tools of plant promoter prediction have been developed. The comparison between NNPP v2.2, TSSP-TCM, Promoter Scan v1.7, Promoter v2.0, Prom-Machine, and PromoBot software demonstrated that PromoBot has the potential to achieve improved accuracy in promoter identification and can be extended to genomes of other eukaryotic species. In addition, TSSPlant, a new tool that identifies TATA and TATA-less promoters in a broad spectrum of plant genomes [102] can achieve a higher precision compared to other programs, above 84% precision with MCC, and approximately 80% precision for promoters without TATA-box,.

5.3 Prediction of transcription factor target genes (TFTGs)

The identification of transcription factor target genes (TFTGs) is the primary step towards understanding the regulatory mechanisms of global gene expression. Screening new transcription factor binding sites (TFBSs) is a common approach for identifying transcription factor target genes. For this purpose, experimental methods have been applied to search for new TFBSs, including promoter dissection along with a loss of function assay, and the high-throughput chromatin immunoprecipitation (ChIP) approach.

Different approaches have been applied for the identification of transcription factor binding sites. The SVM algorithm was developed by Holloway [103] to integrate different features, including binding motifs, evolutionary conservation, and an expression correlation between the transcription factor (TF) gene, target gene, and others. The SVM algorithm is also used to identify the co-occurrence and location of binding sites [104]. The molecular mechanisms underpinning endogenous gene regulation have been modeled by a hidden Markov model that is based on the binding affinity between TF and DNA [105].

A novel systematic computational approach for predicting TFTGs has been developed to identify TFBSs on a genome scale. In this approach, an ML model was created using the DNA binding site of *Arabidopsis thaliana* and the SVM algorithm [106]. The performance of the model was validated by using ten-fold cross-validation, with an area under curve value (0.73). This method was improved by using the reverse sequence of the DNA binding sites.

The DNA binding sites of the upstream region are converted in a numerical vector and are used as features for composing the training model using the SVM algorithm [107].

Deep learning is a novel ML methodology and has been used to identify new protein binding sites. The DeepBind algorithm is a deep learning algorithm, which is superior to many traditional ML algorithms [108]. DeepBind is a versatile tool that can be applied to both microarray, sequencing data, promoters, and protein binding sites. It is a training predictive, fully automatized model, which allows the processing of millions of sequences through parallel programming on a graphics processing unit (GPU). DeepBind uses deep convolutional neural networks, which is a type of neural network with multiple layers, to discover new patterns within unknown sequences. These tools and methodologies can be used or adapted to be used in plant molecular biology.

5.4 Global analysis of gene expression

Cellular activities and biological functions are controlled through complex physical and regulatory interactions of genes, which compose large and sophisticated gene regulatory networks (GRNs). The understanding of GRNs is fundamental to understanding how transcription factors regulate gene expression in a tissue-specific or stress-responsive manner in animals and plants. In plants, the biological analyses of system biology require tools that are capable of parsing a large set of data to understand how regulatory networks behave under biotic and abiotic stress conditions [109]. Transcriptome profiling technologies, such as microarray and high-throughput sequencing platforms, have made it possible to understand functional associations in the expression patterns of genes that are differentially expressed.

ML algorithms are considered to be an innovative tool for analyzing differential networks using differential gene expression data [18]. Currently, different types of ML approaches have been applied using differential gene expression data. The first ML approach, which used microarray expression data, applied SVM for classification and validation of cancer tissue samples [110]. In plants, the first exploratory analyses of expression profile data using ML were developed by Kell [111]. Then, ML was applied for studying the functional genomics of stress responses in Loblolly pine using the express microarray [112]. Other applications of ML algorithms have also allowed the identification of transcription networks

that are regulated by glucose and ABA in Arabidopsis by employing trained models with microarray data and promoters of differentially expressed genes [113]. Trained models using the SVM algorithm have been used to elucidate whether closely related plant genotypes can be distinguished by their gene expression profile [114].

Several statistical and computational methodologies of high complexity have been used to extract original meaning from large datasets. However, ML has been applied as a novel alternative methodology to computationally identify functional relationships among seed-specific genes from microarray datasets towards a better understanding of the seed biology in Arabidopsis [115]. In other studies, ML has also been used to select informative genes that are related to salt tolerance mechanisms in rice by using SVM-RFE (support vector machine recursive feature elimination), which is a variant of the SVM algorithm [116].

GRNs are reprogrammed in diverse tissues or under different conditions, but their identification has limited applicability without any specific biological context. Thus, the application of ML in network inference is crucial to complement the traditional DE analysis and is especially useful in detecting biologically essential genes; however, it is still rarely performed. Some ML-based software tools have been developed for specific purposes. The R package machine learning–based differential network analysis (mlDNA) was implemented to predict candidate stress-related genes [18]. The Beacon GRN inference tool is a method that is based on the SVM algorithm. The Beacon inference tool was trained with the expression levels of developing embryo genes and with previous knowledge of regulatory relationships to predict gene regulatory networks in Arabidopsis seed development [117].

ML provides the possibility for computational and analytical solutions that could perform an integrative analysis of the data and, hence, has been increasingly applied in biology [118]. Nevertheless, there are still a limited number of examples that illustrate the relevance of machine learning approaches to integrate the system biology in the different topics of plant science.

5.5 Machine learning in plant immunity

In plant and animal immunity, machine learning has been used for the prediction of possible structures that can be recognized or used for host defense [119]. In the development of

prediction tools to identify molecular patterns and structures that are involved in defense, the molecular components that were already described with immune activity were used as a positive dataset to train the algorithms of machine learning [120].

Learning algorithms can be necessary in defining many aspects of the innate immune system, which has little definition, in addition to the identification of R-genes that lack conserved protein domains [121]. Recently, machine learning algorithms have been applied to improve the identification of candidate R genes in plants [61,40]. Different types of sequence compositional frequencies (amino acid frequency, dipeptide frequency, tripeptide frequency, multiple frequencies, charge, and hydrophobicity composition) have been used with featuring vectors for training models with the SVM algorithm for NBS-LRR protein prediction.

In addition to the identification of R-genes, different methodologies have been described for the identification of secreted effectors by plant-pathogens. The frequencies of amino acids and the naive Bayes classifier were combined for secreted effector proteins prediction in plant pathogenic fungi [122]. Other tools are described by the same authors to identify the location of plant proteins and effector proteins [123] or to predict whether an effector or plant protein localizes to the apoplast [124]. ML has also been applied for the prediction of effector proteins of the type III (T3S) secretion system of *Xanthomonas euvesicatoria* [125]. In addition to these plant-pathogens, a database that is enriched with machine learning approaches has been dedicated to the plant ssDNA virus of the Geminiviridae family (GeminivirusDW - www.geminivirus.org), along with an approach to predict and classify effector genes and genera in the Geminiviridae family [126].

The identification of protein-protein interactions (PPIs) in the pathogen-plant interaction systems represents a challenge to decipher the intricate molecular mechanism of innate plant immunity and pathogen infection. ML algorithms have been used to investigate the organization of Arabidopsis gene networks in the immune response of PTI and ETI with the ability to distinguish immune response from normal growth or immune-related genes that are associated with PTI and ETI [127]. Dong et al. [127] applied the network-guided forest (NGF) algorithm [128] to identify critical genes/interactions that are involved in immune response. A more intensive application of machine learning will bring a new bias for research in plant immunity, through observation of gene profiles and the discovery of resistance genes.

6. Future directions of ML in plant molecular biology

The volume of datasets in plant molecular biology has steadily grown, thus demanding the development of sophisticated computational tools. Therefore, advancement in graphics processing unit (GPU) technologies is essential to enable the application of complex ML algorithms, such as deep learning. However, more studies are required to apply this technology on large-scale dataset in the field of bioinformatics and computational biology. In this context, we believe that ML tools have a promising future, mainly in systems biology building that is based on new signaling pathways identification, gene expression regulation, and epigenetics modeling.

General workflows of ML methods are well-established and enable the identification of patterns in the process of extraction and evaluation features. Thus, researchers can benefit from this methodology to analyze data sets of DNA or amino acid sequences. Identification of patterns in sequences in the context of plant molecular biology is important, because they can direct laboratory assays or experiments, such as the loss of function assays [12]. In deep learning, the process of extraction of attributes is usually automatic (convolutional layers), and the characteristics are common for any type of data. Therefore, we embrace the potential of this tool for understanding all aspects of plant functional genomics.

7. Conclusion

The utilization of machine learning techniques has been applied in plant molecular biology-related research to analyze, detect and extract information from all levels of studies, including genomics, transcriptomics, proteomics, and metabolomics. Indeed, the transformation of the massive volume of data that was generated by high-throughput technologies into new and valuable knowledge has been one of the most critical challenges in plant molecular biology. Therefore, machine learning has become an important and emerging technology for research in plant molecular biology.

Despite the potential of machine learning, it has been underexploited by plant scientists in the diverse subjects of plant molecular biology. Plant immunity, which represents sophisticated biological systems whose components are not still completely unveiled, is considered to be a major field of research that could benefit enormously from ML approaches. Indeed, the application of this innovative technique to plant immunity may result in the

identification of new molecular pattern recognition receptors (PRR), R resistance genes and molecular mechanisms that determine the interplay between immunity to different plant pathogens in system biology.

8. Acknowledgments

This research was financially supported by the following grants from Brazilian government agencies: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the National Institute of Science and Technology in Plant-Pest Interactions (INCTIPP).

9. Reference

[1] A. Akhtar, E. Fuchs, T. Mitchison, R.J. Shaw, D. S. Johnston, A. Strasser, S. Taylor, C. Walczak, M. Zerial, A decade of molecular cell biology: achievements and challenges, Nat. Rev. Mol. Cell Biol. 12 (2011) 669–674.

[2] D. Eisenberg, E.M. Marcotte, I. Xenarios, T.O. Yeates, Protein function in the post-genomic era, Nature. 405 (2000) 823–826.

[3] Z.D. Stephens, S.Y. Lee, F. Faghri, R.H. Campbell, C. Zhai, M.J. Efron, R. Iyer, M.C. Schatz,S. Sinha, G.E. Robinson, Big Data: Astronomical or Genomical?, PLoS Biol. 13 (2015) e1002195.

[4] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature. 559 (2018) 547–555.

[5] J. Kang, R. Schwartz, J. Flickinger, S. Beriwal, Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective, Int. J. Radiat. Oncol. Biol. Phys. 93 (2015) 1127–1135.

[6] B. Zhang, X. He, F. Ouyang, D. Gu, Y. Dong, L. Zhang, X. Mo, W. Huang, J. Tian, S. Zhang, Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma, Cancer Lett. 403 (2017) 21–27.

[7] J. Rhee, J. Im, Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data, Agric. For. Meteorol. 237-238 (2017) 105–122.

[8] P. Gastaldo, L. Pinna, L. Seminara, M. Valle, R. Zunino, A tensor-based approach to touch modality classification by using machine learning, Rob. Auton. Syst. 63 (2015) 268–278.

[9] K. Fang, C. Shen, D. Kifer, X. Yang, Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network, Geophys. Res. Lett. 44 (2017).

[10] H. Bhaskar, D.C. Hoyle, S. Singh, Machine learning in bioinformatics: A brief survey and recommendations for practitioners, Comput. Biol. Med. 36 (2006) 1104–1125.

[11] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, Brief. Bioinform. 18 (2016) 851-869.

[12] T.F.M. Carvalho, J.C.F. Silva, I.P. Calil, E.P.B. Fontes, F.R. Cerqueira, Rama: a Machine Learning Approach for Ribosomal Protein Prediction in Plants. Sci. rep. 7 (2017) 1-16273.

[13] J.C.F. Silva, T.F.M. Carvalho, E.P.B. Fontes, F.R. Cerqueira, Fangorn Forest (F2): a machine learning approach to classify genes and genera in the family Geminiviridae, BMC Bioinformatics. 18 (2017) 431.

[14] A. Moghimi, C. Yang, M.E. Miller, S.F. Kianian, P.M. Marchetto, A Novel Approach to Assess Salt Stress Tolerance in Wheat Using Hyperspectral Imaging, Front Plant Sci. 9 (2018) 1182.

[15] S. Gutiérrez, J. Fernández-Novales, M.P. Diago, J. Tardaguila, On-The-Go Hyperspectral Imaging Under Field Conditions and Machine Learning for the Classification of Grapevine Varieties, Front Plant Sci. 9 (2018) 1102.

[16] M. Pineda, M.L. Pérez-Bueno, M. Barón, Detection of Bacterial Infection in Melon Plants by Classification Methods Based on Imaging Data, Front Plant Sci. 9 (2018) 164.

[17] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (2015) 321–332.

[18] C. Ma, M. Xin, K.A. Feldmann, X. Wang, Machine Learning-Based Differential Network Analysis: A Study of Stress-Responsive Transcriptomes in Arabidopsis, Plant Cell. 26 (2014) 520–537.

[19] M. Landín, R. Rowe, P. York, Advantages of neurofuzzy logic against conventional experimental design and statistical analysis in studying and developing direct compression formulations, Eur. J. Pharm. Sci. 38 (2009) 325–331.

[20] J. Gago, L. Martínez-Núñez, M. Landín, P. Gallego, Artificial neural networks as an alternative to the traditional statistical methodology in plant research, J. Plant Physiol. 167 (2010) 23–27.

[21] J. Gago, M. Landín, P.P. Gallego, Artificial neural networks modeling the in vitro rhizogenesis and acclimatization of Vitis vinifera L., J. Plant Physiol. 167 (2010) 1226–1231.

[22] D.M. Goodstein, S. Shu, R. Howson, R. Neupane, R.D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, D.S. Rokhsar, Phytozome: a comparative platform for green plant genomics, Nucleic Acids Res. 40 (2011) 1178-1186.

[23] S.Y. Rhee, The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community, Nucleic Acids Res. 31 (2003) 224–228.

[24] L.A. Mueller, The SOL Genomics Network. A Comparative Resource for Solanaceae Biology and Beyond, Plant Physiology. 138 (2005) 1310–1317.

[25] S. Dash, J.D. Campbell, E.K. Cannon, A.M. Cleary, W. Huang, S.R. Kalberer, V. Karingula, A.G. Rice, J. Singh, P.E Umale, N.T Weeks, Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family, Nucleic Acids Res. 44 (2015) 1181–1188.

[26] W. Sanseverino, G. Roma, M.D. Simone, L. Faino, S. Melito, E. Stupka, L. Frusciante, M.R. Ercolano, PRGdb: a bioinformatics platform for plant resistance gene analysis, Nucleic Acids Res. 38 (2009) 814–821.

[27] C.M. Osuna-Cruz, A. Paytuvi-Gallart, A.D. Donato, V. Sundesha, G. Andolfo, A.A. Cigliano, W. Sanseverino, M.R Ercolano, PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes, Nucleic Acids Res. 46 (2017) 1197-1201.

[28] V.G. Vleeshouwers, R. Finkers, D. Budding, M. Visser, M.M Jacobs, R.V. Berloo, M. Pel, N. Champouret, E. Bakker, P. Krenek, H. Rietman, SolRgene: an online database to explore disease resistance genes in tuber-bearing Solanum species, BMC Plant Biol. 11 (2011) 116.

[29] R. Edgar, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (2002) 207–210.

[30] E.A. Weathers, M.E. Paulaitis, T.B. Woolf, J.H. Hoh, Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein, FEBS Letters. 576 (2004) 348–352.

[31] M. Bhasin, G.P.S. Raghava, Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition, J. Biol. Chem. 279 (2004) 23262–23266.

[32] D.S. Horne, Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities, Biopolymers. 27 (1988) 451–477.

[33] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2007) 202–205.

[34] R.R. Sokal, B.A. Thomson, Population structure inferred by local spatial autocorrelation: An example from an Amerindian tribal population, American Journal of Physical Anthropology. 129 (2005) 121–131.

[35] G. Schneider, P. Wrede, The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site, Biophys. J. 66 (1994) 335–344.

[36] K.-C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins. 43 (2001) 246–255.

[37] C. Cai, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nucleic Acids Res. 31 (2003) 3692–3697.

[38] K.-C. Chou, Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect, Biochem Biophys Res Commun. 278 (2000) 477–483.

[39] K.-C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics. 21 (2004) 10–19.

[40] T. Pal, V. Jaiswal, R.S. Chauhan, DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants, Comput. Biol. Med. 78 (2016) 42–48.

[41] D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics. 29 (2013) 960–962.

[42] Z. Chen, P. Zhao, F. Li, A. Leier, T.T Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.C Chou, J. Song, iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics. 34 (2018) 2499–2502.

[43] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.

[44] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.

[45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software, SIGKDD Explor. 11 (2009) 10.

[46] M.B. Kursa, W.R. Rudnicki, Feature Selection with the Boruta Package, J. Stat. Softw. 36 (2010).

[47] M. Gardner, S. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmos. Environ. 32 (1998) 2627–2636.

[48] D.D. Lewis, Naive (Bayes) at forty: The independence assumption in information retrieval, Machine Learning: ECML-98 Lect. Notes Comput. Sc. (1998) 4–15.

[49] E. Frank, L. Trigg, G. Holmes, I.H. Witten, Machine Learning, Mach. Learn. 41 (2000) 5–25.

[50] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), Ann. Stat. 28 (2000) 337–407.

[51] A. Tiwari, P. Aditya, Improving classification of J48 algorithm using bagging, boosting and blending ensemble methods on SONAR dataset using WEKA. Int. J. Eng. Tech. Res. 9 (2014) 207-209.

[52] C. Lindner, P.A. Bromiley, M.C. Ionita, T.F. Cootes, Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting, IEEE Trans Pattern Anal Mach Intell. 37 (2015) 1862–1874.

[53] S. Shevade, S. Keerthi, C. Bhattacharyya, K. Murthy, Improvements to the SMO algorithm for SVM regression, IEEE Trans. Neural Netw. 11 (2000) 1188–1193.

[54] L.J. Cao, S.S. Keerthi, C.J. Ong, J.Q. Zhang, U. Periyathamby, X.J. Fu, H.P. Lee, Parallel Sequential Minimal Optimization for the Training of Support Vector Machines, IEEE Trans Neural Netw. 17 (2006) 1039–1049.

[55] S. Geisser, The Predictive Sample Reuse Method with Applications, J. Am. Stat. Assoc. 70 (1975) 320–328.

[56] B. Efron, C. Stein, The Jackknife Estimate of Variance, Ann. Stat. 9 (1981) 586–596.

[57] D. M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn Tech. 2 (2011) 37-63.

[58] J. Song, J. Zhai, E. Bian, Y. Song, J. Yu, C. Ma, Transcriptome-Wide Annotation of m5C RNA Modifications Using Machine Learning, Front. Plant Sci. 9 (2018) 519.

[59] X. Zhang, M.L. Acencio, N. Lemke, Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review. Front. Physiol. 7 (2016) 75.

[60] N. Zhang, R.S.P. Rao, F. Salvato, J.F. Havelund, I.M. Møller, J.J. Thelen, D. Xu, MU-LOC: A Machine-Learning Method for Predicting Mitochondrially Localized Proteins in Plants, Front. Plant Sci. 9 (2018) 634.

[61] S.K. Kushwaha, P. Chauhan, K. Hedlund, D. Ahrén, NBSPred: a support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction, Bioinformatics. 32 (2015) 1223–1225.

[62] B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta. - Protein Struct. 405 (1975) 442–451.

[63] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, PLoS ONE. 12 (2017) e0177678.

[64] C. Ma, H.H. Zhang, X. Wang, Machine learning for Big Data analytics in plants, Trends Plant Sci. 19 (2014) 798–808.

[65] D. Yu, L. Deng, Deep Learning and Its Applications to Signal and Information Processing Exploratory DSP, IEEE Signal Process Mag. 28 (2011) 145–154.

[66] S. Hashemifar, B. Neyshabur, A.A. Khan, J. Xu, Predicting protein–protein interactions through sequence-based deep learning, Bioinformatics. 34 (2018) i802–i810.

[67] Y. Chen, Y. Li, R. Narayan, A. Subramanian, X. Xie, Gene expression inference with deep learning, Bioinformatics. 32 (2016) 1832–1839.

[68] J.J.A. Armenteros, C.K. Sønderby, S.K. Sønderby, H. Nielsen, O. Winther, DeepLoc: prediction of protein subcellular localization using deep learning, Bioinformatics. 33 (2017) 3387–3395.

[69] J. Wang, H. Cao, J.Z.H. Zhang, Y. Qi, Computational Protein Design with Deep Learning Neural Networks, Sci Rep. 8 (2018).

[70] J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, et al., Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network, J Comput Chem. 35 (2014) 2040–2046.

[71]. S. Hochreiter, M. Heusel, K. Obermayer, Fast model-based protein homology detection without alignment, Bioinformatics. 23 (2007) 1728–1736. doi:10.1093/bioinformatics/btm247.

[72] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, D. Feng, Early diagnosis of Alzheimer's disease with deep learning, 29 (2014) 1015-1018, EEE 11th Proc IEEE Int Symp Biomed Imaging.

[73]. D. Wang, A. Khosla, R. Gargeya, H. Irshad and A.H Beck, Deep learning for identifying metastatic breast cancer, 1 (2016) 1-6. arXiv preprint

[74] A.K. Singh, B. Ganapathysubramanian, S. Sarkar, A. Singh, Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives, Trends Plant Sci. 23 (2018) 883–898.

[75] S. Ghosal, D. Blystone, A.K. Singh, B. Ganapathysubramanian, A. Singh, S. Sarkar, An explainable deep machine vision framework for plant stress phenotyping, Proc. Natl. Acad. Sci. U.S.A. 115 (2018) 4613-4618. doi:10.1073/pnas.1716999115.

[76] H.S. Naik, J. Zhang, A. Lofquist, T. Assefa, S. Sarkar, D. Ackerman, et al., A real-time phenotyping framework using machine learning for plant stress severity rating in soybean, Plant Methods. 13 (2017).

[77] M. Brahimi, M. Arsenovic, S. Laraba, S. Sladojevic, K. Boukhalfa, A. Moussaoui, Deep Learning for Plant Diseases: Detection and Saliency Map Visualisation, Human and Machine Learning Human–Computer Interaction Series. 1 (2018) 93–117.

[78] J.G. Barbedo, Factors influencing the use of deep learning for plant disease recognition, Biosyst Eng. 172 (2018) 84–91.

[79] K.P. Ferentinos, Deep learning models for plant disease detection and diagnosis, Computers and Electronics in Agriculture. 145 (2018) 311–318. doi:10.1016/j.compag.2018.01.009.

[80] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, Brief. Bioinformatics. 5 (2016)851-69.

[81] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives,

IEEE Trans Pattern Anal Mach Intell. 35 (2013) 1798–1828.

[82] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks. 61 (2015) 85–117.

[83] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444.

[84] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, et al., Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, IEEE Signal Process Mag. 29 (2012) 82–97. doi:10.1109/msp.2012.2205597.

[85] I. Sutskever, J. Martens and G.E Hinton, Generating text with recurrent neural networks. Proc Int Conf Mach Learn, 1 (2011) 1017-1024.

[86] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE. 86 (1998) 2278–2324.

[87] S.-R. Liu, J.-J. Zhou, C.-G. Hu, C.-L. Wei, J.-Z. Zhang, MicroRNA-Mediated Gene Silencing in Plant Defense and Viral Counter-Defense, Front. Microbiol. 8 (2017).

[88] F. Li, D. Pignatta, C. Bendix, J.O Brunkard, M.M. Cohn, J. Tung, H. Sun, P. Kumar, B. Baker, MicroRNA regulation of plant innate immune receptors, Proc. Natl. Acad. Sci. U.S.A. 109 (2012) 1790–1795.

[89] H. Cui, J. Zhai, C. Ma, miRLocator: Machine Learning-Based Prediction of Mature MicroRNAs within Plant Pre-miRNA Sequences, PLoS ONE. 10 (2015) e0142753.

[90] P. Xuan, M. Guo, Y. Huang, W. Li, Y. Huang, MaturePred: Efficient Identification of MicroRNAs within Novel Plant Pre-miRNAs, PLoS ONE. 6 (2011) e27422.

[91] Y. Wu, B. Wei, H. Liu, T. Li, S. Rayner, MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences, BMC Bioinformatics. 12 (2011) 107.

[92] N. Karathanasis, I. Tsamardinos, P. Poirazi, MiRduplexSVM: A High-Performing MiRNA-Duplex Prediction and Evaluation Methodology, PLoS ONE. 10 (2015) e0126151.

[93] D. Deb, A. Shrestha, I.B. Maiti, N. Dey, Recombinant Promoter (MUASCsV8CP) Driven Totiviral Killer Protein 4 (KP4) Imparts Resistance Against Fungal Pathogens in Transgenic Tobacco, Front. Plant Sci. 9 (2018) 278.

[94] T. Morton, J. Petricka, D.L. Corcoran, S. Li, C.M. Winter, A. Carda, P.N Benfey, U. Ohler, M. Megraw, Paired-End Analysis of Transcription Start Sites in Arabidopsis Reveals Plant-Specific Promoter Signatures, Plant Cell. 26 (2014) 2746–2760.

[95] A.K.M. Azad, S. Shahid, N. Noman, H. Lee, Prediction of plant promoters based on hexamers and random triplet pair analysis, Algorithm. Mol. Biol. 6 (2011) 19.

[96] U. Ohler, H. Niemann, G.-C. Liao, G.M. Rubin, Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition, Bioinformatics. 17 (2001) 199-206.

[97] I.A. Shahmuradov, Plant promoter prediction with confidence estimation, Nucleic Acids Res. 33 (2005) 1069–1076.

[98] F. Anwar, S.M Baker, T. Jabid, M.M. Hasan, M. Shoyaib, H. Khan, R. Walshe, Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach, BMC Bioinformatics. 9 (2008) 414.

[99] P. Civáň, M. Švec, Genome-wide analysis of rice (Oryza sativa L. subsp. japonica) TATA box and Y Patch promoter elements, Genome. 52 (2009) 294–297.

[100] V. Rangannan, M. Bansal, Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition, Mol. Biosyst. 5 (2009) 1758.

[101] C. Morey, S. Mookherjee, G. Rajasekaran, M. Bansal, DNA Free Energy-Based Promoter Prediction and Comparative Analysis of Arabidopsis and Rice Genomes, Plant Physiol. 156 (2011) 1300–1315.

[102] I.A. Shahmuradov, R.K. Umarov, V.V. Solovyev, TSSPlant: a new tool for prediction of plant Pol II promoters. Nucleic Acids Res, 45 (2017), pp.e65-e65

[103] D. T. Holloway, M. Kon, C. De Lisi, Integrating genomic data to predict transcription factor binding. Genomics Inform. 16 (2005) 83-94.

[104] B. Jiang, M.Q. Zhang, X. Zhang, OSCAR: One-class SVM for accurate recognition of cis-elements, Bioinformatics. 23 (2007) 2823–2828.

[105] H. Dai, R. Umarov, H. Kuwahara, Y. Li, L. Song, X. Gao, Sequence2Vec: a novel embedding approach for modeling transcription factor binding affinity landscape, Bioinformatics. 33 (2017) 3575–3583.

[106] X. Dai, J. He, X. Zhao, A new systematic computational approach to predicting target genes of transcription factors, Nucleic Acids Res. 35 (2007) 4433–4440.

[107] S. Cui, E. Youn, J. Lee, S.J. Maas, An Improved Systematic Approach to Predicting Transcription Factor Target Genes Using Support Vector Machine, PLoS ONE. 9 (2014) e94519.

[108] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nat. Biotechnol. 33 (2015) 831–838.

[109] R. Bonneau, Learning biological networks: from modules to dynamics, Nat. Chem. Biol. 4 (2008) 658–664.

[110] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics. 16 (2000) 906–914.

[111] D.B. Kell, Genomic Computing. Explanatory Analysis of Plant Expression Profiling Data Using Machine Learning, Plant Physiol. 126 (2001) 943–951.

[112] Heath, N. Ramakrishnan, R.R. Sederoff, R.W Whetten, B.I. Chevone, C.A Struble, V.Y. Jouenne, D. Chen, L. Van Zyl, R. Grene, Studying the Functional Genomics of Stress Responses in Loblolly Pine With the Expresso Microarray Experiment Management System Int J Genomics. 3 (2002) 226–243.

[113] Y. Li, Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine, Genome Res. 16 (2006) 414–427.

[114] G. Ancillo, J. Gadea, J. Forment, J. Guerri, L. Navarro, Class prediction of closely related plant varieties using gene expression profiling, J Exp Bot. 58 (2007) 1927–1933.

[115] G.W. Bassel, E. Glaab, J. Marquez, M.J. Holdsworth, J. Bacardit, Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets, Plant Cell. 23 (2011) 3101–3116.

[116] J. Wang, L. Chen, Y. Wang, J. Zhang, Y. Liang, D. Xu, A Computational Systems Biology Study for Understanding Salt Tolerance Mechanism in Rice, PLoS ONE. 8 (2013) e64929.

[117] Y. Ni, D. Aghamirzaie, H. Elmarakeby, E. Collakova, S. Li, R. Grene, L.S.Heath, A Machine Learning Approach to Predict Gene Regulatory Networks in Seed Development in Arabidopsis, Front. Plant Sci. 7 (2016).

[118] X.-T. Yu, T. Zeng, Integrative Analysis of Omics Big Data, In: Huang T. (eds) Computational Systems Biology. Methods Mol. Biol. (2018) 109–135.

[119] G. Wang, X. Li, Z. Wang, APD3: the antimicrobial peptide database as a tool for research and education, Nucleic Acids Res. 44 (2015).

[120] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, K.-C. Chou, iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types, Anal Biochem. 436 (2013) 168–177.

[121] H.U. Stotz, R.D.O. Almeida, N. Davey, V. Steuber, G.T. Valente, Review of combinations of experimental and computational techniques to identify and understand genes involved in innate immunity and effector-triggered defence, Methods. 131 (2017) 120–127.

[122] J. Sperschneider, P.N. Dodds, D.M. Gardiner, K.B. Singh, J.M. Taylor, Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0, Mol. Plant Pathol. 19 (2018) 2094–2110.

[123] J. Sperschneider, A.M. Catanzariti, K. DeBoer, B. Petre, D.M. Gardiner, K.B. Singh, P.N. Dodds, J.M. Taylor, LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell, Sci Rep. 7 (2017).

[124] J. Sperschneider, P.N. Dodds, K.B. Singh, J.M. Taylor, ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning, New Phytol. 217 (2017) 1764–1778.

[125] D. Teper, D. Burstein, D. Salomon, M. Gershovitz, T. Pupko, G. Sessa, Identification of novel Xanthomonas euvesicatoria type III effector proteins by a machine-learning approach, Mol. Plant Pathol. 17 (2015) 398–411.

[126] J.C.F. Silva, T.F. Carvalho, M.F. Basso, M. Deguchi, W.A. Pereira, R.R. Sobrinho, P.M. Vidigal, O.J. Brustolini, F.F. Silva, M. Dal-Bianco, R.L. Fontes, E.P.B Fontes, Geminivirus data warehouse: a database enriched with machine learning approaches, BMC Bioinformatics. 18 (2017).

[127] X. Dong, Z. Jiang, Y.-L. Peng, Z. Zhang, Revealing Shared and Distinct Gene Network Organization in Arabidopsis Immune Responses by Integrative Analysis, Plant Physiol. 167 (2015) 1186–1203.

[128] J. Dutkowski, T. Ideker, Protein Networks as Logic Functions in Development and Cancer, PLoS Comp. Biol. 7 (2011) e1002180.

[129] V. Saravanan, N. Gautham, Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor, OMICS. 19 (2015) 648–658.

[130] K. Chen, L. Kurgan, M. Rahbari, Prediction of protein crystallization using collocation of aminoacid pairs, Biochem Biophys Res Commun. 355 (2007) 764-769.

[131] V. Saravanan, N. Gautham, Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor, OMICS. 19 (2015) 648–658.

[132] T.-Y. Lee, S.-A. Chen, H.-Y. Hung, Y.-Y. Ou, Incorporating Distant Sequence Features and Radial Basis Function Networks to Identify Ubiquitin Conjugation Sites, PLoS ONE. 6 (2011) e17331.

[133] Z. Chen, Y.-Z. Chen, X.-F. Wang, C. Wang, R.-X. Yan, Z. Zhang, Prediction of Ubiquitination Sites by Using the Composition of k-Spaced Amino Acid Pairs, PLoS ONE. 6 (2011) e22930.

[134] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein-protein interactions based only on sequences information, Proc. Natl. Acad. Sci. U.S.A. 104 (2007) 4337–4341.

[135] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein folding class using global description of amino acid sequence., Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 8700–8704.

[136] T.-Y. Lee, Z.-Q. Lin, S.-J. Hsieh, N.A. Bretaña, C.-T. Lu, Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences, Bioinformatics. 27 (2011) 1780–1787.

[137] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.

[138] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne, J. Mol. Biol. 292 (1999) 195–202.

[139] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, A.K. Dunker, Exploiting heterogeneous sequence properties improves prediction of protein disorder, Proteins. 61 (2005) 176–182.

[140] E. Faraggi, Y. Yang, S. Zhang, Y. Zhou, Predicting Continuous Local Structure and the Effect of Its Substitution for Secondary Structure in Fragment-Free Protein Structure Prediction, Structure. 17 (2009) 1515–1527.

[141] P. Radivojac, V. Vacic, C. Haynes, R.R. Cocklin, A. Mohan, J.W Heyen, M.G. Goebl, L.M Iakoucheva, Identification, analysis, and prediction of protein ubiquitination sites,IEEE Trans. Pattern. Anal. Mach. IntellProteins. 78 (2010) 365–380.

[142] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids, Amino Acids. J. Med. Chem. 41 (1998) 2481–2491.

[143] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Series B. Stat. Methodol. 67 (2005) 301–320.

[144] H. Deng, G. Runger, Feature selection via regularized trees, Proc. Int. Jt. Conf. Neural Netw. (2012).

[145] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern. Anal. Mach. Intell. 27 (2005) 1226–1238.

[146] S. Peng, Q. Xu, X.B. Ling, X. Peng, W. Du, L. Chen, Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines, FEBS Lett. 555 (2003) 358–362.

[147] C. Hans, A. Dobra, M. West, Shotgun Stochastic Search for "Largep" Regression, J. Am. Stat. Assoc. 102 (2007) 507–516.

[148] A. Dembo, T. Cover, J. Thomas, Information theoretic inequalities, IEEE Trans. Inf. Theory. 37 (1991) 1501–1518.

[149] T. Phuong, Z. Lin, R. Altman, Choosing SNPs using feature selection, J Bioinform Comput Biol. 4 (2005) 241-57.

[150] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52.

[151] W.G. Cochran, The Comparison Of Percentages In Matched Samples, Biometrika. 37 (1950) 256–266.

[152] K. Kira, L.A. Rendell, A Practical Approach to Feature Selection, Machine Learning Proceedings 1992. (1992) 249–256.

CHAPTER II

RLPredictiOme, a machine learning-derived method for high-throughput prediction of plant receptor-like proteins, reveals novel classes of transmembrane receptors.

Abstract

Cell surface receptors play essential roles in perceiving and processing external and internal signals that arrive at the cell surface in both plants and animals. The receptor-like protein kinase (RLK) and receptor-like protein (RLPs), two major classes of proteins with membrane receptor configuration, play a crucial role in plant development and disease defense. Although RLPs and RLKs share a similar single-pass transmembrane configuration, RLPs harbor short divergent C-terminal regions instead of the conserved kinase domain characteristic of RLKs. This RLP receptor structural configuration precludes the use of sequence comparison algorithms for high-throughput predictions of the RLP family in plant genomes. Here, we developed the RLPredictiOme implemented with machine learning models in combination with Bayesian inference, capable of predicting RLP subfamilies in plant genomes. The ML models were simultaneously trained using six features, along with three stages to distinguish RLPs from noRLPs, RLPs from RLKs and classify new subfamilies of RLPs in plants. The ML models achieved high accuracy, precision, sensitivity, and specificity for predictions RLPs with relatively high probability ranging from 0.79 to 0.99. The capacity prediction of the method was assessed with three datasets, two of them contained LRR-RLPs from Arabidopsis and rice, and the last one consisted of the complete set of previously characterized Arabidopsis RLPs. More than 90% of known RLPs were correctly predicted via RLPredictiOme in these validation tests. In addition to predicting previously characterized RLPs, RLPredictiOme uncovered new RLP subfamilies in the Arabidopsis genome. These include probable lipid transfer (PLT)-RLP, plastocyanin-like-RLP, ring finger-RLP, glycosyl-hydrolase-RLP, and glycerophosphoryl diester phosphodiesterase (GDPD GDPDL L)-RLP subfamilies, yet to be characterized. In comparison with the only ArabidopGDPDL-RLK, molecular evolution studies confirmed that the ectodomain of GDPDL-RLPs might have undergone purifying selection with a predominance of synonymous substitutions. Expression analyses revealed that predicted GDPGL-RLPs display a basal level of expression and respond to developmental and biotic signals. The results of these biological assays substantiate the notion that the members of this subfamily have maintained functional domains during evolution and may play relevant roles in development and plant defense. Therefore, RLPredictiOme can provide new insights into the functional role of surface receptors and their relationships with different biological processes.

1. Introduction

The capacity to transiently regulate cellular processes in response to external environmental signals is crucial to all living organisms. While the downstream regulatory events in a signaling cascade can involve biochemical modifications, including protein phosphorylation, ligand binding, and allosteric regulation, as well as changes in the transcription/translation profiles, the initial sensing event is predominantly mediated by membrane receptor. In plants, two major classes of proteins with membrane receptor structural configuration co-exist as receptor-like kinases (RLK) and receptor-like proteins (RLP) (Tang et al., 2017; He et al., 2018). The receptor-like kinases comprise a large family with more than 420 family members in Arabidopsis (Shiu et al., 2004). These transmembrane receptors harbor a divergent extracellular domain (ectodomain, ECD) at the N-terminal region, followed by a transmembrane segment and a C-terminal cytoplasmic signaling domain. This configuration of a single-pass transmembrane kinase receptor invokes a mechanism of ligand binding-induced homo- or hetero-oligomerization of RLKs as the essential early event for signaling and transduction from the receptor, similarly to the receptor-tyrosine kinases (RTK) of mammalian cells (Ma et al., 2016; Botos et al., 2011). In this scenario, ECD is the stimulus-sensing, ligand recognition domain that induces multimerization, and the kinase domain functions as the phosphorylation-dependent transducing module that relays the signal intracellularly.

Based on the phylogeny of their kinase domain, which organized the ectodomain into clusters of conserved motifs, RLKs are classified into 15 subfamilies. Among them, the leucine-rich repeat (LRR)-RLK subfamily is further subdivided into 13 subfamilies (LRRI-XIII) according to the LRR motif organization ranging from 3 to 26 LRRs (Shiu and Bleecker, 2001, 2003). The RLK family size has been determined in other plant species, which revealed even larger RLK gene families in the genome of soybean, rice, and tomato (Gao and Xua, 2012; Sakamoto et al., 2012; Shiu et al., 2004; Zhou et al., 2016). The complexity of the RLK superfamily may reflect the intricate coordination of plant responses to external signals during plant development and interactions with the biotic and abiotic environment. Accordingly, several RLKs have been functionally characterized in development and plant defenses. Examples of functionally described RLKs include the brassinosteroid receptor BRASSINOSTEROID INSENSITIVE 1 (BRI1) that promotes plant

growth (Li and Chory, 1997), ERECTA and ERECTA-LIKE 1 (ERL1), which recognize the peptides EPIDERMAL PATTERNING FACTOR 1 (EPF1) and EPF2 to regulate stomatal development and patterning (Lee et al., 2012), EXCESS MICROSPOROCYTES 1 (EMS1) which senses the peptide TAPETUM DETERMINANT 1 (TPD1) to control male gametophyte development (Jia et al., 2008), the peptide INFLORESCENCE DEFICIENT IN ABSCISSION (IDA) receptor, designated HAESA (HAE) and HAESA-LIKE 2 (HSL2), which control floral organ abscission and lateral root emergence (IDA) (Cho et al., 2008; Kumpf et al., 2013), the receptor CLAVATA1 (CLV1), which perceives the secreted peptide CLAVATA3 to regulate meristem proliferation (Ogawa et al., 2008), Furthermore, root and meristem development is controlled by the ROOT GROWTH FACTOR (RGF) RECEPTOR 1 (RGFR1)- RGFR5 and xylem differentiation is regulated by PHLOEM INTERCALATED WITH XYLEM [PXY; also known as TDIF RECEPTOR (TDR)], the receptor for the peptide TRACHEARY ELEMENT DIFFERENTIATION INHIBITORY FACTOR (TDIF) (Hirakawa et al., 2008), and PHYTOSULFOKINE (PSK) RECEPTOR 1 (PSKR1) recognizes the sulfated pentapeptide PSK to stimulate plant growth (Wang et al., 2015). In addition to these LRR-RLKs, a single characterized Malectin-RLK, FERONIA (FER), has been shown to perceive the peptide RAPID ALKALINIZATION FACTOR 1 (RALF1) to control root growth (Haruta et al., 2014).

RLKs are also involved in plant immunity, and function as pattern recognition receptors (PRRs), which perceive pathogen-associated molecular patterns (PAMPs) or damage-associated molecular patterns (DMAPs) presented, respectively, by pathogens and plants during infection. Interaction of PRRs with PAMPs/DAMPs initiates PAMP-triggered immunity (PTI), the first layer of the plant innate immune system (Macho and Zipfel, 2014). Examples of LRR-RLKs, characterized as PRRs, include the bacterial flagellin receptor, FLAGELLIN-SENSING 2 (FLS2), (Gómez-Gómez and Boller, 2000), the ELONGATION FACTOR-TU (EF-Tu) RECEPTOR (EFR) that recognizes EF-TU (Zipfel et al., 2006), and the PEP1 RECEPTOR 1 (PEPR1) and PEPR2, which perceive the endogenous plant elicitor peptides (Peps) (Yamaguchi et al., 2006, 2010). Apart from the LRR-RLK subfamily, the Arabidopsis LYSIN MOTIF RECEPTOR KINASE 4 (LYK4), LYK5 and CHITIN ELICITOR RECEPTOR KINASE 1 (CERK1) are characterized PRRs from the Lys-M-RLK subfamily required for Chitin signaling. They harbor extracellular lysine motifs, which recognize microbial peptidoglycans (Miya et al., 2007; Wan et al., 2008, 2012; Petutschnig et

al., 2010; Cao et al., 2014). Additionally, the LIPOOLIGOSACCHARIDE-SPECIFIC REDUCED ELICITATION (LORE) from the lectin S-domain RLK subfamily recognizes the bacterial lipopolysaccharide (LPS) (Ranf et al., 2015).

The subfamily II of the LRR-RLK family contains predominantly members that act as co-receptors of the ligand-binding receptors for signaling throughout development and in plant defense (Ma et al. 2016). Characterized members of this subfamily belong to the SOMATIC EMBRYOGENESIS RECEPTOR KINASE (SERK) class, comprising SER1-SERK5 and NSP-INTERACTING KINASE (NIK) class with NIK1-NIK3 (Sakamoto et al., 2012). SERK3, also designated BRI1-ASSOCIATED RECEPTOR KINASE 1 (BAK1), is the most-well characterized SERK co-receptor and has been shown to form active developmental complexes with BRI1, PSKR, ERECTA and HAESA and active immune complexes with FLS2, EFR and PEPR1 (Ma et al., 2016). Among NIKs, NIK1 is the most-well characterized subfamily member and is required for defense responses against begomovirus triggered by virus-derived nucleic acids (Zorzatto et al., 2015; Teixeira et al., 2019). However, the corresponding receptors for these virus-derived PAMPs remain to be identified. In contrast to its antiviral function, NIK1 has also been shown to modulate PTI negatively by complexing with FLS2 and BAK1 (Li et al., 2019).

The second class of plant transmembrane proteins, RLPs, are built into an N-terminal extracellular domain, which shares similar motifs with RLK ectodomains, an internal single transmembrane segment followed by a short cytoplasmic domain that lacks a transducing-kinase domain (Jamienson et al. 2018). RLPs are structurally similar to Toll-like receptors (TLRs) involved in mammalian immunity, which also contain a leucine rich-repeat ectodomain and a short cytoplasmic tail (Botos et al., 2011). The RLP configuration poses a higher degree of complexity for signaling as they depend on heterodimerization with RLKs or association with receptor-like cytoplasmic kinases (RLCK) for transducing a stimulus from the receptor. Accordingly, the LRR-RLP TOO MANY MOUTHS (TMM) forms complexes with LRR-RLKs ERECTA and ERL1 to perceive the EPF1 and EPF2 peptides for the regulation of stomatal patterning (Lin et al., 2017) and CLAVATA2 RLP is required for the stability of CLV1 RLK (Jeong et al., 1999). Likewise, LysM-RLPs, LYSIN-MOTIF 1 (LYM1) and LYM3, associate with the LysM-RLK CERK1 to recognize bacterial peptidoglycans (Willmann et al., 2011) and the LRR-RLP RLP23 forms a complex with the

LRR-RLK SUPPRESSOR OF BIR1-1 (SOBIR1) that recognizes NECROSIS- AND ETHYLENE-INDUCING PEPTIDE 1 (NEP1)-LIKE PROTEINS (NLPs) to trigger PTI signaling (Albert et al.,2015). In addition to these Arabidopsis RLPs, the first characterized RLP, Cf-9, was identified in tomato as a LRR-RPL and has been shown to trigger effector-triggered immunity (ETI)-like signaling, elicited specifically by the *Cladosporium fulvum* Av9 effector (Jones et al., 1994). The tomato LRR-RLP Cf-4 has also been shown to be required for resistance to *C. fulvum* expressing the *Avr4* gene (Thomas et al., 1997). Both Cf-9 and Cf-4 associate with the RLKs SOBIR1 and BAK1 to initiate receptor endocytosis and plant immunity (Postma et al., 2016).

Although some progress has been made in characterizing RLPs, a biological function has been assigned to only a few plant RLPs, despite their conceptual relevance in cell signaling events. While 15 RLK subfamilies with distinct ECD have been detected in Arabidopsis, currently, only three Arabidopsis RLP subfamilies have been identified based on single-gene identification and functional studies (He et al., 2018). The only genome-wide study of RLPs was restricted to the LRR-RLP subfamily (Jamieson et al., 2018). One possible explanation for the poor characterization of RLPs may be the difficulty of assigning members to this family based on sequence comparison, as they lack the conserved C-terminal serine/threonine kinase domain. Therefore, a complete inventory of the RLP family in the genome of different plant species is lacking, and hence functional studies have been limited. To provide a framework for the identification and prediction of RLP function, we developed the RLPredictiOme as a machine learning method associeted with bayesian inference approaches. In addition to six different features to train ML models, the method used multiple datasets based on RLK ectodomains along with the hypothesis that RLPs lack the kinase domain, but retain the same RLK receptor configuration. The ML models could distinguish RLPs from noRLPs, RLPs from RLKs and classify subfamily with relatively high accuracy, precision, sensitivity, and specificity. To prove the capacity to predict RLP families, we validated the method with biological experiments describing a new RLP family, designated GDPDL-RLP. The RLPredictiOme can facilitate the prediction and provide new insights into the role of RLPs in plants.

2. Methods

In this session, we describe and explain the complete process of the dataset composition and training of ML models with different algorithms (including Bayesian models) under a three step framework for plant RLP classification, as illustrated in Figure 1.

2.1. Reclassification of ectodomain of RLKs in plants for composed datasets

The amino acid sequences of 80 plant species were retrieved from the Phytozome v.11.1 database (https://phytozome.jgi.doe.gov/). We applied filters to remove unknown sequence proteins without functional annotation. The sequences were re-annotated using SMART (smart.embl-heidelberg.de) and Pfam (pfam.sanger.ac.uk) databases. Then, the amino acid sequences containing a predicted kinase domain were selected. The signal peptide was predicted using SignalP v.4.0 (Nielsen, 2017) and Phobius (Käll et al., 2004) software, whereas the transmembrane segment was identified using TMHMM (Sonnhammer et al., 1998) and Phobius software. Then, the sequences were filtered by using the criteria based on the presence of a signal peptide and the presence of a transmembrane segment. Furthermore, the redundant sequences were removed through CD-HIT algorithm (Fu et al., 2012). Subsequently, the amino acid sequences were clustered according to the functional domain of the extracellular ectodomain (LRR-RLK, WAK-RLK, and LysMRLK, for example) (Shiu and Bleecker, 2001, Sakamoto et al., 2012).

2.2. Dataset composition

For the classification of RLPs, we used three steps, two steps of binary classification and one multilabel classification. In summary, the first stage compares RLPs with other families of non-RLP proteins (noRLP); the second one compares RLP with receptor-like kinases (RLKs); and the third one performes the classification of a protein sequence within an RLP subfamily using the functional ectodomain present in RLKs.

In the first stage, the training dataset consisted of amino acid sequences containing the extracellular ectodomain, the region of the membrane segment, and the cytoplasmic region that precedes the kinase domain of the RLKs (but without the kinase domain) as a positive

class (RLP). The negative class is composed of full-length amino acid randomly selected sequences (noRLP); whereas the sequences of the positive class were removed from the negative dataset. To increase the number of negative examples, the dataset was divided into three different data sets.

In the second stage, the positive class contained the training dataset (RLP) and the negative class used the full-length amino acid sequences of RLKs. In the third stage, the data from the positive class of RLPs were labeled according to the reclassification of RLKs based on their ectodomain. In this case, a putative LRR-RLP, for instance, contained an ectodomain of the leucine-rich repeat kinase receptor-like kinase (LRR-RLK), a transmembrane segment and a short cytoplasmic region excluding a kinase domain. Furthermore, the full dataset was distributed into 10 different sub-datasets for working around the limitations of computational time on the training.

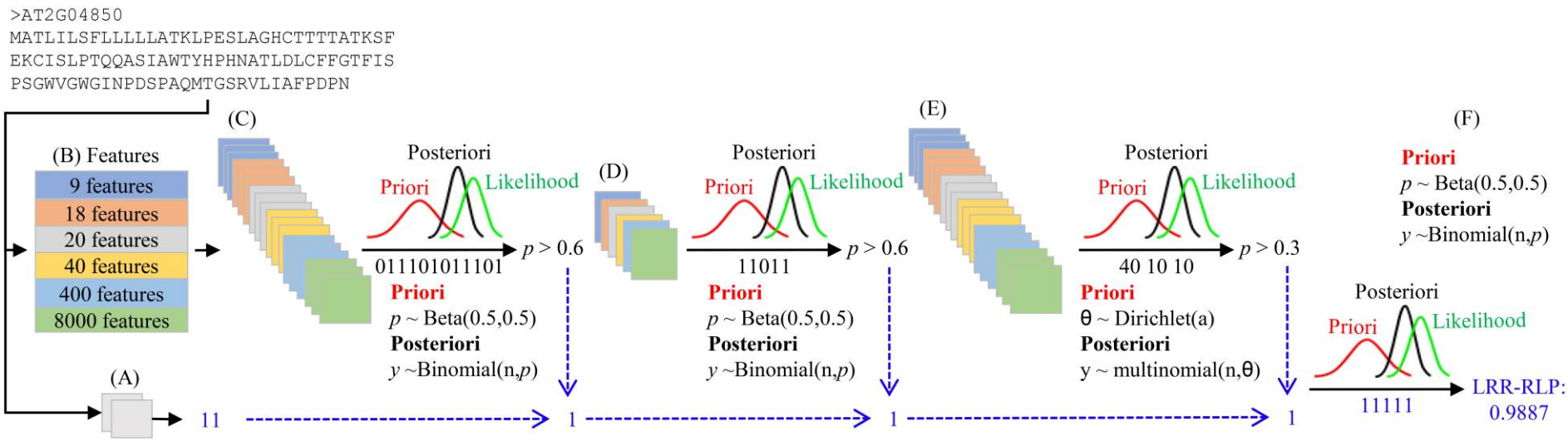

Fig. 1. Schematic representation of the RLPredictiOme method. Amino acid sequences are submitted to the method with the sequential filters, (A) the signal peptide and segment transmembrane prediction, (B) attribute vector provided to ML models, (C) the first step of the classification to distinguish RLP/noRLP, (D) the second step of the classification to distinguish RLPs/RLK, (E) third classification step. The ML Models for subfamily classification, (F) The Bayesian inference for making decision and final prediction.

2.3. Feature extraction

Six feature types for a frequency of sequence composition were calculated for each sequence. These included amino acid composition frequency of full-length sequence, amino acid composition frequency (monopeptide) of the N-terminal and C-terminal regions, dipeptide frequency, tripeptide frequency and frequency of chemical properties of amino acid side chains (CPAASC), and CPAASC2 frequency of the N-terminal and C-terminal regions. For each sequence of positive and negative datasets, a numerical feature vector was created.

1. CPAASC (Carvalho et al., 2017) feature describes the frequency of the chemical properties of amino acid side chains, such as positively charge, negatively charge, polar uncharged, aromatic, nonpolar aliphatic, hydrophobicity, volume and mass of the total number of amino acids in the full-length peptide sequence.
2. CPAASC2 is calculated by the frequency of the chemical properties of amino acid side chains of the N-terminal and C-terminal regions. The full-length sequence is split into two equal (or nearly equal) regions, and, for each one of these regions, the proportion of amino acid composition was also calculated. We consider N-terminal to be the first region of the complete amino acid sequence and C-terminal to be the second region of the full-length sequence.
3. The amino acid composition (Saravanan and Gautham, 2015) feature describes the frequency, $f(t)$, of an individual amino acid type within the total number of amino acids in the full-length peptide sequence as follow:

$$f(t) = \frac{N(t)}{N}, \quad t \in \{A, C, D, \ldots, Y\} \qquad 1$$

where: $N(t)$ is the number of amino acid types, while N is the length of the fulllength peptide sequence. The amino acid composition comprises a total of 20 features (ACDEFGHIKLMNPQRSTVWY).

4. Amino acid composition frequency is calculated by the individual amino acid type of the N-terminal and C-terminal regions. The full-length sequence is split into two equal (or nearly equal) regions and, for each of these regions, the proportion of amino acid composition frequency is also calculated. Amino acid composition frequency in the N-terminal and C-terminal region comprises a total of 40 features.

5. Dipeptide (Saravanan and Gautham, 2015) frequency, $D(r,s)$, describes all combinations of amino acids pairs. It can be calculated as:

$$D(r,s) = \frac{N_{rs}}{N-1}, \ \ r,s \ \in \{A,C,D,\dots Y\} \qquad 2$$

where: N_{rs} is the number of dipeptides represented by the types of amino acids r and s, and N is the number of amino acids of the full-length sequence. The dipeptide frequency comprises a total of 400 features.

6. Tripeptide (Bhasin and Raghava, 2004) frequency, $f(r,s,t)$, describes all combinations of three amino acids. It can be calculated as:

$$f(r,s,t) = \frac{N_{rst}}{N-2}, \ \ r,s,t \ \in \{A,C,D,\dots Y\} \qquad 3$$

where: N_{rst} is the number of tripeptides represented by the types of amino acids r, s, and t. The tripeptide frequency comprises 8000 features.

The six features were used for the training of all classification models in the three proposed steps. In summary, to compare RLPs with non-RLPS proteins (first stage), for each feature type, three training datasets totalizing 18 training sets were created. However, to compare RLPs with RLKs (second stage), one training dataset for each feature type was created. Finally, to classify RLPs within a subfamily (third stage), 10 training datasets for each feature type, resulting in 60 training sets, were created.

2.4. Method for unbalanced datasets

The superfamily RLK in plants has been broadly characterized and are subdivided into different groups with different number of members in the subfamily. The LRR-RLK is the largest subfamily, whereas other subfamilies have lower frequency of members in plants; thereby, we used the SMOTE algorithm (Chawla et al., 2002) to oversample of the minority class, which eliminates the possibility of information loss. The SMOTE creates synthetic samples based on the values of the features from the minor class.

2.5. Machine learning algorithms

The RLPredictiOme method embeds several ML models, which were built with the previously described training sets. In this study, 20 ML algorithms were tested in order to select the one that suits the supervised classification task. Those algorithms are implemented in python library Scikit-learn v.0.22.1 (Pedregosa et al., 2011). We opted to describe here only those that were selected to compose RLPredictiOme.

2.5.1. AdaBoost

The AdaBoost (Adaptive Boosting) is an algorithm based on the ensembled method used for both classification and regression procedures (Freund and Schapire, 1995). AdaBoost is widely used with other algorithms to improve the performance of classification models. The algorithm fits small decision trees across the successive modifications of the data. The data are normalized thought boosting iterations for applying weights w_1, w_2, ..., w_n for each feature of the the training sample. In this context, all the weights are first defined as $w_i = 1 / \sum (x_i)$. Thus, in a first step, the model is trained on the original data. For each iteration, the weights are modified, and the learning algorithm is applied to the reweighted data. The examples incorrectly predicted by the boosted model have their weights increased, and, for correct predictions, the weights are decreased. Each iteration of the algorithm is performed based on examples missed in the previous iteration.

Given a training set T of data x_i and label y_i (0,1) the weights are assigned as:

$$W_i^{(0)} = 1 / \sum_{1} (x_i) \quad 4$$

A set of M classifiers is initially obtained, and an iteration is initialized for each m=1 to M. The m classifier minimizes the weighted sum error for unclassified points as follow:

$$W_e = \sum_{y_i \neq m(x_i)} W_i^m \quad 5$$

Sequentially, an adjusted weight for αm of the classifier is performed according to:

$$\alpha m = \frac{1}{2}\ln\left(\frac{1-e_m}{e_m}\right); \quad e_m \frac{W_e}{W} \tag{6}$$

Newly updated weights for the data are performed in two conditions: if $m(x_i)$ is a mistake, we have:

$$W_i^{(m+1)} = W_i^m e^{\alpha m} \tag{7}$$

Otherwise;

$$W_i^{(m+1)} = W_i^m e^{-\alpha m} \tag{8}$$

2.5.2. Probability calibrated algorithm

The methods of probabilities calibration in machine learning are usually applied for mapping model predictions to posterior probabilities in tasks of binary classification (Platt et al., 1999).

The Probability calibration is performed by using sigmoid function from Platt Calibration algorithm, which is based on prediction results obtained from SVM model performed through sigmoid function:

$$p_i = \frac{1}{1 + exp(Af_i + B)} \tag{9}$$

where: A and B parameters are estimated from training set (f_i,y_i) by maximum likelihood method. In sequence, the gradient descent is applied for A and B as:

$$\underset{A,B}{argmin} = \left\{\sum_i y_i \log(p_i) + (1-y_i)\log(1-p_i)\right\} \tag{10}$$

The Positive N+ and negative N- examples from the dataset are used to avoid overfitting. In this context, Platt calibration uses the values of the response variables y+ and y- to refine the calibration as follow:

$$y_+ = \frac{N_+ + 1}{N_+ + 2}; \quad y_- = \frac{1}{N_- + 2} \tag{11}$$

However, the data used to model fitting are not the same used in the calibration process.

2.5.3. Gradient Boosting algorithm

The Gradient Boosting (GB) is an algorithm usually applied in decision trees with fixed size as base-learners (Friedman, 2002). The GB gradually builds an additive model able to allow the optimization of the loss functions and combines different predictors sequentially with shrinkage methods. This algorithm, originally proposed by Friedman (2001), gets the input of a given dataset $(x,y)_i = i^N$ in a number of iterations M, initially applied of the loss-function Y (y,f) and choice base-learner model $h = (x,o_i)$. Subsequently, the algorithm initializes f_0 with a constant, and for each iteration t in M, the negative gradient $g_t(x)$ is computed, thereby, a new base-learner function $h(x,o)$ is fitted, and the best gradient descent is found according to:

$$pt = arg\ min_p \sum \ddot{Y}\ [\mathrm{y_i} f_{\mathrm{t}-1}(\mathrm{x}-1) + ph(\mathrm{x_i}, \mathrm{o_t})] \tag{12}$$

The update in the function estimation is performed as:

$$f_t = f_{t-1} + p_t h(x, o_t) \tag{13}$$

This process repeats as long as iterations exist in M.

2.5.4. K-Nearest Neighbors Algorithm

The K-Nearest Neighbors Algorithm (KNN) is a non-parametric method used for classification tasks (Samworth et al., 2012). Several nearest neighbor search algorithms have been used due to simplicity of implementation and customization. Based on the input data, KNN initializes a number k of neighbors, and for each example of the dataset, the distance between the observed example and examples of dataset is calculated. Euclidean distance is a widely used metric in classification tasks, and it is given by:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad 14$$

2.5.5. Linear Discriminant (LDA) and Quadratic Discriminant algorithm

The Linear discriminant analysis finds the linear combination between two or more classes based in a feature vector (Hastie and Friedman, 2009). The results of the combinations of linear models can be used for the reduction of dimensionality before later classification. The LDA assumes that all classes have the same covariance matrix with a Gaussian model assigned for each class. The fitted model is used to reduce the dimensionality of the input data. In this process, some mathematical approaches such as singular value decomposition (SVD), least-square solution (LSQR), and eigenvalue decomposition (Eigen) have been successfully used. The LDA algorithm implemented in scikit learning python library uses SVD by default, which is recommended for data with many features.

The LDA algorithm applied in binary classification problems assumes a conditional probability density function, $P(X|y = k)$, for each class, thus allowing the implementation of Bayes theorem as follows:

$$p(y|X) = \frac{P(X|y)P(y)}{P(X)} = \frac{P(X|y)P(y)}{\sum_t P(X|y)P(y)} \qquad 15$$

where: o k is the class, which maximizes conditional probability; and $P(X|y)$ is given by a multivariate Gaussian distribution with the following density function:

$$P(X|y = k) = \frac{1}{2\pi^{d/2} \ |\sum_k|^{1/2}} \exp\left(-\frac{1}{2} \ (X - \mu_k)^t \sum_k^t (X - \mu_k)\right) \qquad 16$$

Where d is the number of features.

The LDA algorithm, in the Gaussian process for each class, assumes the same covariance matrix for all classes. Thus, the linear decision rule is obtained comparing the log-probability ratios as:

$$log\left(\frac{P(y|X)}{P(y|X)}\right) = log\left(\frac{P(X|y)P(y)}{P(X|y)P(y)}\right) = 0$$
$$\Leftrightarrow$$
$$(\mu_k - \mu_l)^t \sum_1^t X = \frac{1}{2}\left(\mu_k^t \sum_1^t \mu_k - \mu_l^t \sum_1^t \mu_l\right) - log\frac{P(y)}{P(y)} \quad 17$$

The LDA analysis trains a univariate linear classifier to discern binary outcomes and provide posterior value for each data point into outcome assignments. In contrast, the quadratic discriminant analysis (QDA) has no assumptions on the covariance matrices $\sum_k$ of the Gaussians model, inducing to the quadratic decision surfaces (Kim et al., 2011).

2.5.6. Logistic Regression

Logistic Regression (LR) is a general model for classification that estimates the effects (w_0, w_1, ..., w_n) of predictor variables (x_1, x_2, ..., x_n) on values obtained from the transformation of original binary outcomes by using a specific function named logit (Schmidt et al., 2017, King and Zeng, 2001). This model can be easily described as:

$$y = w_0 + w_1 * x_1 + w_2 * x_2 + \cdots w_n * x_n \quad 18$$

The LR provides a probability curve by inserting the y values predicted from (18) in the inverse logit function given by:

$$p = Exp(y) \,/\, (1 + Exp(y)) \quad 19$$

where: p is the probability values generated by the model fitting, which allow classifications based on thresholds (specific values of p) that define the elements to be assigned to each

class of interest.

The threshold values that minimize the misclassification rate are obtained by an algorithm based on penalization, which is a kind of dimensional reduction method controlled by two sequential equations (20 and 21) using different weights (w) for predictor variables as follow:

$$\min_{W,C} \frac{1}{2} w^t w + C \sum_{i=1}^{n} log\left(exp\left(-y_i(X_i^T w + c)\right) + 1\right) \quad 20$$

$$\min_{W,C} \|w\|_1 + C \sum_{i=1}^{n} log\left(exp\left(-y_i(X_i^T w + c)\right) + 1\right) \quad 21$$

Additionally, the LR provides information on odd ratios for each predictor variable. These ratios represent the relative importance of these variables in terms of increasing or decreasing effects on original binary outcome.

2.5.7. Deep Neural Network

The Deep Neural Network (DNN) is an artificial neural network, composed by multiples layers between the input data layer and output layer (Hinton et al., 2012). The layers are described by the input layer, two or more hidden layers and an output layer; it is similar to the multilayer perceptron neural network structure. Under this concept, the input data are supplied to the DNNs, which sequentially estimate the weights of predictor variables over the hidden layers to the output layer as follow:

$$a^l = \sigma(w^l a^{l-l} + b^l) \quad 22$$

where: l is the layer, w^l is the weight of predictor variables connecting to the l^{th} layer of neurons, b^l is a vector bias, and σ represents a general class of DNN functions defined as Rectified Linear Unit (ReLU):

$$g(z) = \max\{0, z\} \quad 23$$

In summary, ReLU is a linear function for positive values, and assumes zero for negative values. DNNs are frequently applied to high-dimensional biological data due to computational efficiency reported in different studies on system biology and bioinformatics (Silva et al., 2019).

2.6. Model testing techniques

The classification models were trained using the sub-sampling methods with 10-folds crossvalidation. Thus, the data were divided into 10 subsets, assuming the training with nine datasets and validation with one dataset. This procedure was repeated ten times, whereas the testing for RLPredictiOme method was performed with three independent datasets. One dataset is composed of 44 RLPs already described in the literature, and another datasets with 57 LRR-RLPs and legume-like (L-type) lectins, G-type lectins, calcium-dependent (C-type) lectins, and the lectin-like Lysin-motifs (LysM) describe in *Arabidopsis* (Jamieson et al., Faulkner et al., 2013, Liu et al., 2012). Besides, 100 random amino acid sequences were created by an algorithm in house to demonstrate that the classifiers do not provide random predictions.

2.7. Performance assessment of the models

Evaluation metrics generally used in bioinformatics were applied in order to choice the most efficient algorithms and training models. For each training set and algorithm, we evaluated Accuracy, F-measure, false discovery rate (FDR), Mathew's correlation coefficient (MCC), Precision, Sensitivity, and Specificity. This metrics are calculated based on the confusion matrix (contingence matrix) by means of the number of TP, TN, FP, and FN, which represent the number of true positives, true negatives, false positives, and false negatives, respectively.

The Accuracy defines the percentage of correct classification from the total number of instance positive and negative of the dataset. The Accuracy is defined as:

$$Accuracy = \frac{TP + TN}{(TP + FP + TN + FN)} \quad 24$$

The F-measure (F1 score) is defined as the harmonic mean of the precision and sensitivity as follow:

$$F - measure = \frac{2TN}{(2TN + FP + FN)} \quad 25$$

The Matthew's correlation coefficient (MCC) considers the rate predictions of true and false positives and negatives, respectively defined as sensitivity and specificity. This coefficient is given by:

$$MCC = \frac{(TP * TN) - (FN * FP)}{\sqrt{((TP + FN) * (TN + FP) * (TP + FP) * (TN + FN))}} \quad 26$$

The False discovery rate (FDR) is simply the estimate of the number of false positive results divided by all of the positive results (i.e. type I error estimates). Thus, FDR calculates the probability that a true RLP be classified incorrectly by the used model. FDR is calculated as:

$$FDR = \frac{FP}{(FP + TP)} \quad 27$$

The Precision is a metric that measures a percentage of true predictions related to falsepositive predictions. Thus, it is estimated as:

$$Precision = \frac{TP}{(TP + FP)} \quad 28$$

The Sensitivity measures a percentage of TP correctly predicted in the test, and is given by:

$$Sensitivity = \frac{TP}{(TP + FN)} \quad 29$$

The Specificity measures a percentage of TN correctly predicted. It is calculated as follow:

$$Specificity = \frac{TN}{(TN + FP)} \quad 30$$

Object-oriented programming in python was implemented to evaluate the performance of each model. For multiclass models, the PyCM (Haghighi et al., 2018) python library was used (Multiclass confusion matrix library in Python).

2.8. Bayesian inference in Ensemble Methods

Ensemble methods under a ML approach combine the decisions of several classification models in order to improve the overall performance. Thus, it is possible to avoid misclassification due to reductions in noise, bias, and data variance. In a ensemble method, several models are used to make predictions for each instance of data.

In the binary classification contrasts involving the models RLPs versus noRLPs, and RLPs versus RLKs, we assumed the results provided by *n* independent Bernoulli trials (0 or 1 values) with probability parameter π. Thus, the number of successes (x) derived from these trials follows a binomial distribution (Feller, 2008):

$$x|\pi\ bin(n, \pi) \quad 31$$

whose probability function is given by:

$$p(x|\pi) = \binom{n}{x} \pi^x (1-\pi)^{(1-x)},\ \ x = 0,1, \dots, n \quad 32$$

In this context, we assumed a Beta distribution (Gupta and Nadarajah, 2004) as the prior distribution for π:

$$\pi\ beta(\alpha, \beta) \quad 33$$

whose probability density function is given by:

$$p(\pi) = \frac{1}{B(\alpha, \beta)} \pi^{a-1}(1-\pi)^{\beta-1}, \quad 0 < \pi < 1 \quad 34$$

Under the Bayes theorem, it is shown that the posterior distribution for π (probability of success of classification) is a beta distribution with the following parameters:

$$\pi|x\ beta(\alpha + x, \beta - n - x) \quad 35$$

whose probability density function is given by:

$$p(x) = \binom{n}{x} \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)}, \quad x = 0,1, \ldots, n \quad 36$$

The multilabel models to classify RLPs sub-families have different probabilities of success. Thus, the sum of the classification success for each subfamily follows a multivariate generalization of the binomial distribution, which is named multinomial distribution. We assumed the multinomial distribution as:

$$\boldsymbol{x} \sim Multinomial(\boldsymbol{p}, \boldsymbol{n}) \quad 37$$

where *x* is a response vector, *p* is the probability of observed, and N is a vector of the total counts in each RLP sub-families. The data distribution assuming multinomial model for all trials is given by the following product:

$$P(x_{1\ldots i} \mid p_{1\ldots i}, N_{1\ldots i})\ P(x1) = \prod_i \frac{N_i!}{x_{i1}! \ldots x_{ij}!}\ p_{i1}^{xi1} \ldots p_{ij}^{xij} \quad 38$$

The prior probability widely used for multinomial models is the Dirichlet distribution, which presents the parameters π and θ. The data vector (x) account for the total counts in each RLPs sub-family. The mentioned distribution is denoted as follow:

$$p_i \sim Dirichlet(\pi_{\ k}\theta_k) \quad 39$$

whose probability density function is given by:

$$P(x1, \ldots, xk, a) = \frac{1}{B(\alpha)} \prod\prod x^{a-1} \quad 40$$

The normalizing constant, $B(\alpha)$, is the gamma function given by:

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha i)}{\Gamma(\sum_{i=1}^{k})\alpha i}, \quad \alpha = (\alpha 1, \ldots, \alpha k) \quad 41$$

We perform Bayesian inference using the Bayesian statistical modeling and PyMC3 library python, which uses the Markov chain Monte Carlo (MCMC) algorithms to explore the posterior distributions (Salvatier et al., 2016). Based on previous analyses with MCMC chains, we opted to use a single chain with 10,000 iterations per each amino acid sequence. We set the burn-in to 2,000 interactions and four chains for all models. The Gibbs sampler algorithm was used to generate random samples from the posterior distribution for all analysis (Geman and Geman, 1984).

2.9. RLP subfamilies downstream analysis

The function domain prediction analysis was carried out with the Pfam database v32.0 (http://pfam.xfam.org/) with a Hidden Markov Model (HMM) algorithm implemented in Hummer software. The signal peptide and transmembrane segment were predicted with Signalp v.4.0 and TMHMM software, respectively (Nielsem et al., 2017). The topology diagram was performed with Protter Web server (Omasits et al., 2014). The sequence alignment of the RLP superfamily was conducted using Muscle algorithm (www.ebi.ac.uk/Tools/msa/muscle/). The phylogenetic analysis was performed by the maximum likelihood statistical method with 10.000 bootstraps using FastTree software (Price et al., 2010). The tree was edited using the FigTree (www.tree.bio.ed.ac.uk/software/figtree/) software.

The gene expression of the glycerophosphoryl diester phosphodiesterase RLP subfamily was investigated through for the meta-analysis of transcriptomes using Geneinvestigator V3 (Hruz et al., 2008) and ePlant (Waese et al., 2017) (www.bar.utoronto.ca/eplant) for the expression in tissues and responses to pathogens.

2.10. Protein-protein interaction (PPI) network Analysis

GDPDLs- and SNC4-interacting proteins from Arabidopsis were used as a query term to identify their respective interactions described in the BAR database (Bio-Analytic Resource for Plant Biology, http://bar.utoronto.ca/interactions/). The IntAct and Biogrid databases were selected for searching. The protein-protein interactions (PPI) were visualized into the Cytoscape software (https://cytoscape.org/) to visualize the firework topology of the interactions-network and to measure the network centrality metrics for each protein. We used network centralities, betweenness, closeness, eccentricity, and degree.

Briefly, the betweenness centrality in the PPI network of the graph G = (V,E) was calculated by the number of times a protein interact along the shorter paths among all nodes. The betweenness centrality can be analyzed as:

$$\sum_{s,t\xi V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \qquad 42$$

where: v is the total set of proteins, σ(s,t) is the number of shorter paths of (s,t), and σ(s,t|v) is the number of paths crossing v (Brandes, 2001). The Closeness centrality of a protein v is the sum of the shortest path distances from w to all other proteins and calculated as:

$$\frac{1}{\sum_{w\xi V} dist(v,w)} \qquad 43$$

The eccentricity centrality of a protein v is the maximum distance from v to all other proteins of the graph G and can be analyzed as:

$$\frac{1}{maxdist(v,w): w\xi V} \qquad 44$$

Finally, the degree centrality of protein *v* was calculated by $\sum_i Kdi$ where *d* represents each adjacent protein and *K*, the total number of adjacent proteins.

2.11. Plant Growth, treatment with flg22 and viral infection with TRV and CaLCuV

Arabidopsis thaliana ecotype Columbia (Col-0) of different ages were used in all gene expression experiments. The seeds were germinated on half-strength Murashige and Skoog (MS) plates containing 10% (w/v) sucrose and 0.8% (w/v) agar, sterile and grown under normal growth conditions, at 21◦C under a 16 h light/8 h dark cycle. After growing for 10 days, the seedlings were transferred to a tissue culture plate contained 2 mL of 100 nM flg22 and incubated for 15 minutes. For the viral infection assay with Tobacco rattle vírus (TRV), Agrobacterium cultures containing TRV-RNA1 (pTRV1) and TRV-RNA2 (pTRV2) T-DNA constructs. were infiltrated onto the lower leaf of four-leaf stage *N. benthamiana* plants using a 1-mL needleless syringe. Infected leaves were confirmed by conventional RT-PCR using TRV-RNA2-specific primers. TRV was mechanically inoculated to *A. thaliana* grown in soil in a growth chamber for 14 days, by rubbing the leaves with sap (0.05 M K_2HPO_4 , pH 7.2, 0.01 M Na_2SO_3) from infected *N. benthamiana* leaves. After 2 weeks of inoculation, viral infection was confirmed by RT-PCR. For infection with *Cabbage* leaf curl virus CaLCuV, plants at the seven-leaf stage were inoculated with plasmids containing partial tandem repeats of CaLCuV DNA-A and DNA-B (Fontes et al., 2004), using biolistic delivery. Inoculated plants were transferred to a growth chamber, and infection was confirmed by conventional PCR using CaLCuV DNA-B-specific primers.

2.12. RNA extraction, synthesis of cDNA and qRT-PCR Analysis

For quantitative RT-PCR, total RNA was extracted from frozen leaves or seedlings with TRIzol (Invitrogen) according to the instructions from the manufacturer. RNA was treated with 2 units of RNase-free DNase (Promega). First-strand cDNA was synthesized from 3.5 mg of total RNA using oligo-dT(18) and Transcriptase Reversa M-MLV (Invitrogen), according to the manufacturer's instructions. Real-time RT-PCR reactions were performed on an ABI7500 equipment (Applied Biosystems), using SYBR_ Green PCR Master Mix (Bio-rad). The amplification reactions were performed as follows: 2 min at 50 °C,

10 min at 95 °C, and 40 cycles of 94 °C for 15 s and 60 °C for 1 min. For quantitation of gene expression, we used specific primers for GDPLs e actin 3 (At3g53750) as the endogenous control gene for data normalization. The quantitative gene expression was done using the $2^{-\Delta CT}$ method, and values were normalized to the endogenous control. For quantification of gene expression in flowers, axis of inflorescence, pedicel and roots, the respective Col-0 organs grown in soil for approximately 30 days were used, except for the roots and Col-0 (control), which were 10 days-grown plants in MS medium, under the conditions described above.

3. Results

3.1. Revisiting the ectodomain of the RLK superfamily in plants

We performed a survey in the genome of 80 plant species to identify the functional ectodomains of RLKs based on *in silico* models. A total of 40,418 sequences were retrieved. We identified 100 classes of RLK ectodomains associated with C-terminal kinase domains (Table 1A). However, most of these ectodomains generated subfamilies with less than 10 members. Identity sequences higher than 0.85 were removed through CD-hit software. Additionally, only sequences with a single membrane segment were selected. A total of 14,787 amino acid sequences were recovered, and their ectodomains were used as positive datasets for filtering RLPs/noRLPs and RLPs/RLKs.

Table 1A. Domain localized in ectodomain of the kinase in plants.

Description	Total
LRR-RLK	14087
Unknown-RLK	10020
S-domain-RLK	3859
Malectin-RLK	3299
Salt-stress-response/antifungal-RLK	2345
L-Lectin-RLK	2213
WAK-RLK	1844
B-lectin-RLK	549
LysM-RLK	381
WAK-EGF-RLK	285
EGF-like-RLK	212
WAK-EFG-RLK	177
RCC1-RLK	148
B-Lectin-RLK	145

PAN-RLK	131
C-Lectin-RLK	90
Glycosyl-hydrolases-RLK	90
Thaumatin-RLK	86
NAF-RLK	79
Ethylene-responsive-RLK	74
EF-hand-RLK	50
Cache-RLK	32
Chitinase-RLK	15
PAS-RLK	12
Plastocyanin-like-RLK	12
Ring-finger-RLK	9
Adenovirus E3-RLK	8
CHASE-RLK	8
Cysteine-rich-secretory-RLK	7
GDPDL-RLK	7
Universal-stress-RLK	6
ACT-RLK	5
Probable-lipid-transfer-RLK	5
Ankyrin-Kinase	4
Chromo-RLK	4
PAN-like-Kinase	4
PB1-RLK	4
Sel1-RLK	4
Alpha/beta-hydrolase-RLK	3
Cytochrome P450-RLK	3
Helix-loop-helix-DNA-binding-RLK	3
Histidine-phosphatase-RLK	3
Major-Facilitator-RLK	3
MatE-RLK	3
PPR	3
PPR-RLK	3
Phospholipase-RLK	3
Proline-rich-RLK	3
Sugar-(and other)-transporter-RLK	3
Transmembrane-RLK	3
Alpha-amylase-catalytic-RLK	2
Barwin-RLK	2
C2-RLK	2
CUB-RLK	2
DUF1084-RLK	2
DUF726-RLK	2
Endomembrane-RLK	2
GAF-domain	2
GTPase-RLK	2
Glycosyl hydrolases-RLK	2
Glycosyltransferase-RLK	2
HAD-RLK	2

HAD-hyrolase-like-RLK	2
MSP-RLK	2
NB-ARC-RLK	2
PQQ-enzyme-RLK	2
Peptidase-RLK	2
PfkB-RLK	2
Wnt-and-FGF-inhibitory-regulator-RLK	2
Adenylate-cyclase-associated-(CAP)-N-terminal-RLK	1
Alcohol-dehydrogenase-GroES-like-RLK	1
Aldose-1-epimerase-RLK	1
Ankyrin-RLK	1
Castor-and-Pollux-RLK	1
Cyclic nucleotide-binding-RLK	1
Cyclic-nucleotide-binding-RLK	1
Cytochrome-P450-RLK	1
DEAD/DEAH-box-helicase-RLK	1
DUF1221-RLK	1
DUF674-RLK	1
FAR1-RLK	1
GATA-zinc-finger-RLK	1
Glutathione-S-transferase-RLK	1
Glycosyl-transferase-RLK	1
MRG-RLK	1
Mcm10-RLK	1
Mitochondrial-carrier-RLK	1
NAD(P)-binding-Rossmann-like-RLK	1
NAF-Domain	1
PMR5-RLK	1
Pentatricopeptide-RLK	1
RKF3-like--RLK	1
RWD-RLK	1
Rare-lipoprotein-A-RLK	1
Ribulose-phosphate-RLK	1
Subtilase-RLK	1
TPR-RLK	1
UBX-RLK	1

Table 1B. Total of receptor-like kinase protein in each subfamily.

No.	Label	Count
1	L-Lectin-RLK	980
2	LRR-RLK	5404
3	S-domain-RLK	1626
4	Malectin-RLK	1313
5	Salt-stress-response/antifungal-RLK	1004
6	WAK-RLK	1362
7	B-Lectin-RLK	362
8	Unknown-RLK	3285

10	PAN-RLK	41
11	Ethylene-responsive-RLK	29
12	Thaumatin-RLK	52
13	RCC1-RLK	65
14	Glycosyl-hydrolases-RLK	40
15	C-Lectin-RLK	21
16	Other-RLK	192

In order to represent a higher number of negative examples, three datasets were created. The first dataset contained 14,973 positives and 15,993 negative examples. The second and third ones contained the same a of examples, 14,973 positives and 15,973 negatives. To distinguish RLPs from no-RLPs, we used six features (see Methods sections) from the three datasets, thus implying in a total of 18 training sets. On the other hand, to distinguish RLPs from RLKs, only one dataset with 14,973 positive (ectodomain of the RLKs) and (full-length sequence of the RLKs) negative examples were used, that implied in six training sets based on the assumed number of features.

The RLP subfamily members were assigned according to the ectodomains of RLKs. For each training set, 16 classes named as "others" (Table 1B) were considered by grouping together the smaller subfamilies. In some plant species, uncharacterized RLKs subfamilies contain at least one to 10 members, and they were grouped in one class designated Others-RLPs. LRR-RLKs, unknown-RLK, S-domain-RLK, and WAK-RLKs are over-represented RLK subfamilies in plants, whereas thaumatin, GDPD, and malectin are small subfamilies, which are not represented in all plant species (Sakamoto et al., 2012). For each super represented subfamily, 500 sequences were randomly selected to compose 10 additional datasets. Thus, given the previously mentioned six features, a total of 60 training sets were obtained for training.

3.2. Features analysis

We implemented the RLPredictiOme method using six different types of attributes (Fig. 1). These included the frequency of the chemical properties of amino acid side chains (CPAASC), which have nine features, and CPAASC2 extracted from N-terminal and C-terminal regions with 18 features; the amino acid composition with 20 features and amino acid composition extracted from N-terminal and C-terminal regions with 40 features (Fig.

1B). Furthermore, we used dipeptide and tripeptide compositions resulting in 400 and 8000 features, respectively. The simultaneous use of six features and multiples data sets provided RLPredictiOme with information to apply Bayesian inference (see Methods section) as a powerful ensemble method to make robust predictions.

For the classification models for RLP/noRLPs (first step, Fig. 1C), the tripeptide composition was the feature with the best performance among all tested features of the models built with the RLPs/noRLPs datasets using the logistic regression algorithm (Table 2, highlighted in bold). The three models built with tripeptide composition achieved accuracy (ACC) of 0.953, 0.955, and 0.953, respectively; and Matthews correlation coefficient (MCC) of 0.906, 0.910, and 0.96, respectively. Furthermore, the false discovery rate (FDR) was lower than 0.05.

Table 2. Summarized results of the evaluation models built with the RLPs/noRPs datasets.

Data set	Algorithm	ACC	F1	FDR	MCC	Precision	Sensitivity	Specificity
RLP_vs_noRLP_AAComposition_1	LogisticRegressionCV	0.9173	0.9211	0.0878	0.8343	0.9303	0.9303	0.9032
RLP_vs_noRLP_AAComposition_2	LogisticRegressionCV	0.9205	0.9241	0.0839	0.8407	0.9322	0.9322	0.9078
RLP_vs_noRLP_AAComposition_3	LogisticRegressionCV	0.9209	0.9245	0.0831	0.8416	0.9321	0.9321	0.9088
RLP_vs_noRLP_AAComposition_N_C_terminal_1	MLPClassifier	0.9457	0.9478	0.0534	0.8912	0.9490	0.9490	0.9421
RLP_vs_noRLP_AAComposition_N_C_terminal_2	MLPClassifier	0.9468	0.9487	0.0513	0.8934	0.9487	0.9487	0.9446
RLP_vs_noRLP_AAComposition_N_C_terminal_3	MLPClassifier	0.9482	0.9499	0.0457	0.8964	0.9456	0.9456	0.9511
RLP_vs_noRLP_CPAASC_1	LinearDiscriminantAnalysis	0.9020	0.9102	0.1315	0.8074	0.9561	0.9561	0.8436
RLP_vs_noRLP_CPAASC_2	LinearDiscriminantAnalysis	0.9042	0.9120	0.1282	0.8116	0.9562	0.9562	0.8481
RLP_vs_noRLP_CPAASC_3	LinearDiscriminantAnalysis	0.9040	0.9119	0.1288	0.8113	0.9566	0.9566	0.8473
RLP_vs_noRLP_CPAASC_N_C_terminal_1	LinearDiscriminantAnalysis	0.9104	0.9172	0.1183	0.8232	0.9558	0.9558	0.8614
RLP_vs_noRLP_CPAASC_N_C_terminal_2	LinearDiscriminantAnalysis	0.9132	0.9196	0.1148	0.8284	0.9569	0.9569	0.8660
RLP_vs_noRLP_CPAASC_N_C_terminal_3	LinearDiscriminantAnalysis	0.9140	0.9204	0.1137	0.8301	0.9572	0.9572	0.8674
RLP_vs_noRLP_Dipeptide_1	MLPClassifier	0.9439	0.9457	0.0497	0.8878	0.9412	0.9412	0.9468

RLP_vs_noRLP_Dipeptide_2	MLPClassifier	0.9481	0.9500	0.0501	0.8960	0.9500	0.9500	0.9459
RLP_vs_noRLP_Dipeptide_3	MLPClassifier	0.9447	0.9466	0.0497	0.8894	0.9428	0.9428	0.9468
RLP_vs_noRLP_Tripeptide_1	**LogisticRegressionCV**	**0.9535**	**0.9551**	**0.0410**	**0.9069**	**0.9511**	**0.9511**	**0.9561**
RLP_vs_noRLP_Tripeptide_2	**LogisticRegressionCV**	**0.9550**	**0.9565**	**0.0389**	**0.9100**	**0.9519**	**0.9519**	**0.9584**
RLP_vs_noRLP_Tripeptide_3	**LogisticRegressionCV**	**0.9534**	**0.9549**	**0.0404**	**0.9067**	**0.9502**	**0.9502**	**0.9568**
	Mean	0.9303	0.9342	0.0784	0.8615	0.9480	0.9480	0.9112

For the classification models for RLPs/RLKs (second step, Fig. 1D), the amino acid composition of the N-terminus and C-terminus and tripeptide composition were the features archiving both the best performance, resulting in ACC of 0.97, MCC of 0.95 and FDR lower than 0.05 (Table 3). In the RLP subfamily models built with RLP subfamily data sets (third step, Fig. 1E), the tripeptide composition outperformed the others, with ACC and MCC of 0.984 and 0.866, respectively (Table 4).

Table 3. Summarized results of the evaluation models built with the RLPs/RLKs datasets.

Data set	Algorithm	ACC	F1	FDR	MCC	Precision	Sensitivity	Specificity
RLP_vs_RLK__AAComposition_N_C_terminal	**QuadraticDiscriminantAnalysis**	**0.9775**	**0.9773**	**0.0337**	**0.9552**	**0.9884**	**0.9884**	**0.9670**
RLP_vs_RLK_Tripeptide_	**GradientBoostingClassifier**	**0.9762**	**0.9760**	**0.0367**	**0.9527**	**0.9890**	**0.9890**	**0.9639**
RLP_vs_RLK_CPAASC_N_C_terminal	LinearDiscriminantAnalysis	0.9707	0.9706	0.0479	0.9421	0.9899	0.9899	0.9523
RLP_vs_RLK_CPAASC	LinearDiscriminantAnalysis	0.9647	0.9647	0.0572	0.9304	0.9877	0.9877	0.9426
RLP_vs_RLK_Dipeptide_	MLPClassifier	0.9627	0.9617	0.0344	0.9254	0.9579	0.9579	0.9673
RLP_vs_RLK_AAComposition	QuadraticDiscriminantAnalysis	0.9571	0.9571	0.0627	0.9151	0.9777	0.9777	0.9374
	Mean	0.9681	0.9679	0.0454	0.9368	0.9818	0.9818	0.9551

Table 4. Summarized results of the evaluation models built with the RLPs Subfamily datasets.

Dataset	Algorithm	ACC	F1	MCC	Precision	Sensitivity
RLP_Subfamily_AAComposition_10_SMOTE	LinearDiscriminantAnalysis	0.984	0.872	0.864	0.872	0.872
RLP_Subfamily_AAComposition_1_SMOTE	CalibratedClassifierCV	0.984	0.869	0.861	0.869	0.869
RLP_Subfamily_AAComposition_2_SMOTE	CalibratedClassifierCV	0.984	0.874	0.866	0.874	0.874
RLP_Subfamily_AAComposition_3_SMOTE	LinearDiscriminantAnalysis	0.984	0.873	0.864	0.873	0.873
RLP_Subfamily_AAComposition_4_SMOTE	LinearDiscriminantAnalysis	0.984	0.870	0.862	0.870	0.870
RLP_Subfamily_AAComposition_5_SMOTE	LinearDiscriminantAnalysis	0.983	0.867	0.858	0.867	0.867
RLP_Subfamily_AAComposition_6_SMOTE	LinearDiscriminantAnalysis	0.984	0.871	0.863	0.871	0.871
RLP_Subfamily_AAComposition_7_SMOTE	CalibratedClassifierCV	0.984	0.869	0.861	0.869	0.869
RLP_Subfamily_AAComposition_8_SMOTE	CalibratedClassifierCV	0.985	0.876	0.868	0.876	0.876
RLP_Subfamily_AAComposition_9_SMOTE	LinearDiscriminantAnalysis	0.984	0.875	0.867	0.875	0.875
	Mean	0.984	0.872	0.863	0.872	0.872
RLP_Subfamily__AAComposition_N_C_terminal_10_SMOTE	CalibratedClassifierCV	0.989	0.911	0.905	0.911	0.911
RLP_Subfamily__AAComposition_N_C_terminal_1_SMOTE	CalibratedClassifierCV	0.988	0.904	0.897	0.904	0.904
RLP_Subfamily__AAComposition_N_C_terminal	CalibratedClassifierCV	0.989	0.908	0.902	0.908	0.908

_2_SMOTE						
RLP_Subfamily__AAComposition_N_C_terminal_3_SMOTE	CalibratedClassifierCV	0.988	0.902	0.896	0.902	0.902
RLP_Subfamily__AAComposition_N_C_terminal_4_SMOTE	KNeighborsClassifier	0.989	0.911	0.905	0.911	0.911
RLP_Subfamily__AAComposition_N_C_terminal_5_SMOTE	KNeighborsClassifier	0.989	0.909	0.903	0.909	0.909
RLP_Subfamily__AAComposition_N_C_terminal_6_SMOTE	KNeighborsClassifier	0.988	0.903	0.896	0.903	0.903
RLP_Subfamily__AAComposition_N_C_terminal_7_SMOTE	KNeighborsClassifier	0.988	0.900	0.894	0.900	0.900
RLP_Subfamily__AAComposition_N_C_terminal_8_SMOTE	CalibratedClassifierCV	0.988	0.903	0.897	0.903	0.903
RLP_Subfamily__AAComposition_N_C_terminal_9_SMOTE	CalibratedClassifierCV	0.988	0.907	0.900	0.907	0.907
	Mean	0.988	0.906	0.899	0.906	0.906
RLP_Subfamily_CPAASC_10_SMOTE	LinearDiscriminantAnalysis	0.972	0.778	0.764	0.778	0.778
RLP_Subfamily_CPAASC_1_SMOTE	AdaBoostClassifier	0.971	0.772	0.757	0.772	0.772
RLP_Subfamily_CPAASC_2_SMOTE	AdaBoostClassifier	0.972	0.776	0.761	0.776	0.776
RLP_Subfamily_CPAASC_3_SMOTE	AdaBoostClassifier	0.972	0.773	0.759	0.773	0.773
RLP_Subfamily_CPAASC_4_SMOTE	LinearDiscriminantAnalysis	0.971	0.770	0.755	0.770	0.770
RLP_Subfamily_CPAASC_5_SMOTE	LinearDiscriminantAnalysis	0.972	0.773	0.759	0.773	0.773
RLP_Subfamily_CPAASC_6_SMOTE	LinearDiscriminantAnalysis	0.971	0.771	0.756	0.771	0.771
RLP_Subfamily_CPAASC_7_SMOTE	AdaBoostClassifier	0.972	0.773	0.758	0.773	0.773
RLP_Subfamily_CPAASC_8_SMOTE	LinearDiscriminantAnalysis	0.972	0.778	0.763	0.778	0.778
RLP_Subfamily_CPAASC_9_SMOTE	AdaBoostClassifier	0.972	0.774	0.759	0.774	0.774
	Mean	0.972	0.774	0.759	0.774	0.774
RLP_Subfamily_CPAASC_N_C_terminal_10_SMOTE	AdaBoostClassifier	0.975	0.800	0.787	0.800	0.800
RLP_Subfamily_CPAASC_N_C_terminal_1_SMOTE	LinearDiscriminantAnalysis	0.976	0.810	0.797	0.810	0.810
RLP_Subfamily_CPAASC_N_C_terminal_2_SMOTE	AdaBoostClassifier	0.975	0.803	0.790	0.803	0.803
RLP_Subfamily_CPAASC_N_C_terminal_3_SMOTE	LinearDiscriminantAnalysis	0.976	0.804	0.792	0.804	0.804
RLP_Subfamily_CPAASC_N_C_terminal_4_SMOTE	LinearDiscriminantAnalysis	0.976	0.805	0.793	0.805	0.805
RLP_Subfamily_CPAASC_N_C_terminal_5_SMOTE	AdaBoostClassifier	0.975	0.802	0.789	0.802	0.802
RLP_Subfamily_CPAASC_N_C_terminal_6_SMOTE	LinearDiscriminantAnalysis	0.976	0.808	0.795	0.808	0.808
RLP_Subfamily_CPAASC_N_C_terminal_7_SMOTE	LinearDiscriminantAnalysis	0.976	0.808	0.795	0.808	0.808
RLP_Subfamily_CPAASC_N_C_terminal_8_SMOTE	AdaBoostClassifier	0.975	0.802	0.789	0.802	0.802
RLP_Subfamily_CPAASC_N_C_terminal_9_SMOTE	LinearDiscriminantAnalysis	0.976	0.805	0.792	0.805	0.805
	Mean	0.976	0.805	0.792	0.805	0.805
RLP_Subfamily_Dipeptide_10_SMOTE	KNeighborsClassifier	0.992	0.935	0.931	0.935	0.935
RLP_Subfamily_Dipeptide_1_SMOTE	KNeighborsClassifier	0.992	0.937	0.933	0.937	0.937
RLP_Subfamily_Dipeptide_2_SMOTE	KNeighborsClassifier	0.992	0.935	0.931	0.935	0.935
RLP_Subfamily_Dipeptide_3_SMOTE	KNeighborsClassifier	0.992	0.934	0.930	0.934	0.934
RLP_Subfamily_Dipeptide_4_SMOTE	KNeighborsClassifier	0.991	0.932	0.927	0.932	0.932
RLP_Subfamily_Dipeptide_5_SMOTE	KNeighborsClassifier	0.992	0.934	0.930	0.934	0.934
RLP_Subfamily_Dipeptide_6_SMOTE	KNeighborsClassifier	0.991	0.931	0.926	0.931	0.931
RLP_Subfamily_Dipeptide_7_SMOTE	KNeighborsClassifier	0.992	0.933	0.929	0.933	0.933
RLP_Subfamily_Dipeptide_8_SMOTE	KNeighborsClassifier	0.991	0.925	0.920	0.925	0.925
RLP_Subfamily_Dipeptide_9_SMOTE	KNeighborsClassifier	0.991	0.929	0.925	0.929	0.929
	Mean	0.992	0.932	0.928	0.932	0.932
RLP_Subfamily_Tripeptide_1_SMOTE	KNeighborsClassifier	0.995	0.957	0.954	0.957	0.957
RLP_Subfamily_Tripeptide_2_SMOTE	KNeighborsClassifier	0.994	0.955	0.952	0.955	0.955
RLP_Subfamily_Tripeptide_3_SMOTE	KNeighborsClassifier	0.994	0.956	0.953	0.956	0.956
RLP_Subfamily_Tripeptide_4_SMOTE	KNeighborsClassifier	0.995	0.958	0.955	0.958	0.958
RLP_Subfamily_Tripeptide_5_SMOTE	KNeighborsClassifier	0.995	0.958	0.955	0.958	0.958
RLP_Subfamily_Tripeptide_6_SMOTE	KNeighborsClassifier	0.994	0.954	0.951	0.954	0.954
RLP_Subfamily_Tripeptide_7_SMOTE	KNeighborsClassifier	0.994	0.955	0.952	0.955	0.955
RLP_Subfamily_Tripeptide_8_SMOTE	KNeighborsClassifier	0.994	0.951	0.948	0.951	0.951
RLP_Subfamily_Tripeptide_9_SMOTE	KNeighborsClassifier	0.995	0.958	0.955	0.958	0.958
RLP_Subfamily_Tripeptide_10_SMOTE	KNeighborsClassifier	0.995	0.959	0.957	0.959	0.959
	Mean	**0.994**	**0.956**	**0.953**	**0.956**	**0.956**

3.3. ML model capacity of distinguishing RLPs from no RLPs

The ability of the ML models to distinguish RLPs from noRLPs was assessed through the predictive capacity of the models created with the RLPs/noRLPs datasets (Fig. 1C). The models that classify RLPs/noRLPs were evaluated using 10-Fold cross-validation based on the following metrics: ACC, sensitivity, precision, F-measure, specificity, FDR, and MCC. For each dataset, 21 models (21 algorithms tested) were selected, and the performance results are presented in Table 2. In general, the selected models provided average values for ACC, F-measure, FDR, MCC, precision, sensitivity and specificity equal to 0.93, 0.934, 0.070, 0.861, 0.948, 0.948 and 0.911, respectively.

3.4. ML model abilities to distinguish RLPs from no RLKs

To distinguish RLPs from RLKs, we assessed the generality of models constructed with RLP/RLK datasets (Fig. 1D). The outcome of 10-fold cross-validations and evaluated metrics for RLPs/RLKs models are shown in the Table 3. The quadratic discriminant analysis and gradient boosting classifier with the amino acid composition of the N-terminus, C-terminus and tripeptide features, respectively, outperformed the other ones (Table 3, highlighted in bold). The average performance of the six models provided ACC, F-measure, FDR, MCC, precision, sensitivity and specificity, equal to 0.968, 0.967, 0.04, 0.936, 0.981, 0.981, 0.955, respectively.

3.5. The ability of ML models to classify RLP subfamilies

To classify RLP subfamily, we evaluated models built with RLP subfamily datasets using 10-fold cross-validations. The performance of the models was given by the previously mentioned metrics (Fig. 1E). The tripeptide and dipeptide composition features achieved average MCC values higher than 0.90 when using K-nearest neighbor algorithm. The amino acid composition N-terminus and C-terminus feature achieved an average MCC value of 0.899 using a calibrated classifier and linear discriminant analysis (Table 4). Average values for the classification models were 0.98, 0.87, 0.877 and 0.874, respectively, for ACC

sensitivity, precision, and F-measure while MCC varied from 0.759 to 0.953 (Table 4, highlighted in bold).

3.6. Validation of RLPredictiOme

For RLPredictiOme validation, we tested the ML models in combination with Bayesian inference as an ensemble method approach (Fig. 1). In the first validation, we submitted 47 near-characterized sequences of RLPs against the RLPredictiOme. The validation data set was composed of 39 LRR receptor-like proteins (LRR-RLP), six LysM-RLPs, two WAK-RLPs, and one Salt stress response/antifungal-RLP (Table 5). However, six of these RLPs were not characterized as RLP as they did not have a transmembrane segment (TM). The test resulted in 37 LRR-RLPs correctly classified, two LysM-RLPs were correctly classified, and two LysM-RLPs were classified as undefined due to relative low probability provided by Bayesian inference of the RLP subfamily. The remaining two LysM-RLPs (Q67UE8.1 LYP4 and Q69T51.1 LYP6), one WAK-RLP (AKP45167), and one Salt stress response/antifungal-RLP (LOC_Os04g56430.1) were not classified as RLPs due to the TM absence (Table 5, highlighted in bold).

Table 5. Validation of the RLPs almost characterized

Accession	SP	TM	RLP-noRLP	RLP-noRLP Probability	RLP-RLK	RLP-RLK Probability	RLP-Subfamily	RLP-Subfamily Probability	Classification	Decision probability
NP_001234733.2	Y	Y	RLP	0.9961	RLP	0.5751	LRR-RLP	0.7666	(LRR-RLP)	0.9894
sQ9LNV9.2_RLP1	Y	Y	RLP	0.9961	RLP	0.7161	LRR-RLP	0.7671	(LRR-RLP)	0.9891
sp|Q93ZH0.1|LYM1	Y	Y	RLP	0.8941	RLP	0.9915	LysM-RLP	0.467	(LysM-RLP)	0.989
CAC40826.1_HcrVf2	Y	Y	RLP	0.9961	RLP	0.9895	LRR-RLP	0.8333	(LRR-RLP)	0.9888
AAA65235.1_Cf-9	Y	Y	RLP	0.9965	RLP	0.9906	LRR-RLP	0.8331	(LRR-RLP)	0.9887
AAC78594.1_Hcr2-2A	Y	Y	RLP	0.9965	RLP	0.8569	LRR-RLP	0.849	(LRR-RLP)	0.9885
Q9SSD1.1	Y	Y	RLP	0.9966	RLP	0.991	LRR-RLP	0.4667	(LRR-RLP)	0.9885
AAC15779.1_Cf-2.1	Y	Y	RLP	0.9965	RLP	0.855	LRR-RLP	0.85	(LRR-RLP)	0.9882
sp|Q7FZR1.1|RLP52	Y	Y	RLP	0.9966	RLP	0.9903	LRR-RLP	0.8336	(LRR-RLP)	0.9882
QED40966.1	Y	Y	RLP	0.9962	RLP	0.7168	LRR-RLP	0.8506	(LRR-RLP)	0.9881
CAC40827.1_HcrVf3	Y	Y	RLP	0.9964	RLP	0.9909	LRR-RLP	0.8501	(LRR-RLP)	0.988
sp|Q9LJS0.1|RLP42	Y	Y	RLP	0.9966	RLP	0.9911	LRR-RLP	0.8502	(LRR-RLP)	0.988
AAC78593.1_Hcr2-0B	Y	Y	RLP	0.9962	RLP	0.991	LRR-RLP	0.8495	(LRR-RLP)	0.9879
Q9FK66.1_RLP55	Y	Y	RLP	0.9958	RLP	0.9915	LRR-RLP	0.6669	(LRR-RLP)	0.9879
sQ9SN38.1_RLP5	Y	Y	RLP	0.9963	RLP	0.9912	LRR-RLP	0.8497	(LRR-RLP)	0.9879
AAC78596.1_Hcr2-5D	Y	Y	RLP	0.9959	RLP	0.9909	LRR-RLP	0.85	(LRR-RLP)	0.9878
BAE95828.1 (LysM)	Y	Y	RLP	0.9964	RLP	0.99	Undefined	0.4169	**(Undefined)**	0.9878
Q9LJS2.1	Y	Y	RLP	0.9964	RLP	0.9906	LRR-RLP	0.8505	(LRR-RLP)	0.9878
AJG42080.1_RLM2	Y	Y	RLP	0.9963	RLP	0.9908	LRR-RLP	0.8493	(LRR-RLP)	0.9877
CAA05269.1_Hcr9-4E	Y	Y	RLP	0.9962	RLP	0.9893	LRR-RLP	0.8332	(LRR-RLP)	0.9877
AJG42091.1_LEPR3	Y	Y	RLP	0.9967	RLP	0.9911	LRR-RLP	0.8508	(LRR-RLP)	0.9875
Q9M2Y3.1_RLP44	Y	Y	RLP	0.9962	RLP	0.9902	LRR-RLP	0.7503	(LRR-RLP)	0.9875
CAC40825.1_HcrVf1	Y	Y	RLP	0.9965	RLP	0.9921	LRR-RLP	0.8166	(LRR-RLP)	0.9874
NP_001234474.2	Y	Y	RLP	0.9963	RLP	0.991	LRR-RLP	0.8332	(LRR-RLP)	0.9874

Solyc08g016270.1.1	Y	Y	RLP	0.9961	RLP	0.72	LRR-RLP	0.6335	(LRR-RLP)	0.9874
AAC78595.1_Hcr2-5B	Y	Y	RLP	0.9963	RLP	0.8517	LRR-RLP	0.85	(LRR-RLP)	0.9873
O80809.1_CLV2	Y	Y	RLP	0.9964	RLP	0.991	LRR-RLP	0.8496	(LRR-RLP)	0.9873
sp|O23006.1|LYM2	Y	Y	RLP	0.9962	RLP	0.9908	Undefined	0.5005	**(Undefined)**	0.9873
sp|O48849.1|RLP23	Y	Y	RLP	0.9959	RLP	0.9906	LRR-RLP	0.7833	(LRR-RLP)	0.9873
AAC78592.1_Hcr2-0A	Y	Y	RLP	0.9966	RLP	0.8518	LRR-RLP	0.8513	(LRR-RLP)	0.9872
sp|Q6NPN4.1|LYM3	Y	Y	RLP	0.9452	RLP	0.99	LysM-RLP	0.4501	(LysM-RLP)	0.9872
AAC78591.1	Y	Y	RLP	0.9966	RLP	0.9899	LRR-RLP	0.8507	(LRR-RLP)	0.9871
AJV90937.1	Y	Y	RLP	0.9968	RLP	0.8507	LRR-RLP	0.8332	(LRR-RLP)	0.9871
AUT14025.1	Y	Y	RLP	0.9962	RLP	0.8537	LRR-RLP	0.7329	(LRR-RLP)	0.987
AAC15780.1_Cf-2.2	Y	Y	RLP	0.9961	RLP	0.8555	LRR-RLP	0.8491	(LRR-RLP)	0.9863
AGI92782.1_RLP1.813	Y	Y	RLP	0.9963	RLP	0.9906	LRR-RLP	0.4005	(LRR-RLP)	0.9862
NP_187187.1	Y	Y	RLP	0.9964	RLP	0.9913	LRR-RLP	0.6497	(LRR-RLP)	0.986
AKR80573.1_I-7	Y	Y	RLP	0.9963	RLP	0.8605	LRR-RLP	0.65	(LRR-RLP)	0.9855
NP_001362850.1_EIX2	Y	Y	RLP	0.9961	RLP	0.8581	LRR-RLP	0.6005	(LRR-RLP)	0.985
sp|Q9SHI4.1|RLP3	N	Y	RLP	0.9965	RLP	0.9904	LRR-RLP	0.8328	(LRR-RLP)	0.8015
NP_001355132.1	N	Y	RLP	0.9965	RLP	0.9903	LRR-RLP	0.5163	(LRR-RLP)	0.8012
Q940E8.1_FEA2	Y	N	RLP	0.9487	RLP	0.8554	LRR-RLP	0.849	noRLP	0.2048
sp|Q67UE8.1|LYP4	Y	N	RLP	0.7894	RLP	0.8564	Undefined	0.0	noRLP	0.2017
AFB75328.1	Y	N	RLP	0.9472	RLP	0.857	LRR-RLP	0.5667	noRLP	0.2012
AKP45167.1	Y	N	RLP	0.9462	RLP	0.8543	Undefined	0.4495	noRLP	0.201
sp|Q69T51.1|LYP6	Y	N	RLP	0.8422	RLP	0.8544	Undefined	0.0	noRLP	0.2007
LOC_Os04g56430.1	Y	N	RLP	0.9471	RLP	0.8518	Salt-stress-response/ antifungal-RLP	0.4334	noRLP	0.1986

In the second validation, we used the data of a genome-wide study of RLPs restricted to the LRR-RLP subfamily (Jamieson et al., 2018). The 57 LRR-RLPs of Arabidopsis were submitted to the RLPredictiOme predictor. As a result, 47 LRR-RLPs were classified correctly, although 13 LRR-RLPs did not account for signal peptide (SP). One LRR-RLP harboring SP was undefined, and the remaining nine LRR-RLPs were not classified as RLP due to the TM absence (Table 6, highlighted in bold). Interestingly, the AtRLP4 protein was previously classified as LRR-RLP; however, the RLPredictiOme classified it as Malectin-RLP due to the presence of the one di-glucose binding domain within the endoplasmic reticulum associated with an LRR domain.

Table 6. Validation of the RLPs of genome-wide study of RLPs restricted to the LRR-RLP subfamily.

Accession	SM	TM	RLP-noRLP	RLP-noRLP Probability	RLP-RLK	RLP-RLK Probability	RLP-Subfamily	RLP-Subfamily Probability	Classification	Decision probability
AT1G65380.1	Y	Y	RLP	0.9962	RLP	0.9907	LRR-RLP	0.8505	(LRR-RLP)	0.9902
AT1G17240.1	Y	Y	RLP	0.9962	RLP	0.9913	LRR-RLP	0.8497	(LRR-RLP)	0.9886
AT4G13880.1	Y	Y	RLP	0.9963	RLP	0.9899	LRR-RLP	0.8001	(LRR-RLP)	0.9884
AT5G27060.1	Y	Y	RLP	0.9962	RLP	0.991	LRR-RLP	0.6669	(LRR-RLP)	0.9884
AT3G23110.1	Y	Y	RLP	0.9964	RLP	0.9912	LRR-RLP	0.6502	(LRR-RLP)	0.9883
AT1G80080.1	Y	Y	RLP	0.9961	RLP	0.9911	LRR-RLP	0.5506	(LRR-RLP)	0.9883
AT2G32680.1	Y	Y	RLP	0.9967	RLP	0.9918	LRR-RLP	0.7838	(LRR-RLP)	0.9882
AT1G74180.1	Y	Y	RLP	0.9959	RLP	0.858	LRR-RLP	0.8163	(LRR-RLP)	0.988
AT3G05370.1	Y	Y	RLP	0.9962	RLP	0.8556	LRR-RLP	0.6337	(LRR-RLP)	0.988
AT3G11080.1	Y	Y	RLP	0.9962	RLP	0.991	LRR-RLP	0.8496	(LRR-RLP)	0.988
AT3G28890.1	Y	Y	RLP	0.9966	RLP	0.8561	LRR-RLP	0.6336	(LRR-RLP)	0.988

AT2G25440.1	Y	Y	RLP	0.9962	RLP	0.9902	LRR-RLP	0.4832	(LRR-RLP)	0.9878
AT5G45770.1	Y	Y	RLP	0.9965	RLP	0.99	LRR-RLP	0.683	(LRR-RLP)	0.9878
AT2G42800.1	Y	Y	RLP	0.9963	RLP	0.9908	LRR-RLP	0.6665	(LRR-RLP)	0.9876
AT3G05360.1	Y	Y	RLP	0.9967	RLP	0.9913	LRR-RLP	0.6668	(LRR-RLP)	0.9876
AT5G65830.1	Y	Y	RLP	0.9966	RLP	0.8566	LRR-RLP	0.667	(LRR-RLP)	0.9876
AT1G28340.1	Y	Y	RLP	0.8425	RLP	0.9905	**Malectin-RLP**	0.4502	**(Malectin-RLP)**	0.9875
AT1G74190.1	Y	Y	RLP	0.9959	RLP	0.8564	LRR-RLP	0.8499	(LRR-RLP)	0.9871
AT2G15080.1	Y	Y	RLP	0.9965	RLP	0.9904	LRR-RLP	0.8502	(LRR-RLP)	0.987
AT3G05650.1	Y	Y	RLP	0.9964	RLP	0.9906	LRR-RLP	0.6664	(LRR-RLP)	0.9868
AT1G45616.1	Y	Y	RLP	0.9961	RLP	0.9913	LRR-RLP	0.7665	(LRR-RLP)	0.9868
AT3G05660.1	Y	Y	RLP	0.9966	RLP	0.8557	LRR-RLP	0.85	(LRR-RLP)	0.9866
AT1G58190.1	Y	Y	RLP	0.9962	RLP	0.8521	LRR-RLP	0.6663	(LRR-RLP)	0.9866
AT3G49750.1	Y	Y	RLP	0.9963	RLP	0.9909	LRR-RLP	0.7502	(LRR-RLP)	0.9865
AT4G13920.1	Y	Y	RLP	0.9967	RLP	0.9911	LRR-RLP	0.8498	(LRR-RLP)	0.9865
AT5G25910.1	Y	Y	RLP	0.9964	RLP	0.9899	LRR-RLP	0.8501	(LRR-RLP)	0.9864
AT2G33060.1	Y	Y	RLP	0.9966	RLP	0.9914	LRR-RLP	0.8332	(LRR-RLP)	0.9863
AT4G04220.1	Y	Y	RLP	0.9962	RLP	0.9911	LRR-RLP	0.8506	(LRR-RLP)	0.9863
AT2G33050.1	Y	Y	RLP	0.9964	RLP	0.9915	LRR-RLP	0.7498	(LRR-RLP)	0.986
AT1G71400.1	Y	Y	RLP	0.996	RLP	0.8563	LRR-RLP	0.6831	(LRR-RLP)	0.9851
AT4G18760.1	Y	Y	RLP	0,9967	RLP	0,9903	LRR-RLP	0,8495	(LRR-RLP)	0,9885
AT1G71390.1	N	Y	RLP	0.9966	RLP	0.99	LRR-RLP	0.6667	(LRR-RLP)	0.8021
AT2G25470.1	N	Y	RLP	0.9964	RLP	0.8556	LRR-RLP	0.8502	(LRR-RLP)	0.8014
AT1G47890.1	N	Y	RLP	0.9967	RLP	0.9908	LRR-RLP	0.8501	(LRR-RLP)	0.8001
AT4G13810.1	N	Y	RLP	0.9964	RLP	0.9907	LRR-RLP	0.833	(LRR-RLP)	0.7997
AT3G23010.1	N	Y	RLP	0.9965	RLP	0.9908	LRR-RLP	0.667	(LRR-RLP)	0.7995
AT1G74170.1	N	Y	RLP	0.9964	RLP	0.8561	LRR-RLP	0.7164	(LRR-RLP)	0.7994
AT3G24982.1	N	Y	RLP	0.9963	RLP	0.989	LRR-RLP	0.8512	(LRR-RLP)	0.7993
AT1G17250.1	N	Y	RLP	0.9965	RLP	0.9911	LRR-RLP	0.8496	(LRR-RLP)	0.799
AT3G23120.1	N	Y	RLP	0.997	RLP	0.9905	LRR-RLP	0.6835	(LRR-RLP)	0.7976
AT3G53240.1	N	Y	RLP	0.9961	RLP	0.9905	LRR-RLP	0.783	(LRR-RLP)	0.7973
AT1G07390.1	N	Y	RLP	0.9957	RLP	0.7119	LRR-RLP	0.7826	(LRR-RLP)	0.7969
AT3G11010.1	N	Y	RLP	0.9961	RLP	0.9902	LRR-RLP	0.6665	(LRR-RLP)	0.7958
AT1G34290.1	Y	Y	RLP	0.9964	RLP	0.9898	**Undefined**	0.2166	**(Undefined)**	0.7949
AT5G49290.1	N	Y	RLP	0.9966	RLP	0.9901	LRR-RLP	0.6833	(LRR-RLP)	0.7941
AT2G32660		N								
AT2G33020		N								
AT2G33030		N								
AT2G33080		N								
AT3G24900		N								
AT3G25010		N								
AT4G13900		N								
AT5G40170		N								
AT3G25020		N								

In a third validation, we selected 148 LRR-RLPs described in a genome-wide study of rice RLPs (Fritz-Laylin et al., 2005) (Table S1). The results show that 78 LRR-RLPs with SP and TM were correctly classified with relatively high probability (greater than 0.98). Additionally, from 73 LRR-RLPs with a single TM, 71 were correctly classified, whereas two of them were classified as Other-RLPs with estimated probability ranging from 0.792 to 0.805. Only four predicted LRR-RLPs from rice were classified as noRLPs; two of them lack both SP and TM; and two do not harbor TM.

The fourth validation was carried out to ensure that RLPredictiOme does not randomly classify proteins. For this, 100 randomly generated sequences were confronted against RLPredictiOme, and all sequences were classified as noRLP in the first step (Table 7).

Table 7. Random sequence confronted against RLPredictiOme.

Accession	SM	TM	RLP-noRLP	RLP-noRLP Probability	RLP-RLK	RLP-RLK Probability	RLP-Subfamily	RLP-Subfamily Probability	Classification	Decision probability
Alien_71_464	Y	Y	noRLP	0.0532	RLP	0.7145	Other-RLP	0.4166	noRLP	0.4033
Alien_78_801	Y	Y	noRLP	0.0532	RLP	0.857	WAK-RLP	0.3169	noRLP	0.4014
Alien_88_471	N	Y	noRLP	0.369	RLP	0.855	Unknown	0.2837	noRLP	0.2068
Alien_90_956	N	Y	noRLP	0.0527	RLK-like	0.5721	Other-RLP	0.3499	noRLP	0.2064
Alien_94_666	N	Y	noRLP	0.0535	RLP	0.8558	S-domain-RLP	0.3164	noRLP	0.2045
Alien_11_789	N	Y	noRLP	0.0524	RLK-like	0.4288	Other-RLP	0.4331	noRLP	0.2034
Alien_34_248	N	Y	noRLP	0.2093	RLP	0.8571	Other-RLP	0.4004	noRLP	0.2022
Alien_70_660	N	Y	noRLP	0.3677	RLP	0.8564	Unknown	0.2491	noRLP	0.2002
Alien_59_959	N	Y	noRLP	0.052	RLK-like	0.576	S-domain-RLP	0.417	noRLP	0.1994
Alien_20_195	Y	N	noRLP	0.3704	RLP	0.8544	Unknown	0.2671	noRLP	0.1987
Alien_23_503	N	Y	noRLP	0.3698	RLP	0.8596	Unknown	0.3	noRLP	0.1987
Alien_69_854	N	Y	noRLP	0.0542	RLP	0.7198	Other-RLP	0.4327	noRLP	0.1985
Alien_2_750	N	Y	noRLP	0.0526	RLK-like	0.5768	Other-RLP	0.3331	noRLP	0.1956
Alien_66_528	N	N	noRLP	0.0001	RLP	0.8549	S-domain-RLP	0.3829	noRLP	0.0195
Alien_1_268	N	N	noRLP	0.0002	RLP	0.8536	Other-RLP	0.3831	noRLP	0.0093
Alien_51_917	N	N	noRLP	0.0002	RLK-like	0.573	Unknown	0.283	noRLP	0.0044
Alien_79_429	N	N	noRLP	0.3166	RLP	0.8588	Other-RLP	0.3001	noRLP	0.0041
Alien_61_779	N	N	noRLP	0.0002	RLP	0.7131	S-domain-RLP	0.3834	noRLP	0.0036
Alien_67_112	N	N	noRLP	0.1591	RLP	0.7131	Other-RLP	0.3342	noRLP	0.0035
Alien_42_363	N	N	noRLP	0.316	RLP	0.8576	S-domain-RLP	0.3336	noRLP	0.003
Alien_4_417	N	N	noRLP	0.0002	RLK-like	0.5712	WAK-RLP	0.4337	noRLP	0.0029
Alien_24_102	N	N	noRLP	0.4222	RLP	0.861	WAK-RLP	0.3498	noRLP	0.0027
Alien_9_882	N	N	noRLP	0.0002	RLP	0.7132	S-domain-RLP	0.3664	noRLP	0.0019
Alien_7_199	N	N	noRLP	0.3166	RLP	0.8564	WAK-RLP	0.3504	noRLP	0.0018
Alien_29_460	N	N	noRLP	0.2089	RLP	0.8554	Unknown	0.284	noRLP	0.0017
Alien_50_474	N	N	noRLP	0.0009	RLP	0.8548	Unknown	0.2495	noRLP	0.0017
Alien_72_442	N	N	noRLP	0.0002	RLP	0.8498	Unknown	0.2333	noRLP	0.0017
Alien_97_120	N	N	noRLP	0.3685	RLP	0.8566	Unknown	0.2999	noRLP	0.0017
Alien_38_893	N	N	noRLP	0.0003	RLK-like	0.5771	S-domain-RLP	0.4499	noRLP	0.0016
Alien_73_528	N	N	noRLP	0.0002	RLP	0.857	S-domain-RLP	0.3665	noRLP	0.0016
Alien_83_641	N	N	noRLP	0.0003	RLP	0.7085	Other-RLP	0.3502	noRLP	0.0016
Alien_44_248	N	N	noRLP	0.0003	RLP	0.7133	S-domain-RLP	0.3833	noRLP	0.0015
Alien_62_945	N	N	noRLP	0.0002	RLK-like	0.5733	S-domain-RLP	0.4834	noRLP	0.0015
Alien_16_855	N	N	noRLP	0.0002	RLK-like	0.4308	Unknown	0.2658	noRLP	0.0014
Alien_40_703	N	N	noRLP	0.0002	RLP	0.711	S-domain-RLP	0.3499	noRLP	0.0014
Alien_45_534	N	N	noRLP	0.0002	RLP	0.8553	WAK-RLP	0.3165	noRLP	0.0014
Alien_74_665	N	N	noRLP	0.0001	RLP	0.8547	Unknown	0.2503	noRLP	0.0014
Alien_18_925	N	N	noRLP	0.0001	RLK-like	0.5679	Other-RLP	0.4166	noRLP	0.0013
Alien_33_955	N	N	noRLP	0.0003	RLK-like	0.4348	Unknown	0.2332	noRLP	0.0013
Alien_39_171	N	N	noRLP	0.1577	RLP	0.8516	Unknown	0.2665	noRLP	0.0012
Alien_49_350	N	N	noRLP	0.0002	RLP	0.8573	S-domain-RLP	0.4842	noRLP	0.0012

Alien_63_622	N	N	noRLP	0.0002	RLP	0.8555	Unknown	0.2664	noRLP	0.0012
Alien_89_627	N	N	noRLP	0.0002	RLP	0.8567	Other-RLP	0.3835	noRLP	0.0012
Alien_91_929	N	N	noRLP	0.0003	RLK-like	0.573	Other-RLP	0.4331	noRLP	0.0012
Alien_14_450	N	N	noRLP	0.3148	RLP	0.7157	WAK-RLP	0.333	noRLP	0.0011
Alien_15_536	N	N	noRLP	0.0007	RLP	0.8566	Unknown	0.2668	noRLP	0.0011
Alien_22_586	N	N	noRLP	0.001	RLP	0.8562	S-domain-RLP	0.3993	noRLP	0.0011
Alien_3_226	N	N	noRLP	0.0003	RLK-like	0.431	Unknown	0.2991	noRLP	0.0011
Alien_57_326	N	N	noRLP	0.3151	RLP	0.8605	Unknown	0.2502	noRLP	0.0011
Alien_13_137	N	N	noRLP	0.2113	RLK-like	0.5764	Unknown	0.1667	noRLP	0.001
Alien_35_659	N	N	noRLP	0.0002	RLK-like	0.5687	Other-RLP	0.3829	noRLP	0.001
Alien_37_440	N	N	noRLP	0.0003	RLK-like	0.5743	Unknown	0.2666	noRLP	0.001
Alien_48_571	N	N	noRLP	0.0002	RLP	0.8586	Unknown	0.2999	noRLP	0.001
Alien_54_839	N	N	noRLP	0.0004	RLP	0.7158	Unknown	0.2674	noRLP	0.001
Alien_12_553	N	N	noRLP	0.3185	RLP	0.858	Unknown	0.2335	noRLP	0.0009
Alien_17_304	N	N	noRLP	0.3169	RLP	0.8541	Unknown	0.2828	noRLP	0.0009
Alien_25_176	N	N	noRLP	0.0003	RLP	0.8568	Unknown	0.2667	noRLP	0.0009
Alien_30_623	N	N	noRLP	0.0002	RLP	0.8547	Other-RLP	0.3833	noRLP	0.0009
Alien_32_240	N	N	noRLP	0.1576	RLP	0.8531	Unknown	0.2499	noRLP	0.0009
Alien_53_589	N	N	noRLP	0.0006	RLP	0.7103	Unknown	0.3	noRLP	0.0009
Alien_58_715	N	N	noRLP	0.0001	RLK-like	0.5748	S-domain-RLP	0.3842	noRLP	0.0009
Alien_82_456	N	N	noRLP	0.0001	RLP	0.855	S-domain-RLP	0.3165	noRLP	0.0009
Alien_85_415	N	N	noRLP	0.0004	RLP	0.715	Unknown	0.2167	noRLP	0.0009
Alien_8_947	N	N	noRLP	0.0001	RLK-like	0.5689	Unknown	0.25	noRLP	0.0009
Alien_10_555	N	N	noRLP	0.0002	RLP	0.8536	Unknown	0.2996	noRLP	0.0008
Alien_19_229	N	N	noRLP	0.0003	RLP	0.8599	PAN-RLP	0.3336	noRLP	0.0008
Alien_27_824	N	N	noRLP	0.0002	RLP	0.7111	Unknown	0.3337	noRLP	0.0008
Alien_41_731	N	N	noRLP	0.0004	RLP	0.7117	Unknown	0.2666	noRLP	0.0008
Alien_43_686	N	N	noRLP	0.0001	RLP	0.7129	S-domain-RLP	0.3662	noRLP	0.0008
Alien_47_420	N	N	noRLP	0.0004	RLP	0.8546	Other-RLP	0.4172	noRLP	0.0008
Alien_52_779	N	N	noRLP	0.0003	RLK-like	0.4383	Unknown	0.2999	noRLP	0.0008
Alien_55_478	N	N	noRLP	0.0002	RLP	0.7179	Other-RLP	0.3997	noRLP	0.0008
Alien_60_817	N	N	noRLP	0.0002	RLP	0.7135	Unknown	0.2999	noRLP	0.0008
Alien_64_626	N	N	noRLP	0.0002	RLP	0.7138	Other-RLP	0.4	noRLP	0.0008
Alien_75_673	N	N	noRLP	0.0002	RLP	0.8548	Unknown	0.2832	noRLP	0.0008
Alien_81_442	N	N	noRLP	0.0003	RLK-like	0.5736	S-domain-RLP	0.4833	noRLP	0.0008
Alien_87_495	N	N	noRLP	0.0005	RLP	0.8555	S-domain-RLP	0.3838	noRLP	0.0008
Alien_93_110	N	N	noRLP	0.3149	RLP	0.8597	WAK-RLP	0.467	noRLP	0.0008
Alien_99_622	N	N	noRLP	0.0002	RLP	0.8568	Unknown	0.25	noRLP	0.0008
Alien_21_499	N	N	noRLP	0.0002	RLP	0.86	S-domain-RLP	0.3498	noRLP	0.0007
Alien_31_429	N	N	noRLP	0.0002	RLP	0.7128	Unknown	0.2996	noRLP	0.0007
Alien_46_860	N	N	noRLP	0.0002	RLK-like	0.571	Unknown	0.2995	noRLP	0.0007
Alien_56_859	N	N	noRLP	0.0005	RLK-like	0.5724	S-domain-RLP	0.3328	noRLP	0.0007
Alien_5_855	N	N	noRLP	0.0003	RLK-like	0.572	Unknown	0.2997	noRLP	0.0007
Alien_65_609	N	N	noRLP	0.0002	RLK-like	0.4257	Unknown	0.2667	noRLP	0.0007
Alien_6_529	N	N	noRLP	0.0001	RLP	0.8565	Unknown	0.2504	noRLP	0.0007
Alien_86_232	N	N	noRLP	0.1581	RLP	0.8535	Other-RLP	0.3495	noRLP	0.0007
Alien_92_960	N	N	noRLP	0.0005	RLK-like	0.5741	Other-RLP	0.3168	noRLP	0.0007
Alien_95_597	N	N	noRLP	0.157	RLP	0.8588	Unknown	0.2833	noRLP	0.0007
Alien_96_597	N	N	noRLP	0.3704	RLP	0.8544	WAK-RLP	0.3999	noRLP	0.0007
Alien_0_119	N	N	noRLP	0.0528	RLP	0.7163	PAN-RLP	0.4339	noRLP	0.0006
Alien_26_112	N	N	noRLP	0.5285	RLP	0.8585	Unknown	0.2664	noRLP	0.0006

Alien_76_327	N	N	noRLP	0.0003	RLP	0.7066	Other-RLP	0.4002	noRLP	0.0006
Alien_77_685	N	N	noRLP	0.0002	RLK-like	0.569	Unknown	0.2494	noRLP	0.0006
Alien_98_323	N	N	noRLP	0.1046	RLP	0.7172	Other-RLP	0.5328	noRLP	0.0006
Alien_28_468	N	N	noRLP	0.0001	RLP	0.8563	Unknown	0.2831	noRLP	0.0005
Alien_36_821	N	N	noRLP	0.0001	RLP	0.717	Unknown	0.2337	noRLP	0.0005
Alien_68_626	N	N	noRLP	0.0002	RLP	0.8541	Unknown	0.2835	noRLP	0.0005
Alien_80_637	N	N	noRLP	0.0002	RLK-like	0.5715	S-domain-RLP	0.4333	noRLP	0.0005
Alien_84_494	N	N	noRLP	0.1614	RLP	0.8574	S-domain-RLP	0.3501	noRLP	0.0005

3.7. High throughput prediction of RLPs in the Arabidopsis genome using RLPredictiOme

We performed high throughput prediction submitting the Arabidopsis sequences against RLPredictiOme. The cutoff tuning for the probability filter was assumed as 0.6 in the first two-step; and as 0.7 in the final step (Fig. 1F). In the third step, the probability estimates were more flexible to predict the RLP subfamilies.

From this genome-wide prediction, RLPredictiOme classified 176 RLP sequences into 15 subfamilies (Table S2). Table 8 summarizes the correct predictions within the subfamily, which are highlighted in black bold. The number of proteins with unknown functions are highlighted in red, whereas the blue description represents the RLPs subfamilies predicted into other subfamilies. The LRR-RLPs subfamily contained 49 members. Three new members (AT5G37360, AT5G19230, and AT4G28560) predicted with relatively high probability were not classified into a known subfamily, whereas two sequences were incorrectly classified. Interestingly, AtRLP4 has two different domains, one LRR and one Di-glucose binding within the endoplasmic reticulum, which characterizes malectin proteins. In summary, it is relevant to a report that the proposed RLPredictiOme method classified the AtRLP4 into Malectin-RLP subfamily (see Table S2).

Table 8. Number of RLPs and RLKs predicted.

Class	RLP	in bold	in red	in blue	in black (mistake)	RLKs in Arabidopsis
LRR-RLP	49	46	3	0	2	235
L-Lectin-RLP	5	0	5		5	45
Salt stress response/antifungal-RLP	9	3	1	5	0	44
WAK-RLP	6	5	1		4	42
S-domain-RLP	1	1			1	37
Unknown-RLP (Extensin,PERK,RKF3,URKI)	43	43			11	28
Malectin-RLP	6	2	3	1	5	15
RCC1-RLP	4		4			8
LysM-RLP	4	2	2			3
B-lectin-RLP	1			1		2
C-Lectin-RLP	0					2

Ethylene-responsive-RLP		3	3			3	2
PAS-RLP		0					2
Thaumatin-RLP		6	6				2
PPR-RLP		0					1
Glycosyl-hydrolases-RLP		3		3			0
PAN-RLP		1		1		1	0
Other-RLP		35	11	24		13	0
Undefined		78					
	Total	176	122	47	7	45	468

The candidate sequences with a legume lectin domain were classified into two RLP subfamilies, B-Lectin-RLP, and L-Lectin-RLP (Table S2). Only one member was classified as B-Lectin-RLP with an unknown function, whereas six members were classified into the L-Lectin-RLP subfamily, also designated as unknown function proteins. Seven proteins were classified incorrectly into this subfamily. The 20 Lysin motif-containing candidate proteins were classified as LysM-RLP (Table S2). Two (AT1G77630.1 and AT2G17120.1) of the three previously characterized LysM-RLPs (Buendia et al., 2018), two classified LysM-RLPs (AT3G06360.1 and AT5G26270.1 belong to subfamilies previously identified as unknow function subfamilies and one sequence (AT1G63550.1) belongs to family Salt stress response/antifungal-RLP. The other 15 sequences may belong to the lipid transfer protein family, not yet characterized. Additionally, the ectodamain lipid transfer family associated with a kinase domain was allocated in the Other-RLP group as probable lipid transfer-RLK. A total of 12 sequences were classified as probable lipid transfer-RLP; however, this misclassification occurred in the LysM-RLP and Unknown-RLP groups, which can be functionally similar to the exact role of these proteins. It may be due to the over representability of these two mentioned groups.

In the Malectin-RLP subfamily, RLPredictiOme correctly classified two members previously characterized (AT1G28340.1 and AT1G24485.1, reference). Four candidate members were identified into subfamilies of unknown function, and seven sequences were incorrectly predicted (Table S2). Furthermore, the third previously identified Malectin-RLP (AT3G46240.1) was predicted as an RCC1-RLP. This subfamily has seven predicted members without known function. One Salt stress response/antifungal-RLP was predicted within this family. The Salt stress response/antifungal-RLP had four members correctly classified, and four predicted within other subfamilies (three in WAK-RLP and one in RCC1-RLP). The S-domain-RLP had one correctly, and one incorrectly predicted sequence (Table S2).

As for the Thaumatin-RLP subfamily, all six members were correctly predicted (Table S2), whereas the WAK-RLP subfamily predicted correctly five members, but also incorporated one candidate sequence with unknown function and three Salt stress response/antifungal-RLP. Ectodomains without a functional domain were classified within a subfamily designated Unknown-RLPs. This group also includes RLPs harboring the ectodomais PERK-like, extensin, RKF3-like, CrRLK1, and RLK10-like, which are proline-rich proteins. RLPredictiOme predicted 46 sequences with unknown function classified as Unknown-RLP subfamily (Table S2). The protein sequences that are not classified correctly or that have a low relative probability of subfamily classification are designated as undefined and are not considered RLPs. In summary, a total of 78 proteins were classified in this group (Table S2).

RLPredictiOme identified probable lipid transfer-RLPs, which were considered as a novel RLP class associated with RLKs, yet to be characterized. Furthermore, three new classes of RLPs were predicted: Plastocyanin-like-RLP, ring finger-RLP and glycosyl-hydrolase-RLP, where contained eight, five and seven members, respectively. Interestingly, five members of the glycerophosphoryl diester phosphodiesterase family (GDPDL) were predicted as Others-RLPs. As a rare protein family in plants, we selected GDPDL-RLP to carry out an experimental validation for these receptor-like proteins candidates. The number of predicted RLPs in each subfamily is shown in Table 8.

3.8. GDPDL family downstream analysis

Phylogenetic analysis of the kinase domain of the RLKs family, and the kinase domain of IRE1A and IRE1B, endoplasmic reticulum-specific protein kinase, clustered the kinase domain of GDPDL-RLK and Thaumatin in the same group distinct from the ER kinases (Figure 2A). These results suggest that GDPDL-RLKs are not ER transmembrane proteins. The secondary structure and the topology of GDPDL show that the N-terminal region of GDPDL-RLK is composed of a signal peptide, a GDPD domain, and more than 10 candidate sites for N-glycosylation (Fig. 2B). As an RLK, GDPDL-RLK contains an ectodomain facing the extracellular space, a transmembrane segment, and a cytoplasmic portion harboring the kinase domain. The topology of classified GDPDL-RLPs fits a typical RLP configuration with N-terminal peptide signal, the glycerophosphoryl diester phosphodiesterase ectodomain,

the transmembrane segment but it lacks a short C-terminal cytoplasmic domain. GDPDL1 and GDPDl6 harbor two glycerophosphoryl diester phosphodiesterase domain, whereas, GDPDL3/4/5 has a single domain localized in a similar position compared with GDPDL-RLK. expression profile of the GDPDL-RLPs in response to pathogens. (D) The expression profile of the GDPDL-RLPs in different organs and developmental stages.

The molecular evolution of the new GDPDLs and the ectodomain of GDPDL-RLK were investigated by calculating the ratio between non-synonymous and synonymous substitutions (Ka/Ks). When compared to the full-length sequence of GDPDL-RLK, only the gene pair GDPDL-RLK/GDPDL6 with a ratio of Ka/Ks > 1 may have undergone a positive selection (Table 9). The ectodomain sequence of GDPDL-RLK compared with gene pairs GDPL1/3/4 was submitted to purifying selection, as suggested by their Ka/Ks ratio < 1 and p-value < 0.05. The divergence time GDPL1/3/4 was of ~ 23.7, 32.5, and 120.1 Mya. These results suggest that, despite the divergence time of GDPL1/3/4 compared to the GDPDL-RLK ectodomain, they led to synonymous mutations that possibly maintained the functional characteristics of GDPL1/3/4 with ectodomain GDPDL-RLK.

Table 9. Evolution molecular analysis of the GDPDLs.

Sequence	Ka	Ks	Ka/Ks	Selection	Date (Mya)	P-Value
GDPDL5-GDPDL3	0.382	1.578	0.242	Purifying	129.316	7.98E-49
GDPD (ectodomain)-GDPDL4	0.214	1.466	0.146	Purifying	120.193	2.22E-45
GDPDL4-GDPD-RLK	0.214	1.288	0.166	Purifying	105.602	9.31E-45
GDPDL1-GDPDL4	0.180	0.940	0.192	Purifying	77.037	1.60E-51
GDPDL3-GDPDL4	0.164	0.852	0.192	Purifying	69.822	1.12E-46
GDPDL4-GDPDL6	0.646	0.802	0.805	Purifying	65.744	0.146094
GDPD-RLK-GDPDL6	0.695	0.638	1.090	Positive	52.286	0.109708
GDPD (ectodomain)-GDPDL3	0.170	0.397	0.428	Purifying	32.525	4.56E-13
GDPDL3-GDPD-RLK	0.167	0.394	0.423	Purifying	32.333	3.06E-13
GDPD-RLK-GDPDL3	0.167	0.394	0.423	Purifying	32.333	3.06E-13
GDPDL1-GDPDL3	0.141	0.390	0.363	Purifying	31.961	1.05E-17
GDPDL1-GDPD-RLK	0.120	0.327	0.368	Purifying	26.786	5.38E-16
GDPD-RLK-GDPDL1	0.120	0.327	0.368	Purifying	26.786	5.38E-16
GDPDL1-GDPD (ectodomain)	0.125	0.326	0.384	Purifying	26.730	5.08E-15

3.9. Identification of GDPDLs- and SNC4-interacting proteins from Arabidopsis

Protein-protein interactions between the GDPDLs and GDPDL-RLK, also designated SUPRESSOR OF NPR1, CONSITUTIVE 4 (SNC4). and the Arabidopsis proteins were identified *in silico* through the protein-protein interactome using several database (BioGRID database, Arabidopsis interactome database, and the String database) utilizing Cytoscape

software. This procedure identified the PPI network containing GDPDLs and directly interacting Arabidopsis proteins (Fig. 3). The GDPDL6 formed the largest hub (degree 38). Among the GDLDL6-interacting proteins, the glycogen synthase kinase 3/SHAGGY-like kinases (GSKs-AT1G57870) may represent a candidate protein for signaling (Fig. 3A, Table 10). Although GSKs have recently been discovered in plants, evidence suggests that they are involved in different biological processes such as brassinosteroid signaling, flower development, and injury responses (Jonak & Hirt, 2002). The node-hub GDPDL5 contains the AtMLP328 pathogenesis-related protein and other proteins of unknown function (Fig. 3A, Table 10). The AtMLP328 is a member of the major latex protein-like (MLPL) gene family responsible for promoting vegetative growth and delaying flowering.

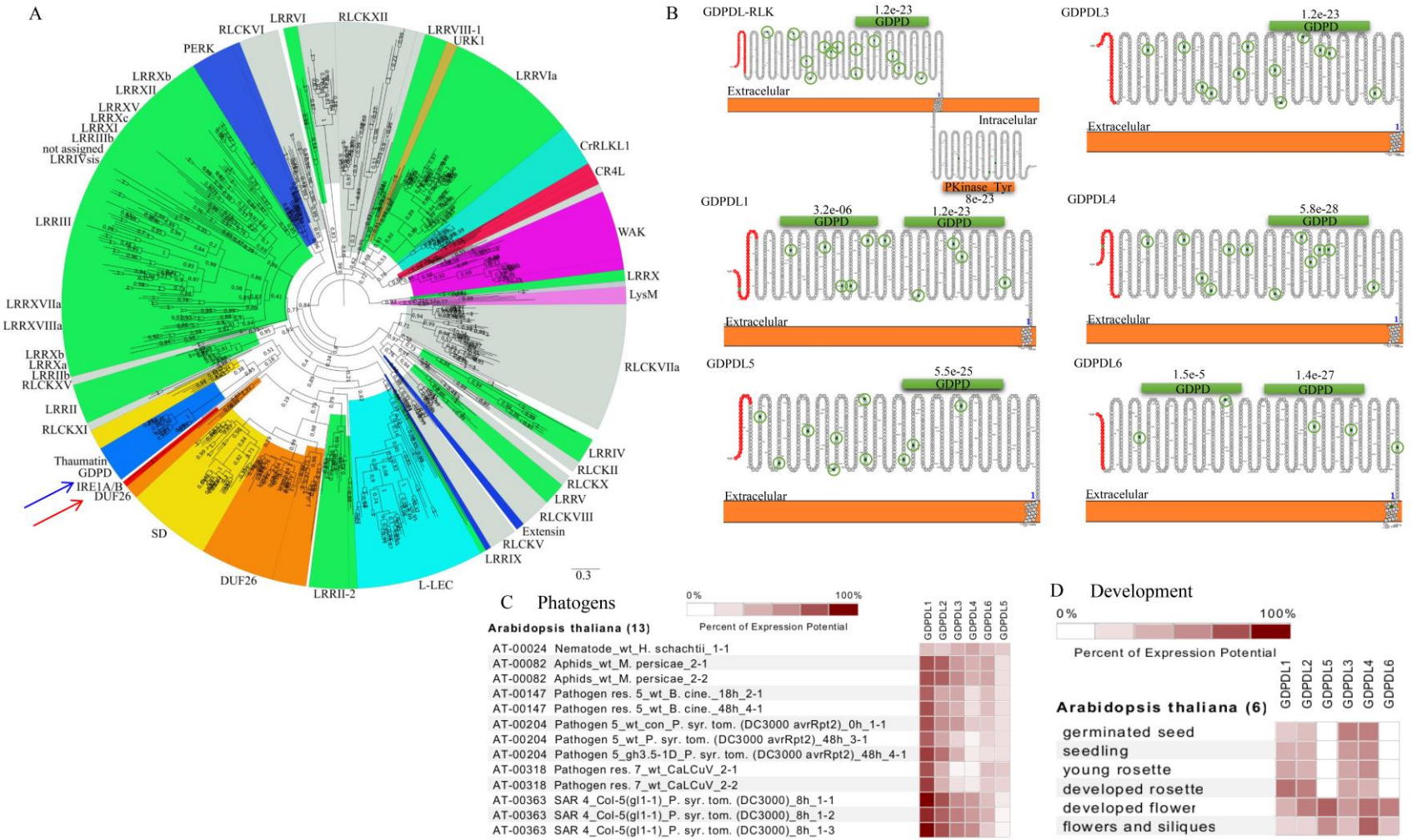

Fig. 2. Analysis in silic of the GDPDL-RLPs. (A) Phylogenetic tree of kinase catalytic domain. (B) The topology of GDPDL-RLP. (C) The expression profile of the GDPDL-RLPs in pathogens response. (D) The expression profile of the GDPDL-RLPs in development.

Table 10. Protein-protein interactions between the GDPDLs proteins and the Arabidopsis proteins. The colors indicated the hubs presented in Figure 3A.

Name	Betweenness Centrality	Closeness Centrality	Degree	Eccentricity	Description
SNC4	0.19234075	0.37614679	12	3	glycerophosphoryl diester phosphodiesterase family protein, putative, expressed
RLP51	0.0	0.27516779	2	4	leucine rich repeat family protein, putative, expressed
SNC1	3.0111E-4	0.27702703	4	4	rp3 protein, putative, expressed
SUA	1.0037E-4	0.27702703	4	4	RNA recognition motif family protein, expressed
DRT111	1.0037E-4	0.27702703	4	4	G-patch domain containing protein, expressed
AT2G20050	0.0	0.27424749	1	4	AGC_PKA/PKG_like.1 - ACG kinases include homologs to PKA, PKG and PKC, expressed
AT1G59780	0.0	0.27424749	1	4	NBS-LRR disease resistance protein, putative, expressed
AT3G55350	0.0	0.27609428	3	4	trp repressor/replication initiator, putative, expressed
BPA1	0.30818366	0.51898734	6	2	RNA recognition motif containing protein, putative, expressed
AT4G17720	0.30818366	0.51898734	6	2	RNA recognition motif, putative, expressed
AT1G22920	0.0	0.27424749	2	4	COP9 signalosome complex subunit 5b, putative, expressed
GDPDL5	0.17835276	0.37104072	10	3	glycerophosphoryl diester phosphodiesterase family protein, putative, expressed
MLP328	0.0	0.27702703	7	4	pathogenesis-related Bet v I family protein, putative, expressed
AGL46	0.0	0.27702703	7	4	OsMADS89 - MADS-box family gene with M-gamma type-box, expressed
AT2G47115	0,04302	0.2779661	8	4	expressed protein
AT1G29660	0,04302	0.2779661	8	4	GDSL-like lipase/acylhydrolase, putative, expressed
AT5G51950	0,04302	0.2779661	8	4	HOTHEAD precursor, putative, expressed
AT1G20680	0,04302	0.2779661	8	4	Ser/Thr-rich protein T10 in DGCR region, putative, expressed
AT2G17710	0,04302	0.2779661	8	4	expressed protein
AT5G42530	0,04302	0.2779661	8	4	
BPA1	0.30818366	0.51898734	6	2	RNA recognition motif containing protein, putative, expressed
AT4G17720	0.30818366	0.51898734	6	2	RNA recognition motif, putative, expressed
GDPDL3	0.1693342	0.37104072	10	3	glycerophosphoryl diester phosphodiesterase family protein, putative, expressed
SHV2	0.0	0.27516779	5	4	COBRA-like protein 7 precursor, putative, expressed
MRH1	0.0	0.27516779	5	4	MRH1, putative, expressed
BST1	0.0	0.27516779	5	4	endonuclease/exonuclease/phosphatase family domain containing protein, expressed
MRH6	0.0	0.27516779	5	4	universal stress protein domain containing protein, putative, expressed
MRH2	0.0	0.27516779	5	4	kinesin motor domain containing protein, expressed
ATCOAE	0.0	0.27152318	1	4	dephospho-CoA kinase, putative, expressed
AT3G23750	0.0	0.27152318	1	4	receptor protein kinase TMK1 precursor, putative, expressed
BPA1	0.30818366	0.51898734	6	2	RNA recognition motif containing protein, putative, expressed
AT4G17720	0.30818366	0.51898734	6	2	RNA recognition motif, putative, expressed
GDPDL1	0.12794717	0.37442922	10	3	glycerophosphoryl diester phosphodiesterase family protein, putative, expressed
AT1G49750	0.0	0.27333333	1	4	uncharacterized protein At4g06744 precursor, putative, expressed
AT3G45710	0.0	0.27333333	1	4	peptide transporter PTR2, putative, expressed
PLDGAMMA1	0.00779455	0.29181495	3	4	phospholipase D, putative, expressed
MAP18	0.0	0.27333333	1	4	Unknown function
CDS1	0.0	0.28275862	2	4	phosphatidate cytidylyltransferase, putative, expressed
BPA1	0.30818366	0.51898734	6	2	RNA recognition motif containing protein, putative, expressed
AT4G17720	0.30818366	0.51898734	6	2	RNA recognition motif, putative, expressed
GDPDL4	0.21573054	0.38497653	14	3	glycerophosphoryl diester phosphodiesterase family protein, putative, expressed
AT5G38480	0.0	0.27891156	1	4	14-3-3 protein, putative, expressed
FLA7	0.00445805	0.29390681	6	4	fasciclin domain containing protein, expressed
SKU5	0.0	0.2877193	4	4	monocopper oxidase, putative, expressed
FLA8	0.0	0.2877193	4	4	fasciclin-like arabinogalactan protein, putative, expressed
ZW9	0.00445805	0.29390681	6	4	ubiquitin carboxyl-terminal hydrolase, putative, expressed
AT1G32860	0.00853443	0.29496403	2	4	glycosyl hydrolases family 17, putative, expressed
AT3G56370	0.0	0.27891156	1	4	receptor-like protein kinase precursor, putative, expressed
AT4G09000	0.0	0.27891156	1	4	14-3-3 protein, putative, expressed
BG_PPAP	0.0	0.27891156	1	4	glycosyl hydrolases family 17, putative, expressed
AT1G01080	0.06480132	0.39047619	3	4	RNA recognition motif containing protein, putative, expressed
AT5G65430	0.0	0.27891156	1	4	14-3-3 protein, putative, expressed
BPA1	0.30818366	0.51898734	6	2	RNA recognition motif containing protein, putative, expressed
AT4G17720	0.30818366	0.51898734	6	2	RNA recognition motif, putative, expressed
GDPDL6	0.67455299	0.4969697	38	3	glycerophosphoryl diester phosphodiesterase family protein, putative, expressed
AT4G11860	0.0	0.27891156	1	4	ubiquitin interaction motif family protein, expressed

AT3G23410	0.0	0.33333333	1	4	alcohol oxidase, putative, expressed
AT4G23400	0.0	0.33333333	1	4	aquaporin protein, putative, expressed
AT4G30850	0.0	0.33333333	1	4	haemolysin-III, putative, expressed
AT1G57870	0.0	0.33333333	1	4	CGMC_GSK.5 - CGMC includes CDA, MAPK, GSK3, and CLKC kinases, expressed
AT1G31812	0.0	0.33333333	1	4	acyl CoA binding protein, putative, expressed
AT1G14360	0.0	0.33333333	1	4	solute carrier family 35 member B1, putative, expressed
AT5G06320	0.0	0.33333333	1	4	harpin-induced protein 1 domain containing protein, expressed
AT1G07550	0.0	0.33333333	1	4	senescence-induced receptor-like serine/threonine-protein kinase precursor, putative, expressed
AT5G07340	0.0	0.33333333	1	4	calreticulin precursor protein, putative, expressed
AT2G41705	0.0	0.33333333	1	4	crcB-like protein, expressed
AT3G12180	0.0	0.33333333	1	4	cornichon protein, putative, expressed
AT5G11890	0.0	0.33333333	1	4	harpin-induced protein 1 domain containing protein, expressed
AT1G14020	0.0	0.33333333	1	4	auxin-independent growth promoter protein, putative, expressed
AT1G34640	0.0	0.33333333	1	4	expressed protein
AT3G66654	0.0	0.33333333	1	4	peptidyl-prolyl cis-trans isomerase, putative, expressed
AT2G22425	0.0	0.33333333	1	4	signal peptidase complex subunit 1, putative, expressed
AT2G27290	0.0	0.33333333	1	4	protein of unknown function DUF1279 domain containing protein, expressed
AT5G49540	0.0	0.33333333	1	4	transmembrane protein 93, putative, expressed
AT1G13770	0.0	0.33333333	1	4	DUF647 domain containing protein, putative, expressed
AT1G29060	0.0	0.33333333	1	4	expressed protein
AT4G14455	0.0	0.33333333	1	4	SNARE domain containing protein, putative, expressed
AT4G25360	0.0	0.33333333	1	4	leaf senescence related protein, putative, expressed
AT4G12250	0.0	0.33333333	1	4	UDP-glucuronate 4-epimerase, putative, expressed
AT5G35460	0.0	0.33333333	1	4	integral membrane protein, putative, expressed
AT1G16170	0.0	0.33333333	1	4	expressed protein
AT5G03345	0.0	0.33333333	1	4	expressed protein
AT1G47640	0.0	0.33333333	1	4	SSA2 - 2S albumin seed storage family protein precursor, putative, expressed
AT5G52420	0.0	0.33333333	1	4	expressed protein
BPA1	0.30818366	0.51898734	6	2	RNA recognition motif containing protein, putative, expressed
AT4G17720	0.30818366	0.51898734	6	2	RNA recognition motif, putative, expressed

The cluster of GDPDL3-interacting proteins includes the BRI1-ASSOCIATED RECEPTOR KINASE1, also designated SOMATIC EMBRYOGENESIS RECEPTOR KINASE3 (BAK1/SERK3). BAK1 has been shown to function as a co-receptor for many RLKs, including the recruitment of receptor-like proteins and SOBIR to form a heterodimeric complex in recognition of ligands by RLPs, for example, RLP23-SOBIR1-BAK1, cf-4-BAK1/SERK3 (Albert et al., 2015, Postma et al., 2016) (Fig. 3A, Table 10).

The interactions of GDPDLs- and SNC4 converge to centralized-hubs represented by BPA1, AT1G01080, and AT4G17720 (BPL1), which contain an RNA binding motif (Fig. 3A, Table 10). The BPA1 protein has been shown to interact with Arabidopsis ACD11, which induces the expression of genes associated with disease resistance and genes involved in the ROS-mediated response defense, upon recognition of fungal elicitors (Li et al., 2019, Petersen et al., 2009). Furthermore, BPA1 and BPL1 are induced during geminivirus infection (Ascencio-Ibáñez et al., 2008). The GDPDLs-Arabidopsis PPI network is enriched for proteins involved in plant defense response to pathogens and in vegetative growth, indicating that this new RLP family may be involved in immunity and in developmental signaling.

To gain further insights into the cellular processes involved by GDPDLs, we performed functional enrichment analyses of their direct interactors. We identify in all three categories, Biological Process, Molecular Function, and Cellular Component ontology, enriched GO terms with p-value <0.05. Under molecular function, we identified enriched terms for Glycerophosphodiester phosphodiesterase activity, nucleotide binding, purine ribonucleotide binding, and hydrolase activity, which are unusual enzyme activities associated with membrane receptor activity (Table 9). Under the cellular component ontology, we observed an over-representation of proteins from plasma membrane term, membrane-bounded term, and plant-type cell wall term, which may suggest that the location and functional activities of these hubs are specific to transmembrane proteins. (Fig. 3B). Under the biological process ontology, the response to defense response, response to external stimulus, and developmental growth term represented significantly enriched GO terms, which show that this family of proteins may be related to immunity and plant development (Table S3).

3.10. The expression profile of the GDPDLs in response to pathogens and in different organs

To gain insights into the potential defense response of the GDPDLs genes and to validate these candidate receptor-like protein as expressed genes, we investigated their expression profiles thought free expression datasets using the gene investigator (www.genevestigator.ethz.ch- academic free license) (Fig. 2C). From these microarray data, GDPDL1-RLK has been shown to be induced by aphids, the bacteria *Pseudomonas syringae,* and the begomovirus *Cabbage leaf curl virus* (CaCuLV), but not by nematodes. Likewise, GDPDL2-RLP is induced by bacteria and aphids and to a less extent by begomovirus. GDPDL3-RLP and GDPDL4-RLP are up-regulated by aphids and bacteria and down-regulated by begomovirus. GDPDL5 and GDPDL6 are not induced by aphids and bacteria, but down-regulated by CaCuLV. For organ-specific expression, GDPDL5 and GDPDL6 are only expressed in flower and siliques, whereas GDPDLl-RLK and the remaining GDPDL-RLPs are expressed in all organs analyzed.

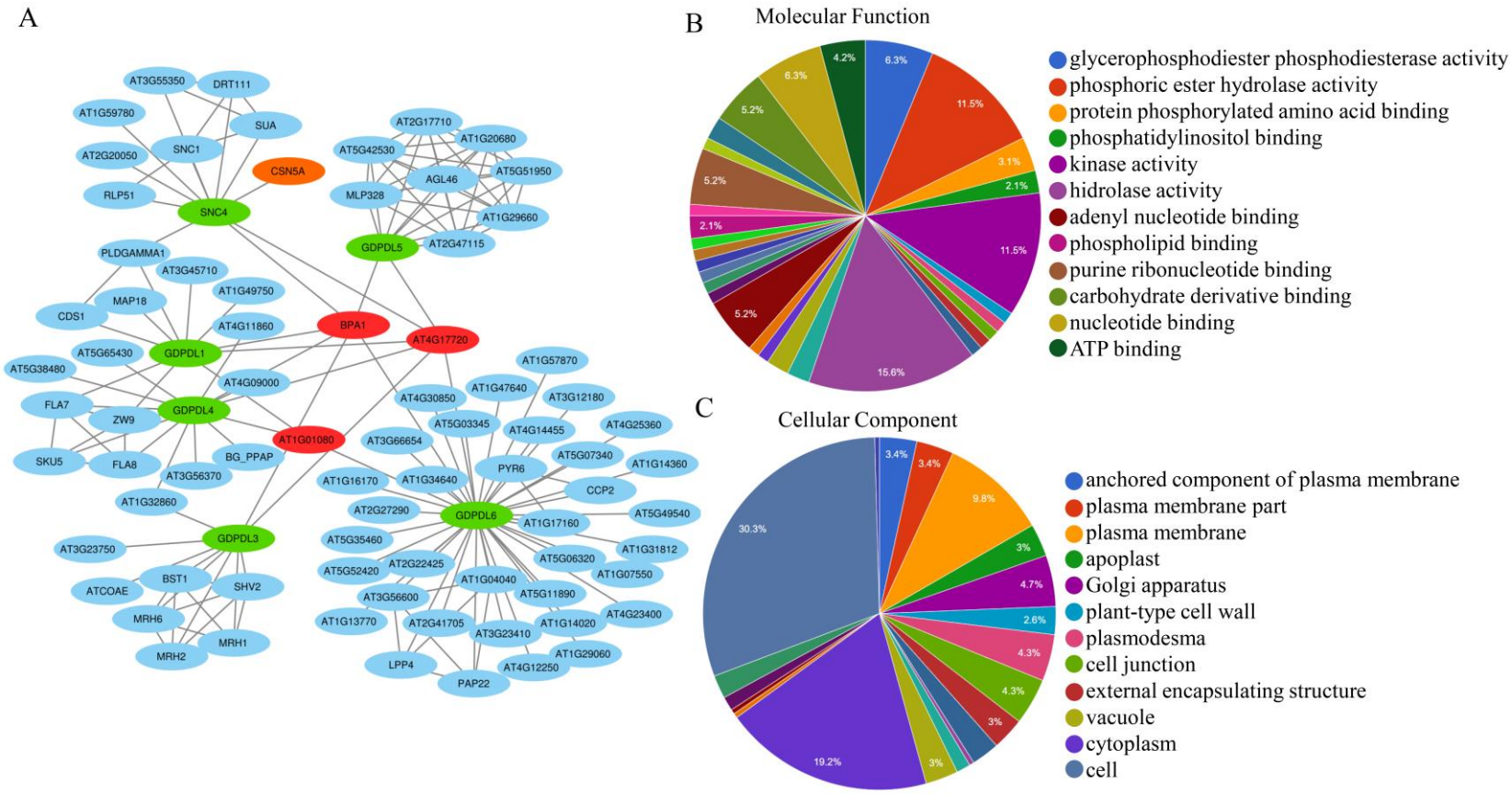

Fig. 3. GDPDL-RLPs- interacting Arabidopsis proteins. (A) GDPDL-RLP-interacting proteins were identified in the Arabidopsis interactome and the network was assembled by the Cytoscape software. GDPDL-RLPs and SNC4 are indicated in green, GDPDL-specifically interacting proteins in ligth blue, RNA-binding proteins, which interact with all 6 GDPDLs, including GDPDL_RLK, are shown in red. In orange, CSN5A which represents a cental hub of plant-pathogen interactions (B) Gene enrichment of proteins under the molecular function term from the GDPD-RLP-Arabidopsis PPI network. (C) Gene enrichment of proteins from the GDPD-RLP-Arabidopsis PPI network under the cellular component term.

Pathogen-induced and organ-specific expressions of the predicted GDPDL-RLP genes were confirmed by qRT-PCR (Figures 4 and 5). We monitored the expression of the GDPDL-RLP genes in response to infections with the viruses *Tobacco rattle virus* (TRV) and CaCuLV. We also activated antibacterial immune responses (PTI) with flg22 and monitored the expression of GDPDLs (Figure 4). Consistent with the microarray data, GDPL5 and GDPL6 were not induced by flg22 and was down-regulated by CaLCuV, whereas all the other GDPDLs were induced by flg22. Interesting all 5 GDPDLs analyzed by qRT-PCR were induced by TRV, a plant RNA virus. Remarkably, these GDPDL proteins are interconnected to each other via interactions with RNA recognition motif proteins, which form centralized hubs in the network interaction (Fig. 3A, Table 8). This result may suggest an involvement of GDPDLs in the antiviral response induced by RNA virus.

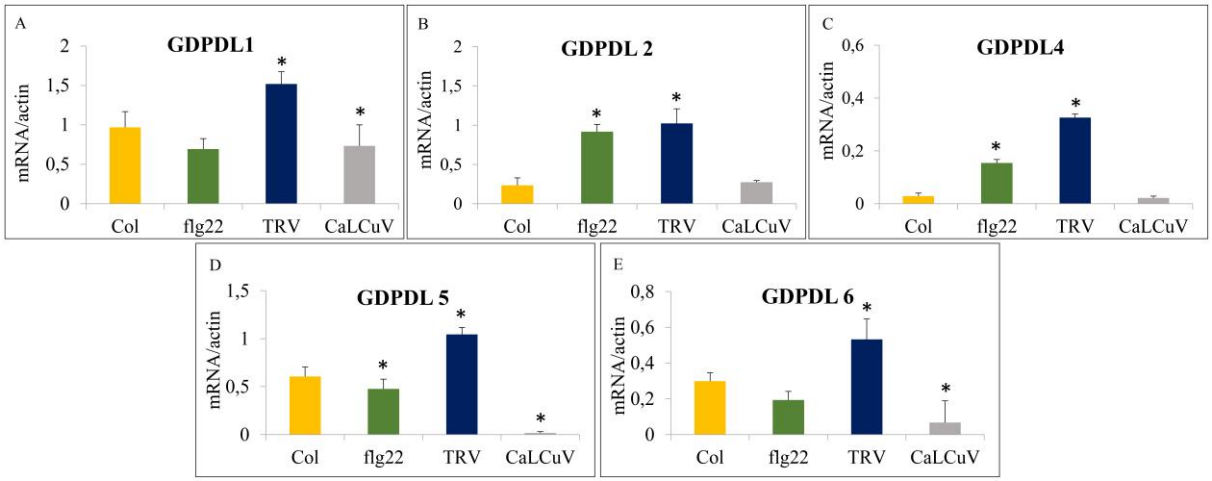

Fig. 4. Expression analysis of the GDPDL genes in response do biotic signal. Total RNA was extracted from Arabidopsis seedlings treated with flg22 or infected with the viruses TRV and CaLCuV and transcript accumulation of the indicated genes was monitored by quantitative RT-PCR with gene-specific primers. The gene expression was calculated by the $2^{-\Delta CT}$ method using actin as an endogenous control. The error or standard bars indicate the mean ± SD of three independent experiments. *P < 0.05; Tukey test.

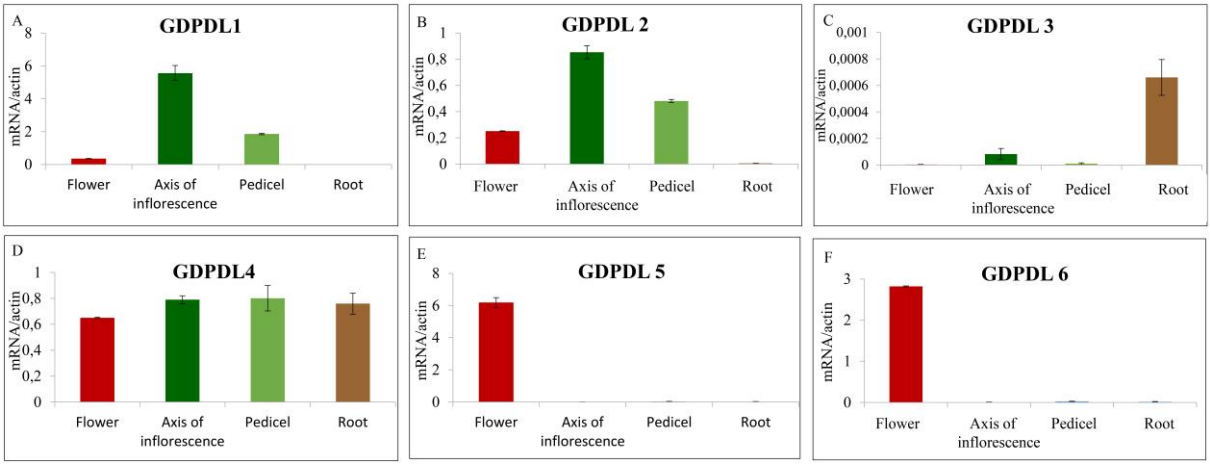

Fig. 5. Organ-specific expression of the GDPDL genes. Total RNA was extracted from different organs of Arabidopsis seedlings and the transcript levels of the indicated genes were determined by qRT-PCR using gene-specific primers. The gene expression was calculated by the $2^{-\Delta CT}$ method using actin as an endogenous control. The error or standard bars indicate the mean ± SD of three independent experiments.

We also confirmed the expression profile of these GDPDL genes in different tissues by qRT-PCR. We used tissues of the root, pedicel, axis of inflorescence, and flower. The expression levels of GDPDL1 and GDPDL2 are similar in all tissues (Fig. 5A, B). The highest expression levels were identified in the axis of inflorescence and pedicel, suggesting different functions in development. Likewise, GDPDL3 is most expressed in roots and barely detected in other tissues (Fig. 5C). Interestingly, the expression levels of GDPDL4 are regular in all tissues, showing that this protein may have a varied role during development (Fig. 5D), whereas, qRT-PCR confirmed that the GDPDL5 and GDPDL6 transcripts accumulated to high levels in flowers (Fig. 5E, F). Collectively, these gene expression analyses confirmed that GDPDL-RLPs are expressed in response to stimuli and development, substantiating the argument that they form a new class of RLPs most likely involved in immunity and developmental signaling.

4. Discussion

Due to the relevance of the RLK family in several biological processes and molecular functions, this large family has been extensively studied in different plant species (Shiu et al., 2001, Sakamoto et al., 2012, Liu et al., 2015, Yan et al. 2017, Dezhsetan et al., 2017, Zuo et al., 2019, Yan et al., 2018). Nevertheless, far less is known about the plant RLP family, in spite of their conceptual relevance in signaling modules. In fact, RLPs can perceive external signals, but depend on association with RLKs for signal transduction due to the lack of a cytoplasmic kinase domain at the C-terminus. The absence of a conserved kinase domain precludes the use of sequence comparison algorithms for genome-wide studies of the plant RLP family. Thus, the identification of RLPs in plant genomes is a challenge, and few RLPs have been described in plant species. Moreover, a large-scale RLP prediction tool has not been developed. Here, we developed the RLPredictiOme method based on machine learning approaches and bayesian inference for throughout prediction of RLPs.

Typically, the ML classification models applied in plant molecular biology require real data to train ML supervised algorithm (Silva et al., 2019, Pal et al., 2016, Ni et al., 2016, Kushwaha et al., 2016), The RLPredictiOme can predict RLP subfamilies using only the RLK ectodomain along with six features simultaneously used during all the prediction process. The prediction model consists of three steps subsequently built with trained models and different algorithms capable of distinguishing RLP from noRLP, RLP from RLKs and finally predicting an RLP subfamily. The combination of several ML models with different algorithms has not been applied for protein and viral sequence classification (Carvalho et al., 2017, Silva et al., 2017). The use of different classifiers requires methods that compile the results of the classifiers into a single final prediction. Some methods have used different techniques for model's combination, including a majoritarian vote of the classifiers or using an average probabilityfor the classifications (Dietterich et al., 2000, Carvalho et al., 2017). The approaches applied in the RLPredictiOme for combining models are based on the results of the success and failure of predictions, which are modeled with bayesian inference. In each step after the classifications, the Bayesian inference is applied. The validation results of the RLPredictiOme showed relatively high probabilities for classifying RLPs proteins (See Table 6, columns RLP-noRLP Probability, RLP-RLK Probability, and RLP-Subfamily Probability). In contrast, noRLP proteins were predicted with a relatively lower probability (Table 7).

Finally, based on the probability of Bayesian inferences for each step, a final step is used as a decision-making process for the prediction of RLPs (Fig.1 F). The RLPredictiOme predicts RLP proteins with a probability ranging from 0.79 to 0.99 (See Tables 6, 7 and 8, column Decision probability). Thus, the ML models can be successfully combined with Bayesian inference to perform robust high-throughput predictions of RLPs in plant genomes.

The RLPredictiOme was able to predict new RLP subfamilies with relatively higher probability in all steps, although groups less-represented were classified into a corresponding subfamily with relatively lower probability. Furthermore, groups less-represented of RLPs tended to be classified within other-RLP subfamilies. This was the case of the probable lipid transfer-RLP subfamily, which shares similar functional characteristics with LysM-RLP. The lipid transfer proteins (LTPs) already described as non-specific lipid transfer proteins (nsLTPs) contains an eight-cysteine motif that is stabilized by four disulfide bonds (Wang et. Al, 2019). The probable lipid transfer family (PLT)-RLPs identified by RLPredictiOme harbor a five-cysteine motif (CC-Xn-CXC-Xn-C) in the TP_2 functional domain differently from the typical nsLTPs (Wang, et al., 2020). Phylogenetics relations, structure, and genome-wide distribution of LTPs have been described in cucumber, which have been demonstrated to be involved in response to nematode (Wang et. Al, 2019). Furthermore, PLTs have been shown to play an important role in regulating various plant biological processes and responding to biotic and abiotic stress (Torres-Schumann et al., 1992, Kapoor et al., 2019). Due to evidence of association with kinases, PTL-RLPs may be classified as a new subfamily of RLPs or may represent an expansion of the LysM-RLP subfamily, which exhibit similar functional roles.

In silico and *in vitro* analyses of GDPDL-RLPs confirmed the efficiency of the RLPredictiOme to identify a new family of RLPs based on the ectodomain of GDPDL-RLK sequences. The GDPDL-RLK is a reduced class of RLKs in plants. Among all the plant species analysed, they were found only in *Arabidopsis halleri* (Araha.28943s0001.1), *Arabidopsis lyrat*a (475793), *Arabidopsis thaliana* (AT1G66980.1), *Boechera stricta* (Bostr.26959s0213.1, Bostr.26959s0216.1), *Brassica rapa* (Brara.K00110.1), all from the *Brassicaceae* family and *Capsella grandiflora* (Cagra.0792s0001.1) and *Panicum virgatum*, (Pavir.6NG294600.1) from the *Poaceae* family Despite only one GDPDL-RLK in the Arabidopsis genome (Bi et al., 2010), RLPreditiOme identified five sequences as GDPDL-

RLP. Furthermore, the GDPDL-RLK subfamily has been maintained in only few plant species thereby it is likely that this family is suffering a reduction in size and distribution. The GDPDL2-RLK (AT1G66980) has been previously characterized as SNC4, an atypical receptor-like kinase with predicted extracellular GDPD domain, which are involved in the regulation of plant immunity (Zhang et al., 2014). The Glycerophosphodiester phosphodiesterase (GDPD) hydrolyzes glycerophosphodiesters into sn-glycerol-3-phosphate (G-3-P) and plays an important role in various biological processes (Zhang et al., 2014). The GDPDL2-RLK ectodomain is structurally similar to the predicted GDPDL-RLPs (Fig. 2B). Molecular evolution investigated by calculating ka/ks of GDPDL-RLP-GDPDL-RLK pairs revealed a significant rate of synonymous substitutions indicating that although the loss of the kinase domain has occurred, the functional characteristics of the ectodomain remained conserved among evolution (Table 9).

A common feature of the RLK subfamilies is that they are often larger than the RLP subfamily counterparts are, which suggests that some members of the RLK subfamilies have lost their conserved C-terminal kinase domain during evolution. In contrast, RLPredictiOme identified a new RLP subfamily, GDPDL-RLP, which seems to have expanded in comparison with the corresponding GDPDL-RLK subfamily. Therefore, it was of our interest to examine the expression profile of the GDPDL-RLP members to ensure a basal level of expression during development or in response to pathogens. *In silico* analyses from publicly available expression databases indicated that the RLP members display different expression profiles in different organs and in response to pathogens, indicating that may be involved in development and immunity.

GDPDL1 (GDPGL-RLP) has been previously shown to be expressed in the rosettes of Arabidopsis plants (Duruflé et al., 2017). We confirmed by qRT-PCR that GDPDL1 is expressed in the pedicels of the rosette and also in the flowers. GDPDL1 has also been shown to be involved in processes that confer rigidity to the cell wall (Duruflé et al., 2017), which are related to defense against insects, nematodes and oomycetes, as demonstrated by the previously published microarray data. We showed here by qRT-PCR that GDPDL1 is induced in the response to the plant virus TRV.

The expression profile of the five RLP members were also examined. GDPDL1 and GDPDL2 displayed the highest expressed in pedicels and flower stems and were highly

induced by flg22. Among all members of this new GDPDL family, GDPDL3 was barely detected in all organs examined, except roots, consistent with its role in root morphogenesis (Hayashi et al., 2008). GDPL4 was uniformly expressed in all organs evaluated. GDPDL4 has been described as a highly expressed gene in rosettes and it is involved in the development of root hair (Duruflé et al., 2017; Salazar-Henao et al., 2016). In reponse to biotic stimuli, the accumulation of GDPDL4 transcripts was very similar to GDPDL2.

Two members of this family, GDPDL4 and GDPDL5, displayed high levels of expression in flowers, which indicates that both genes may be individual in the development of reproductive organs and structures. Regarding the activation of the expression of these genes by the pathogens tested, they were only induced by infection with TRV however, microarray data indicate that GDPDL4 and GDPDL5 are induced in response to nematodes. The expression pattern and evolution studies of members of GDPGL-RLP subfamily further substantiate the notion that the members of this subfamily have maintained functional domains and may play relevant roles in development and in plant defense.

5. References

ALBERT, I. et al. An RLP23–SOBIR1–BAK1 complex mediates NLP-triggered immunity. **Nature Plants** 1.10, pp. 1–9. 2015.

ASCENCIO-IBÁÑEZ, J. T. et al. Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle duringgeminivirus infection. **Plant physiology,** 148.1, pp. 436–454. 2008.

BHASIN, M. AND GAJENDRA P.S. Classification of nuclear receptors based on amino acid composition and dipeptide composition. **Journal of Biological Chemistry,** 279.22, pp. 23262–23266. 2004.

BI, D., et al. Activation of plant immune responses by a gain-of-function mutation in an atypical receptor-like kinase. **Plant Physiology**, 153.4, pp. 1771–1779. 2010.

BOTOS, I., SEGAL D. M., AND DAVIES D. R. The structural biology of Toll-like receptors. **Structure**, 19.4, pp. 447–459. 2011.

BUENDIA, LUIS et al. LysM receptor-like kinase and LysM receptor-like protein families: an update on phylogeny and functional characterization. **Frontiers in Plant Science**, 9, p. 1531. 2018.

CAO, YANGRONG et al. The kinase LYK5 is a major chitin receptor in Arabidopsis and forms a chitin-induced complex with related kinase CERK1. **Elife**, 3, e03766. 2014.

CARVALHO, THALES FRANCISCO MOTA et al. Rama: a machine learning approach for ribosomal protein prediction in plants. **Scientific Reports**, 7.1, pp. 1–13. 2017.

CHAWLA, NITESH V et al. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, 16, pp. 321–357. 2002.

CHO, SUNG KI et al. Regulation of floral organ abscission in Arabidopsis thaliana. **Proceedings of the National Academy of Science**s, 105.40, pp. 15629–15634. 2008.

DEZHSETAN, SARA. Genome scanning for identification and mapping of receptor-like kinase RLK gene superfamily in Solanum tuberosum. **Physiology and Molecular Biology of Plants**, 23.4, pp. 755–765. 2017.

DIETTERICH, THOMAS G. Ensemble methods in machine learning. International workshop on multiple classifier systems. **Springer**, pp. 1–15. 2000.

DURUFLÉ, HAROLD et al. Cell wall modifications of two Arabidopsis thaliana ecotypes, Col and Sha, in response to sub-optimal growth conditions: an integrative study. **Plant Science**, 263, pp. 183–193. 2017.

FAULKNER, CHRISTINE et al. LYM2-dependent chitin perception limits molecular flux via plasmodesmata. **Proceedings of the National Academy of Sciences**, 110.22, pp. 9166–9170. 2013.

FELLER, WILLLIAM. An introduction to probability theory and its applications. Vol. 2. **John Wiley & Sons**. 2008.

FREUND, YOAV AND ROBERT E SCHAPIRE. A desicion-theoretic generalization of online learning and an application to boosting. European conference on computational learning theory. **Springer**, pp. 23–37. 1995.

FRIEDMAN, JEROME H. Stochastic gradient boosting**. Computational Statistics & Data Analysis**, 38.4, pp. 367–378. 2002.

FU, LIMIN et al. CD-HIT: accelerated for clustering the next-generation sequencing data. **Bioinformatics**, 28.23, pp. 3150–3152. 2012.

GAO, LIN-LIN AND HONG-WEI XUE. Global analysis of expression profiles of rice receptor-like kinase genes. **Molecular Plant**, 5.1, pp. 143–153. 2012.

GEMAN, STUART AND DONALD GEMAN. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 6, pp. 721–741. 1984.

GÓMEZ-GÓMEZ, LOURDES AND THOMAS BOLLER. FLS2: an LRR receptor–like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. **Molecular Cell**, 5.6, pp. 1003–1011. 2000.

GUPTA, ARJUN K AND SARALEES NADARAJAH. Handbook of beta distribution and its applications. **CRC Press**. 2004.

HAGHIGHI, SEPAND et al. PyCM: Multiclass confusion matrix library in Python. **Journal of Open Source Software**, 3.25, p. 729. 2018.

HARUTA, MIYOSHI et al. A peptide hormone and its receptor protein kinase regulate plant cell expansion. **Science**, 343.6169, pp. 408–411. 2014

HASTIE, TREVOR, ROBERT TIBSHIRANI, AND JEROME FRIEDMAN. The elements of statistical learning: data mining, inference, and prediction. **Springer Science & Business Media**. 2009.

HAYASHI, SHIMPEI et al. The glycerophosphoryl diester phosphodiesterase-like proteins SHV3 and its homologs play important roles in cell wall organization. **Plant and Cell Physiology**, 49.10, pp. 1522–1535. 2008.

HE, YUNXIA et al. Plant cell surface receptor-mediated signaling–A common theme amid diversity. **J Cell Sci**, 131.2, jcs209353. 2018.

HINTON, GEOFFREY et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. **IEEE Signal Processing Magazine**, 29.6, pp. 82–97. 2012

HIRAKAWA, YUKI et al. Non-cell-autonomous control of vascular stem cell fate by a CLE peptide/receptor system. **Proceedings of the National Academy of Sciences**, 105.39, pp. 15208–15213. 2008.

HRUZ, TOMAS et al. Genevestigator v3: a reference expression database for themetaanalysis of transcriptomes. **Advances in Bioinformatics**. 2008.

JAMIESON, PIERCE A, LIBO SHAN, AND PING HE. Plant cell surface molecular cypher: receptor-like proteins and their roles in immunity and development. **Plant Science**, 274, pp. 242–251. 2018.

JEONG, SANGHO, AMY E TROTOCHAUD, AND STEVEN E CLARK. The Arabidopsis CLAVATA2 gene encodes a receptor-like protein required for the stability of the CLAVATA1 receptorlike kinase. **The Plant Cell**, 11.10, pp. 1925–1933. 1999.

JIA, GENGXIANG et al. Signaling of cell fate determination by the TPD1 small protein and EMS1 receptor kinase. **Proceedings of the National Academy of Sciences**, 105.6, pp. 2220–2225. 2008.

JONAK, CLAUDIA AND HERIBERT HIRT. Glycogen synthase kinase 3/SHAGGY-like kinases in plants: an emerging family with novel functions. **Trends in Plant Science**, 7.10, pp. 457–461. 2002.

JONES, DAVID A et al. Isolation of the tomato Cf-9 gene for resistance to Cladosporium fulvum by transposon tagging. **Science**, 266.5186, pp. 789–793. 1994.

JOSÉ-ESTANYOL, MATILDE, F XAVIER GOMIS-RÜTH, AND PERE PUIGDOMÈNECH. Theeightcysteine motif, a versatile structure in plant proteins. **Plant Physiology and Biochemistry**, 42.5, pp. 355–365. 2004.

KÄLL, LUKAS, ANDERS KROGH, AND ERIK LL SONNHAMMER. A combined transmembrane topology and signal peptide prediction method. **Journal of Molecular Biology**, 338.5, pp. 1027–1036. 2004.

KAPOOR, RITU et al. Genome-Wide Analysis and Expression Profiling of Rice Hybrid Proline-Rich Proteins in Response to Biotic and Abiotic Stresses, and Hormone Treatment. **Plants**, 8.9, p. 343. 2019.

KIM, KANG SOO et al. Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. **Current Applied Physics**, 11.3, pp. 740–745. 2011.

KING, GARY AND LANGCHE ZENG. Logistic regression in rare events data. **Political Analysis**, 9.2, pp. 137–163. 2001.

KUMPF, ROBERT P et al. Floral organ abscission peptide IDA and its HAE/HSL2 receptors control cell separation during lateral root emergence. **Proceedings of the National Academy of Sciences**, 110.13, pp. 5235–5240. 2013.

KUSHWAHA, SANDEEP K et al. NBSPred: a support vector machine-basedhighthroughput pipeline for plant resistance protein NBSLRR prediction. **Bioinformatics**, 32.8, pp. 1223–1225. 2016.

LEE, JIN SUK et al. Direct interaction of ligand–receptor pairs specifying stomatal patterning. **Genes & Development**, 26.2, pp. 126–136. 2012.

LI, BO et al. The receptor-like kinase NIK1 targets FLS2/BAK1 immune complex and inversely modulates antiviral and antibacterial immunity. **Nature Communications**, 10.1, pp. 1–14. 2019.

LI, JIANMING AND JOANNE CHORY. A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. **Cell**, 90.5, pp. 929–938. 1997.

LI, QI et al. A phytophthora capsici effector targets ACD11 binding partners that regulate ROS-mediated defense response in arabidopsis. **Molecular Plant**, 12.4, pp. 565– 581. 2019.

LIN, GUANGZHONG et al. A receptor-like protein acts as a specificity switch for the regulation of stomatal development. **Genes & Development**, 31.9, pp. 927–938. 2017.

LIU, BING et al. Lysin motif–containing proteins LYP4 and LYP6 play dual roles in peptidoglycan and chitin perception in rice innate immunity. **The Plant Cell**, 24.8, pp. 3406–3419. 2012.

LIU, JINYI et al. Soybean kinome: functional classification and gene expression patterns. **Journal of Experimental Botany**, 66.7, pp. 1919–1934. 2015.

MA, XIYU et al. SERKing coreceptors for receptors. **Trends in Plant Science**, 21.12, pp. 1017–1033. 2016.

MACHO, ALBERTO P AND CYRIL ZIPFEL. Plant PRRs and the activation of innate immune signaling. **Molecular Cell**, 54.2, pp. 263–272. 2014.

MIYA, AYAKO et al. CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. **Proceedings of the National Academy of Sciences**, 104.49, pp. 19613–19618. 2007.

NI, YING ET AL. A machine learning approach to predict gene regulatory networks in seed development in Arabidopsis. **Frontiers in Plant Science**, 7, p. 1936. 2016.

NIELSEN, HENRIK Predicting secretory proteins with SignalP. Protein function prediction. **Springer**, pp. 59–73. 2017.

OGAWA, MARI et al. Arabidopsis CLV3 peptide directly binds CLV1 ectodomain. **Science**, 319.5861, pp. 294–294. 2008.

OMASITS, ULRICH et al. Protter: interactive protein feature visualization and integration with experimental proteomic data. **Bioinformatics**, 30.6, pp. 884–886. 2014.

PAL, TARUN, VARUN JAISWAL, AND RAJINDER S CHAUHAN, DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. **Computers in Biology and Medicine**, 78, pp. 42–48. 2016.

PEDREGOSA, FABIAN et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, 12.Oct, pp. 2825–2830. 2011.

PETERSEN, NIKOLAJ HT et al. Identification of proteins interacting with Arabidopsis ACD11. **Journal of Plant Physiology**, 166.6, pp. 661–666. 2009.

PETUTSCHNIG, ELENA K et al. The lysin motif receptor-like kinase LysM-RLK CERK1 is a major chitin-binding protein in Arabidopsis thaliana and subject to chitin-induced phosphorylation. **Journal of Biological Chemistry**, 285.37, pp. 28902–28911. 2010.

PLATT, JOHN et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. **Advances in Large Margin Classifiers**, 10.3, pp. 61– 74. 1999.

POSTMA, JELLE et al. Avr4 promotes Cf-4 receptor-like protein association with the BAK1/SERK3 receptor-like kinase to initiate receptor endocytosis and plant immunity. **New Phytologist**, 210.2, pp. 627–642. 2016.

PRICE, MORGAN N, PARAMVIR S DEHAL, AND ADAM P Arkin, FastTree 2–approximately maximum-likelihood trees for large alignments. **PLoS One**, 5.3. 2010.

RANF, STEFANIE et al. A lectin S-domain receptor kinase mediates lipopolysaccharide sensing in Arabidopsis thaliana. **Nature Immunology**, 16.4, p. 426. 2015.

SAKAMOTO, TETSU et al. The tomato RLK superfamily: phylogeny and functional predictions about the role of the LRRII-RLK subfamily in antiviral defense. **BMC Plant Biology**, 12.1, p. 229. 2012.

SALAZAR-HENAO, JORGE E, WEN-DAR LIN, AND WOLFGANG SCHMIDT, Discriminative gene co-expression network analysis uncovers novel modules involved in the formation of phosphate deficiency-induced root hairs in Arabidopsis. **Scientific Reports**, 6, p. 26820. 2016.

SALVATIER, JOHN, THOMAS V WIECKIÂ, AND CHRISTOPHER FONNESBECK PyMC3: Python probabilistic programming framework. **ASCL**, ascl–1610. 2016.

SAMWORTH, RICHARD J et al. Optimal weighted nearest neighbour classifiers. **The Annals of Statistics**, 40.5, pp. 2733–2763. 2012.

SARAVANAN, VIJAYAKUMAR AND NAMASIVAYAM GAUTHAM, Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid

composition-based feature descriptor. **Omics: a Journal of Integrative Biology**, 19.10, pp. 648–658. 2015.

SCHMIDT, MARK, NICOLAS LE ROUX, AND FRANCIS BACH, Minimizing finite sums with the stochastic average gradient. **Mathematical Programming**, 162.1-2, pp. 83–112. 2017.

SHIU, SHIN-HAN AND ANTHONY B BLEECKER, Plant receptor-like kinase gene family: diversity, function, and signaling. **Sci. STKE**, 2001.113, re22–re22. 2001a.

SHIU, SHIN-HAN AND ANTHONY B BLEECKER, Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. **Proceedings of the National Academy of Sciences,** 98.19, pp. 10763–10768. 2001b.

SHIU, SHIN-HAN AND ANTHONY B BLEECKER. Expansion of the receptor-like kinase/Pellegene family and receptor-like proteins in Arabidopsis. **Plant Physiology**, 132.2, pp. 530–543. 2003.

SHIU, SHIN-HAN, WOJCIECH M KARLOWSKI, et al. Comparative analysis of thereceptorlike kinase family in Arabidopsis and rice. **The Plant Cell**, 16.5, pp. 1220–1234. 2004.

SILVA, JOSE CLEYDSON F et al. Machine learning approaches and their current application in plant molecular biology: A systematic review. **Plant Science**, 284, pp. 37– 47. 2019.

SILVA, JOSÉ CLEYDSON F et al. Fangorn Forest F2: a machine learning approach to classify genes and genera in the family Geminiviridae. **BMC Bioinformatics**, 18.1, p. 431. 2017.

SONNHAMMER, ERIK LL, GUNNAR VON HEIJNE, ANDERS KROGH, et al. A hidden Markov model for predicting transmembrane helices in protein sequences. **Ismb**. Vol. 6, pp. 175–182. 1998.

TANG, DINGZHONG, GUOXUN WANG, AND JIAN-MIN ZHOU. Receptor kinases inplantpathogen interactions: more than pattern recognition. **The Plant Cell**, 29.4, pp. 618–637. 2017.

TEIXEIRA, RUAN M et al. Virus perception at the cell surface: revisiting the roles of receptor-like kinases as viral pattern recognition receptors. **Molecular Plant Pathology**, 20.9, pp. 1196–1202. 2019.

THOMAS, COLWYN M et al. Characterization of the tomato Cf-4 gene for resistance to Cladosporium fulvum identifies sequences that determine recognitional specificity in Cf-4 and Cf-9. In: **The Plant Cell**, 9.12, pp. 2209–2224. 1997.

TORRES-SCHUMANN, SONIA, JOSÉ A GODOY, AND JOSÉ A PINTOR-TORO, A probable lipid transfer protein gene is induced by NaCl in stems of tomato plants. **Plant Molecular Biology**, 18.4, pp. 749–757. 1992.

WAESE, JAMIE et al. ePlant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. **The Plant Cell**, 29.8, pp. 1806–1821. 2017.

WAN, JINRONG, KIWAMU TANAKA, et al. LYK4, a lysin motif receptor-like kinase, is important for chitin signaling and plant innate immunity in Arabidopsis. **Plant Physiology**, 160.1, pp. 396–406. 2012.

WAN, JINRONG, XUE-CHENG ZHANG, et al. A LysM receptor-like kinase plays a critical role in chitin signaling and fungal resistance in Arabidopsis. **The Plant Cell**, 20.2, pp. 471–481. 2008.

WANG, JIZONG et al. Allosteric receptor activation by the plant peptide hormone phytosulfokine. **Nature**, 525.7568, pp. 265–268. 2015.

WANG, XING et al. Genome-wide analysis of putative lipid transfer protein LTP_2 gene family reveals CsLTP_2 genes involved in response of cucumber against root-knot nematode Meloidogyne incognita. **Genome Ja**, 2020.

WILLMANN, ROLAND et al. Arabidopsis lysin-motif proteins LYM1 LYM3 CERK1 mediate bacterial peptidoglycan sensing and immunity to bacterial infection. **Proceedings of the National Academy of Sciences**, 108.49, pp. 19824–19829. 2011.

YAMAGUCHI, YUBE, ALISA HUFFAKER, et al. PEPR2 is a second receptor for the Pep1 and Pep2 peptides and contributes to defense responses in Arabidopsis. **The Plant Cell**, 22.2, pp. 508–522. 2010.

YAMAGUCHI, YUBE, GREGORY PEARCE, AND CLARENCE A RYAN. The cell surfaceleucinerich repeat receptor for AtPep1, an endogenous peptide elicitor in Arabidopsis, is functional in transgenic tobacco cells. **Proceedings of the National Academy of Sciences**, 103.26, pp. 10104–10109. 2006.

YAN, JUN, GUILIN LI, et al. Genome-wide classification, evolutionary analysis and gene expression patterns of the kinome in Gossypium. **PLoS One**, 13.5. 2018.

YAN, JUN, PEISEN SU, et al. Genome-wide identification, classification, evolutionary analysis and gene expression patterns of the protein kinase gene family in wheat and Aegilopstauschii. **Plant Molecular Biology**, 95.3, pp. 227–242. 2017.

ZHANG, ZHIBIN et al. Splicing of receptor-like kinase-encoding SNC4 and CERK1 is regulated by two conserved splicing factors that are required for plant immunity. **Molecular Plant**, 7.12, pp. 1766–1775. 2014.

Zhou, Fulai, Yong Guo, and Li-Juan Qiu.ZHOU, FULAI, YONG GUO, AND LI-JUAN QIU 2016. Genome-wide identification and evolutionary analysis of leucine-rich repeat receptor-like protein kinase genes in soybean. **BMC Plant Biology**, 16.1, p. 58. 2016.

ZIPFEL, CYRIL et al. Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts Agrobacterium-mediated transformation. **Cell**, 125.4, pp. 749–760. 2006.

ZORZATTO, CRISTIANE et al. NIK1-mediated translation suppression functions as a plant antiviral immunity mechanism. **Nature**, 520.7549, pp. 679–682. 2015.

ZUO, CUNWU et al. Genome-Wide Analysis of the Apple Malus domesticaCysteineRich Receptor-Like Kinase CRK Family: Annotation, Genomic Organization, and Expression Profiles in Response to Fungal Infection. **Plant Molecular Biology Reporter**, pp. 1–11. 2019

6. Appendix

Table S1. Validation of the LRR-RLP subfamily of genome-wide rice.

1SM	TM	RLP-noRLP	RLP-noRLP Probability	RLP-RLK	RLP-RLK Probability	RLP-Subfamily	RLP-Subfamily Probability	Classification	Decision probability
Y	Y	RLP	0.9964	RLP	0.9909	LRR-RLP	0.8335	(LRR-RLP)	0.991
Y	Y	RLP	0.9966	RLP	0.9917	LRR-RLP	0.8514	(LRR-RLP)	0.9898
Y	Y	RLP	0.9966	RLP	0.9913	LRR-RLP	0.8507	(LRR-RLP)	0.9893
Y	Y	RLP	0.9958	RLP	0.7127	LRR-RLP	0.8501	(LRR-RLP)	0.9893
Y	Y	RLP	0.9961	RLP	0.8586	LRR-RLP	0.7664	(LRR-RLP)	0.9893
Y	Y	RLP	0.9968	RLP	0.8564	LRR-RLP	0.8506	(LRR-RLP)	0.9891
Y	Y	RLP	0.9962	RLP	0.9895	LRR-RLP	0.8501	(LRR-RLP)	0.9891
Y	Y	RLP	0.9967	RLP	0.8576	LRR-RLP	0.8493	(LRR-RLP)	0.9891
Y	Y	RLP	0.9965	RLP	0.8494	LRR-RLP	0.8331	(LRR-RLP)	0.989
Y	Y	RLP	0.9958	RLP	0.9896	LRR-RLP	0.8502	(LRR-RLP)	0.9889
Y	Y	RLP	0.9958	RLP	0.9913	LRR-RLP	0.8339	(LRR-RLP)	0.9889
Y	Y	RLP	0.9962	RLP	0.9897	LRR-RLP	0.8509	(LRR-RLP)	0.9888
Y	Y	RLP	0.9961	RLP	0.8568	LRR-RLP	0.7677	(LRR-RLP)	0.9888
Y	Y	RLP	0.9967	RLP	0.9899	LRR-RLP	0.8493	(LRR-RLP)	0.9888
Y	Y	RLP	0.9472	RLP	0.9907	LRR-RLP	0.8505	(LRR-RLP)	0.9888
Y	Y	RLP	0.9965	RLP	0.8566	LRR-RLP	0.8497	(LRR-RLP)	0.9887
Y	Y	RLP	0.9967	RLP	0.8593	LRR-RLP	0.8485	(LRR-RLP)	0.9886
Y	Y	RLP	0.9961	RLP	0.8544	LRR-RLP	0.833	(LRR-RLP)	0.9886
Y	Y	RLP	0.9967	RLP	0.9915	LRR-RLP	0.8501	(LRR-RLP)	0.9883
Y	Y	RLP	0.9963	RLP	0.8562	LRR-RLP	0.5677	(LRR-RLP)	0.9883
Y	Y	RLP	0.9964	RLP	0.991	LRR-RLP	0.8501	(LRR-RLP)	0.9882
Y	Y	RLP	0.9959	RLP	0.9901	LRR-RLP	0.8511	(LRR-RLP)	0.9881
Y	Y	RLP	0.9966	RLP	0.9889	LRR-RLP	0.8496	(LRR-RLP)	0.988
Y	Y	RLP	0.9968	RLP	0.8577	LRR-RLP	0.8494	(LRR-RLP)	0.988
Y	Y	RLP	0.9965	RLP	0.9918	LRR-RLP	0.8502	(LRR-RLP)	0.9878
Y	Y	RLP	0.9959	RLP	0.8523	LRR-RLP	0.8511	(LRR-RLP)	0.9878
Y	Y	RLP	0.9949	RLP	0.9899	LRR-RLP	0.8504	(LRR-RLP)	0.9878
Y	Y	RLP	0.9961	RLP	0.9917	LRR-RLP	0.85	(LRR-RLP)	0.9877
Y	Y	RLP	0.9961	RLP	0.9891	LRR-RLP	0.8	(LRR-RLP)	0.9877
Y	Y	RLP	0.9963	RLP	0.7165	LRR-RLP	0.7998	(LRR-RLP)	0.9877
Y	Y	RLP	0.9962	RLP	0.8512	LRR-RLP	0.8511	(LRR-RLP)	0.9876
Y	Y	RLP	0.9968	RLP	0.7135	LRR-RLP	0.8328	(LRR-RLP)	0.9876
Y	Y	RLP	0.9964	RLP	0.9892	LRR-RLP	0.8501	(LRR-RLP)	0.9876
Y	Y	RLP	0.9964	RLP	0.8568	LRR-RLP	0.8495	(LRR-RLP)	0.9876
Y	Y	RLP	0.996	RLP	0.9917	LRR-RLP	0.85	(LRR-RLP)	0.9876
Y	Y	RLP	0.9971	RLP	0.8592	LRR-RLP	0.7994	(LRR-RLP)	0.9875

Y	Y	RLP	0.996	RLP	0.8597	LRR-RLP	0.833	(LRR-RLP)	0.9875
Y	Y	RLP	0.9964	RLP	0.9903	LRR-RLP	0.8485	(LRR-RLP)	0.9875
Y	Y	RLP	0.9965	RLP	0.8578	LRR-RLP	0.8497	(LRR-RLP)	0.9874
Y	Y	RLP	0.9963	RLP	0.9905	LRR-RLP	0.8491	(LRR-RLP)	0.9873
Y	Y	RLP	0.9968	RLP	0.9909	LRR-RLP	0.8517	(LRR-RLP)	0.9872
Y	Y	RLP	0.9968	RLP	0.8554	LRR-RLP	0.8337	(LRR-RLP)	0.9871
Y	Y	RLP	0.9963	RLP	0.9902	LRR-RLP	0.6662	(LRR-RLP)	0.9871
Y	Y	RLP	0.9968	RLP	0.9905	LRR-RLP	0.7171	(LRR-RLP)	0.9871
Y	Y	RLP	0.9963	RLP	0.9907	LRR-RLP	0.6671	(LRR-RLP)	0.9871
Y	Y	RLP	0.8973	RLP	0.9916	LRR-RLP	0.4337	(LRR-RLP)	0.987
Y	Y	RLP	0.9962	RLP	0.9905	LRR-RLP	0.8497	(LRR-RLP)	0.9869
Y	Y	RLP	0.9963	RLP	0.9925	LRR-RLP	0.8496	(LRR-RLP)	0.9869
Y	Y	RLP	0.9959	RLP	0.7173	LRR-RLP	0.7996	(LRR-RLP)	0.9868
Y	Y	RLP	0.9962	RLP	0.9894	LRR-RLP	0.8491	(LRR-RLP)	0.9868
Y	Y	RLP	0.9965	RLP	0.9907	LRR-RLP	0.8502	(LRR-RLP)	0.9868
Y	Y	RLP	0.9967	RLP	0.8521	LRR-RLP	0.8332	(LRR-RLP)	0.9867
Y	Y	RLP	0.948	RLP	0.9917	LRR-RLP	0.8338	(LRR-RLP)	0.9867
Y	Y	RLP	0.9969	RLP	0.9909	LRR-RLP	0.8497	(LRR-RLP)	0.9866
Y	Y	RLP	0.9971	RLP	0.9912	LRR-RLP	0.6671	(LRR-RLP)	0.9866
Y	Y	RLP	0.9473	RLP	0.858	LRR-RLP	0.8508	(LRR-RLP)	0.9866
Y	Y	RLP	0.996	RLP	0.991	LRR-RLP	0.8334	(LRR-RLP)	0.9866
Y	Y	RLP	0.9971	RLP	0.8539	LRR-RLP	0.8496	(LRR-RLP)	0.9866
Y	Y	RLP	0.9971	RLP	0.9922	LRR-RLP	0.6664	(LRR-RLP)	0.9866
Y	Y	RLP	0.9962	RLP	0.9913	LRR-RLP	0.8168	(LRR-RLP)	0.9866
Y	Y	RLP	0.9967	RLP	0.858	LRR-RLP	0.8496	(LRR-RLP)	0.9865
Y	Y	RLP	0.9968	RLP	0.8505	LRR-RLP	0.833	(LRR-RLP)	0.9864
Y	Y	RLP	0.9963	RLP	0.9904	LRR-RLP	0.8504	(LRR-RLP)	0.9864
Y	Y	RLP	0.9965	RLP	0.8564	LRR-RLP	0.833	(LRR-RLP)	0.9863
Y	Y	RLP	0.8923	RLP	0.576	LRR-RLP	0.8166	(LRR-RLP)	0.9862
Y	Y	RLP	0.9958	RLP	0.9912	LRR-RLP	0.7004	(LRR-RLP)	0.9862
Y	Y	RLP	0.9968	RLP	0.9905	LRR-RLP	0.8503	(LRR-RLP)	0.9861
Y	Y	RLP	0.9961	RLP	0.8515	LRR-RLP	0.8163	(LRR-RLP)	0.9861
Y	Y	RLP	0.9966	RLP	0.8542	LRR-RLP	0.8504	(LRR-RLP)	0.9861
Y	Y	RLP	0.997	RLP	0.9916	LRR-RLP	0.8163	(LRR-RLP)	0.9861
Y	Y	RLP	0.9968	RLP	0.9902	LRR-RLP	0.8496	(LRR-RLP)	0.9859
Y	Y	RLP	0.9962	RLP	0.9906	LRR-RLP	0.8333	(LRR-RLP)	0.9859
Y	Y	RLP	0.9968	RLP	0.9897	LRR-RLP	0.8507	(LRR-RLP)	0.9857
Y	Y	RLP	0.9966	RLP	0.9903	LRR-RLP	0.6996	(LRR-RLP)	0.9856
Y	Y	RLP	0.9963	RLP	0.9912	LRR-RLP	0.8498	(LRR-RLP)	0.9855
Y	Y	RLP	0.9968	RLP	0.9908	LRR-RLP	0.85	(LRR-RLP)	0.9855
Y	Y	RLP	0.9963	RLP	0.9911	Unknown	0.3165	(Unknown)	0.9855
Y	Y	RLP	0.9965	RLP	0.9906	LRR-RLP	0.8009	(LRR-RLP)	0.9852
Y	Y	RLP	0.996	RLP	0.99	LRR-RLP	0.6337	(LRR-RLP)	0.9844
Y	Y	RLP	0.996	RLP	0.9911	LRR-RLP	0.8157	(LRR-RLP)	0.9825
N	Y	RLP	0.9964	RLP	0.9908	LRR-RLP	0.8497	(LRR-RLP)	0.8062

N	Y	RLP	0.9968	RLP	0.9906	LRR-RLP	0.6496	(LRR-RLP)	0.805
N	Y	RLP	0.9962	RLP	0.9908	LRR-RLP	0.8329	(LRR-RLP)	0.8049
N	Y	RLP	0.9957	RLP	0.9905	LRR-RLP	0.3511	(LRR-RLP)	0.8049
N	Y	RLP	0.9966	RLP	0.7123	LRR-RLP	0.8488	(LRR-RLP)	0.8048
N	Y	RLP	0.9955	RLP	0.9902	LRR-RLP	0.3835	(LRR-RLP)	0.8043
N	Y	RLP	0.9966	RLP	0.9894	LRR-RLP	0.8332	(LRR-RLP)	0.8041
N	Y	RLP	0.7891	RLP	0.7116	LRR-RLP	0.5184	(LRR-RLP)	0.804
N	Y	RLP	0.9964	RLP	0.7185	LRR-RLP	0.8498	(LRR-RLP)	0.8038
N	Y	RLP	0.9965	RLP	0.991	LRR-RLP	0.849	(LRR-RLP)	0.8033
N	Y	RLP	0.9963	RLP	0.9908	LRR-RLP	0.6168	(LRR-RLP)	0.8031
N	Y	RLP	0.9963	RLP	0.9918	LRR-RLP	0.634	(LRR-RLP)	0.8031
N	Y	RLP	0.9963	RLP	0.9917	LRR-RLP	0.8163	(LRR-RLP)	0.8028
N	Y	RLP	0.9963	RLP	0.8557	LRR-RLP	0.85	(LRR-RLP)	0.8021
N	Y	RLP	0.9961	RLP	0.9912	LRR-RLP	0.4171	(LRR-RLP)	0.8018
N	Y	RLP	0.997	RLP	0.9921	LRR-RLP	0.851	(LRR-RLP)	0.8014
N	Y	RLP	0.9966	RLP	0.9914	LRR-RLP	0.8497	(LRR-RLP)	0.8014
N	Y	RLP	0.9968	RLP	0.991	LRR-RLP	0.6508	(LRR-RLP)	0.8013
N	Y	RLP	0.9959	RLP	0.9903	LRR-RLP	0.8505	(LRR-RLP)	0.8012
N	Y	RLP	0.996	RLP	0.9912	LRR-RLP	0.851	(LRR-RLP)	0.8011
N	Y	RLP	0.947	RLP	0.9906	LRR-RLP	0.6663	(LRR-RLP)	0.8011
N	Y	RLP	0.9961	RLP	0.7127	LRR-RLP	0.8489	(LRR-RLP)	0.8009
N	Y	RLP	0.9962	RLP	0.9892	LRR-RLP	0.8501	(LRR-RLP)	0.8007
N	Y	RLP	0.9968	RLP	0.8573	LRR-RLP	0.8164	(LRR-RLP)	0.8006
N	Y	RLP	0.9959	RLP	0.9913	LRR-RLP	0.8329	(LRR-RLP)	0.8004
N	Y	RLP	0.9966	RLP	0.9909	LRR-RLP	0.8494	(LRR-RLP)	0.8004
N	Y	RLP	0.9963	RLP	0.991	LRR-RLP	0.6495	(LRR-RLP)	0.8002
N	Y	RLP	0.9966	RLP	0.855	LRR-RLP	0.85	(LRR-RLP)	0.8001
N	Y	RLP	0.9968	RLP	0.9914	LRR-RLP	0.6665	(LRR-RLP)	0.8
N	Y	RLP	0.9961	RLP	0.8561	LRR-RLP	0.8491	(LRR-RLP)	0.8
N	Y	RLP	0.9959	RLP	0.7158	LRR-RLP	0.8509	(LRR-RLP)	0.7999
N	Y	RLP	0.9958	RLP	0.9901	LRR-RLP	0.5669	(LRR-RLP)	0.7999
N	Y	RLP	0.9967	RLP	0.9901	LRR-RLP	0.8489	(LRR-RLP)	0.7999
N	Y	RLP	0.9967	RLP	0.9912	LRR-RLP	0.6007	(LRR-RLP)	0.7991
N	Y	RLP	0.9962	RLP	0.9909	LRR-RLP	0.8512	(LRR-RLP)	0.7986
N	Y	RLP	0.9967	RLP	0.9893	LRR-RLP	0.8332	(LRR-RLP)	0.7986
N	Y	RLP	0.9473	RLP	0.9905	LRR-RLP	0.6664	(LRR-RLP)	0.7986
N	Y	RLP	0.9958	RLP	0.9911	LRR-RLP	0.8173	(LRR-RLP)	0.7985
N	Y	RLP	0.9963	RLP	0.9904	LRR-RLP	0.5822	(LRR-RLP)	0.7985
N	Y	RLP	0.9967	RLP	0.9915	LRR-RLP	0.6675	(LRR-RLP)	0.7984
N	Y	RLP	0.9963	RLP	0.8591	LRR-RLP	0.8174	(LRR-RLP)	0.7984
N	Y	RLP	0.9968	RLP	0.9908	LRR-RLP	0.6674	(LRR-RLP)	0.7983
N	Y	RLP	0.9964	RLP	0.9928	LRR-RLP	0.8505	(LRR-RLP)	0.7982
N	Y	RLP	0.9968	RLP	0.7155	LRR-RLP	0.8331	(LRR-RLP)	0.7981
N	Y	RLP	0.9457	RLP	0.9907	LRR-RLP	0.8507	(LRR-RLP)	0.798
N	Y	RLP	0.9968	RLP	0.9907	LRR-RLP	0.7002	(LRR-RLP)	0.798

N	Y	RLP	0.9968	RLP	0.8562	LRR-RLP	0.8509	(LRR-RLP)	0.7978
N	Y	RLP	0.9968	RLP	0.991	LRR-RLP	0.8329	(LRR-RLP)	0.7975
N	Y	RLP	0.9961	RLP	0.9909	LRR-RLP	0.4158	(LRR-RLP)	0.7975
N	Y	RLP	0.996	RLP	0.9898	LRR-RLP	0.8319	(LRR-RLP)	0.7974
N	Y	RLP	0.6299	RLP	0.9904	Other-RLP	0.3164	(Other-RLP)	0.7971
N	Y	RLP	0.9967	RLP	0.9909	LRR-RLP	0.8168	(LRR-RLP)	0.7971
N	Y	RLP	0.9961	RLP	0.8608	LRR-RLP	0.7661	(LRR-RLP)	0.7971
N	Y	RLP	0.9962	RLP	0.9919	LRR-RLP	0.8493	(LRR-RLP)	0.797
N	Y	RLP	0.7911	RLP	0.99	Other-RLP	0.3841	(Other-RLP)	0.797
N	Y	RLP	0.996	RLP	0.9916	LRR-RLP	0.6673	(LRR-RLP)	0.797
N	Y	RLP	0.9966	RLP	0.9915	LRR-RLP	0.6666	(LRR-RLP)	0.7969
N	Y	RLP	0.6834	RLP	0.7095	LRR-RLP	0.8495	(LRR-RLP)	0.7969
N	Y	RLP	0.996	RLP	0.9911	LRR-RLP	0.6671	(LRR-RLP)	0.7968
N	Y	RLP	0.996	RLP	0.9919	LRR-RLP	0.8508	(LRR-RLP)	0.7967
N	Y	RLP	0.9956	RLP	0.9901	LRR-RLP	0.7003	(LRR-RLP)	0.7966
N	Y	RLP	0.9964	RLP	0.9901	LRR-RLP	0.8497	(LRR-RLP)	0.7959
N	Y	RLP	0.9965	RLP	0.9887	LRR-RLP	0.8488	(LRR-RLP)	0.7958
N	Y	RLP	0.9958	RLP	0.9908	LRR-RLP	0.7502	(LRR-RLP)	0.7957
N	Y	RLP	0.9961	RLP	0.9911	LRR-RLP	0.6836	(LRR-RLP)	0.7951
N	Y	RLP	0.9969	RLP	0.99	LRR-RLP	0.8001	(LRR-RLP)	0.795
N	Y	RLP	0.9963	RLP	0.9896	LRR-RLP	0.8499	(LRR-RLP)	0.7947
N	Y	RLP	0.9962	RLP	0.9922	LRR-RLP	0.8168	(LRR-RLP)	0.7939
N	Y	RLP	0.9962	RLP	0.9909	LRR-RLP	0.6679	(LRR-RLP)	0.7938
N	Y	RLP	0.9964	RLP	0.9919	LRR-RLP	0.8502	(LRR-RLP)	0.7936
N	Y	RLP	0.9959	RLP	0.991	LRR-RLP	0.6161	(LRR-RLP)	0.7935
N	Y	RLP	0.9969	RLP	0.9906	LRR-RLP	0.6662	(LRR-RLP)	0.7934
N	Y	RLP	0.996	RLP	0.9898	LRR-RLP	0.8497	(LRR-RLP)	0.7922
N	Y	RLP	0.9961	RLP	0.9901	LRR-RLP	0.8505	(LRR-RLP)	0.792
Y	N	RLP	0.9499	RLP	0.8572	LRR-RLP	0.7006	noRLP	0.2033
N	N	RLP	0.9468	RLP	0.8516	Unknown	0.2667	noRLP	0.2008
Y	N	RLP	0.9468	RLP	0.8537	LRR-RLP	0.5839	noRLP	0.1974
N	N	noRLP	0.1575	RLP	0.5637	Ethylene-responsive-RLP	0.4992	noRLP	0.0012

Table S2. Higth throughput prediction of the Arabidopsis proteins of RLPrefictiOme

Accession	SM	TM	RLP-noRLP	RLP-noRLP Probability	RLP-RLK	RLP-RLK Probability	RLP-Subfamily	RLP-Subfamily Probability	Classification	Decision probability	Description
AT1G65380.1	Y	Y	RLP	0.9962	RLP	0.9907	LRR-RLP	0.8505	(LRR-RLP)	0.9902	**AtRLP10**
AT1G17240.1	Y	Y	RLP	0.9962	RLP	0.9913	LRR-RLP	0.8497	(LRR-RLP)	0.9886	**AtRLP2**
AT4G18760.1	Y	Y	RLP	0.9967	RLP	0.9903	LRR-RLP	0.8495	(LRR-RLP)	0.9885	**AtRLP51**
AT4G13880.1	Y	Y	RLP	0.9963	RLP	0.9899	LRR-RLP	0.8001	(LRR-RLP)	0.9884	**AtRLP48**
AT5G27060.1	Y	Y	RLP	0.9962	RLP	0.991	LRR-RLP	0.6669	(LRR-RLP)	0.9884	**AtRLP53**
AT3G23110.1	Y	Y	RLP	0.9964	RLP	0.9912	LRR-RLP	0.6502	(LRR-RLP)	0.9883	**AtRLP37**
AT1G80080.1	Y	Y	RLP	0.9961	RLP	0.9911	LRR-RLP	0.5506	(LRR-RLP)	0.9883	**AtRLP17**
AT2G32680.1	Y	Y	RLP	0.9967	RLP	0.9918	LRR-RLP	0.7838	(LRR-RLP)	0.9882	**AtRLP23**
AT3G11080.1	Y	Y	RLP	0.9962	RLP	0.991	LRR-RLP	0.8496	(LRR-RLP)	0.988	**AtRLP35**
AT1G74180.1	Y	Y	RLP	0.9959	RLP	0.858	LRR-RLP	0.8163	(LRR-RLP)	0.988	**AtRLP14**
AT3G05370.1	Y	Y	RLP	0.9962	RLP	0.8556	LRR-RLP	0.6337	(LRR-RLP)	0.988	**AtRLP31**
AT3G28890.1	Y	Y	RLP	0.9966	RLP	0.8561	LRR-RLP	0.6336	(LRR-RLP)	0.988	**AtRLP43**
AT5G45770.1	Y	Y	RLP	0.9965	RLP	0.99	LRR-RLP	0.683	(LRR-RLP)	0.9878	**AtRLP55**
AT2G25440.1	Y	Y	RLP	0.9962	RLP	0.9902	LRR-RLP	0.4832	(LRR-RLP)	0.9878	**AtRLP20**
AT5G65830.1	Y	Y	RLP	0.9966	RLP	0.8566	LRR-RLP	0.667	(LRR-RLP)	0.9876	**ATRLP57**
AT3G05360.1	Y	Y	RLP	0.9967	RLP	0.9913	LRR-RLP	0.6668	(LRR-RLP)	0.9876	**AtRLP30**
AT2G42800.1	Y	Y	RLP	0.9963	RLP	0.9908	LRR-RLP	0.6665	(LRR-RLP)	0.9876	**AtRLP29**
AT5G37360.1	Y	Y	RLP	0.6322	RLP	0.9916	LRR-RLP	0.4	(LRR-RLP)	0.9875	Subfamily not named and unknown function
AT2G33020.1	Y	Y	RLP	0.9966	RLP	0.9905	LRR-RLP	0.8161	(LRR-RLP)	0.9873	**AtRLP24**
AT1G74190.1	Y	Y	RLP	0.9959	RLP	0.8564	LRR-RLP	0.8499	(LRR-RLP)	0.9871	**AtRLP15**
AT2G15080.1	Y	Y	RLP	0.9965	RLP	0.9904	LRR-RLP	0.8502	(LRR-RLP)	0.987	**AtRLP19**
AT5G19230.1	Y	Y	RLP	0.9969	RLP	0.9904	LRR-RLP	0.4334	(LRR-RLP)	0.987	Subfamily not named and unknown function
AT1G45616.1	Y	Y	RLP	0.9961	RLP	0.9913	LRR-RLP	0.7665	(LRR-RLP)	0.9868	**AtRLP6**

AT3G05650.1	Y	Y	RLP	0.9964	RLP	0.9906	LRR-RLP	0.6664	(LRR-RLP)	0.9868	**AtRLP32**
AT3G05660.1	Y	Y	RLP	0.9966	RLP	0.8557	LRR-RLP	0.85	(LRR-RLP)	0.9866	**AtRLP33**
AT1G58190.1	Y	Y	RLP	0.9962	RLP	0.8521	LRR-RLP	0.6663	(LRR-RLP)	0.9866	**AtRLP9**
AT4G13920.1	Y	Y	RLP	0.9967	RLP	0.9911	LRR-RLP	0.8498	(LRR-RLP)	0.9865	**AtRLP50**
AT3G49750.1	Y	Y	RLP	0.9963	RLP	0.9909	LRR-RLP	0.7502	(LRR-RLP)	0.9865	**AtRLP44**
AT5G25910.1	Y	Y	RLP	0.9964	RLP	0.9899	LRR-RLP	0.8501	(LRR-RLP)	0.9864	**AtRLP52**
AT4G04220.1	Y	Y	RLP	0.9962	RLP	0.9911	LRR-RLP	0.8506	(LRR-RLP)	0.9863	**AtRLP46**
AT2G33060.1	Y	Y	RLP	0.9966	RLP	0.9914	LRR-RLP	0.8332	(LRR-RLP)	0.9863	**AtRLP27**
AT2G33050.1	Y	Y	RLP	0.9964	RLP	0.9915	LRR-RLP	0.7498	(LRR-RLP)	0.986	**AtRLP26**
AT4G28560.1	Y	Y	RLP	0.9965	RLP	0.9908	LRR-RLP	0.6164	(LRR-RLP)	0.986	**RIC7**
AT1G71400.1	Y	Y	RLP	0.996	RLP	0.8563	LRR-RLP	0.6831	(LRR-RLP)	0.9851	**AtRLP12**
AT1G71390.1	N	Y	RLP	0.9966	RLP	0.99	LRR-RLP	0.6667	(LRR-RLP)	0.8021	**AtRLP11**
AT2G25470.1	N	Y	RLP	0.9964	RLP	0.8556	LRR-RLP	0.8502	(LRR-RLP)	0.8014	**AtRLP21**
AT4G13810.1	N	Y	RLP	0.9964	RLP	0.9907	LRR-RLP	0.833	(LRR-RLP)	0.7997	**AtRLP47**
AT3G23010.1	N	Y	RLP	0.9965	RLP	0.9908	LRR-RLP	0.667	(LRR-RLP)	0.7995	**AtRLP36**
AT3G24982.1	N	Y	RLP	0.9963	RLP	0.989	LRR-RLP	0.8512	(LRR-RLP)	0.7993	**AtRLP40**
AT4G25750.1	N	Y	RLP	0.7363	RLP	0.9922	LRR-RLP	0.4168	(LRR-RLP)	0.7992	ABC-2 type transporter domain containing protein. expressed
AT1G17250.1	N	Y	RLP	0.9965	RLP	0.9911	LRR-RLP	0.8496	(LRR-RLP)	0.799	**AtRLP3**
AT3G21580.1	N	Y	RLP	0.8424	RLP	0.9903	LRR-RLP	0.3832	(LRR-RLP)	0.7981	cobalt ion transporter. putative. expressed
AT3G23120.1	N	Y	RLP	0.997	RLP	0.9905	LRR-RLP	0.6835	(LRR-RLP)	0.7976	**AtRLP38**
AT3G53240.1	N	Y	RLP	0.9961	RLP	0.9905	LRR-RLP	0.783	(LRR-RLP)	0.7973	**AtRLP45**
AT1G07390.1	N	Y	RLP	0.9957	RLP	0.7119	LRR-RLP	0.7826	(LRR-RLP)	0.7969	**AtRLP1**
AT3G11010.1	N	Y	RLP	0.9961	RLP	0.9902	LRR-RLP	0.6665	(LRR-RLP)	0.7958	**AtRLP34**
AT3G44070.1	N	Y	RLP	0.9961	RLP	0.9899	LRR-RLP	0.5335	(LRR-RLP)	0.7951	Subfamily not named and unknown function
AT5G49290.1	N	Y	RLP	0.9966	RLP	0.9901	LRR-RLP	0.6833	(LRR-RLP)	0.7941	**ATRLP56**
AT1G34290.1	Y	Y	RLP	0.9964	RLP	0.9898	**(Undefined)**	0.2166	**(Undefined)**	0.7949	**AtRLP5**

AT1G74170.1	N	Y	RLP	0.9964	RLP	0.8561	LRR-RLP	0.7164	(LRR-RLP)	0.7994	**AtRLP13**
AT1G47890.1	N	Y	RLP	0.9967	RLP	0.9908	LRR-RLP	0.8501	(LRR-RLP)	0.8001	**AtRLP7**
AT2G20520.1	Y	Y	RLP	0.8429	RLP	0.9908	LysM-RLP	0.6839	(LysM-RLP)	0.9885	fasciclin domain containing protein. expressed
AT2G48130.1	Y	Y	RLP	0.996	RLP	0.991	LysM-RLP	0.3337	(LysM-RLP)	0.9883	LTPL78 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT2G27130.1	Y	Y	RLP	0.8431	RLP	0.9899	LysM-RLP	0.3001	(LysM-RLP)	0.9881	LTPL78 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT5G64080.1	Y	Y	RLP	0.7392	RLP	0.9901	LysM-RLP	0.4837	(LysM-RLP)	0.988	LTPL78 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT2G04780.1	Y	Y	RLP	0.8962	RLP	0.99	LysM-RLP	0.3665	(LysM-RLP)	0.9878	fasciclin domain containing protein. expressed
AT1G77630.1	Y	Y	RLP	0.9457	RLP	0.9912	LysM-RLP	0.5002	(LysM-RLP)	0.9876	**LysM-RLP**
AT4G14815.1	Y	Y	RLP	0.9475	RLP	0.9905	LysM-RLP	0.3501	(LysM-RLP)	0.9876	LTPL76 - Protease inhibitor / Other-RLP (Probable-lipid-transfer-RLK)
AT2G45470.1	Y	Y	RLP	0.9466	RLP	0.9911	LysM-RLP	0.5328	(LysM-RLP)	0.9875	fasciclin-like arabinogalactan protein. putative. expressed
AT2G32300.1	Y	Y	RLP	0.7903	RLP	0.9904	LysM-RLP	0.3502	(LysM-RLP)	0.9875	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT3G46550.1	Y	Y	RLP	0.9961	RLP	0.9896	LysM-RLP	0.3003	(LysM-RLP)	0.9874	fasciclin-like arabinogalactan protein. putative. expressed
AT3G06360.1	Y	Y	RLP	0.9963	RLP	0.9911	LysM-RLP	0.4834	(LysM-RLP)	0.9873	Subfamily not named and unknown function
AT4G14805.1	Y	Y	RLP	0.9962	RLP	0.9908	LysM-RLP	0.5666	(LysM-RLP)	0.9871	LTPL47 - Protease inhibitor / Other-RLP (Probable-lipid-transfer-RLK)
AT1G63550.1	Y	Y	RLP	0.9962	RLP	0.9906	LysM-RLP	0.4004	(LysM-RLP)	0.987	Salt stress response/antifungal
AT2G44300.1	Y	Y	RLP	0.7878	RLP	0.9888	LysM-RLP	0.3835	(LysM-RLP)	0.9868	LTPL82 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT4G22666.1	Y	Y	RLP	0.7913	RLP	0.9917	LysM-RLP	0.3499	(LysM-RLP)	0.9867	LTPL85 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT4G12360.1	Y	Y	RLP	0.996	RLP	0.9905	LysM-RLP	0.5332	(LysM-RLP)	0.9865	LTPL85 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT1G73890.1	Y	Y	RLP	0.9479	RLP	0.9913	LysM-RLP	0.5662	(LysM-RLP)	0.9862	LTPL82 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT1G73550.1	Y	Y	RLP	0.946	RLP	0.9903	Unknown	0.2503	(LysM-RLP)	0.7999	LTPL85 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT5G26270.1	N	Y	RLP	0.7363	RLP	0.991	LysM-RLP	0.4834	(LysM-RLP)	0.7991	Subfamily not named and unknown function
AT2G17120.1	Y	Y	RLP	0.9959	RLP	0.9904	Unknown	0.2992	(LysM-RLP)	0.8015	**LysM-RLP**

AT2G19440.1	Y	Y	RLP	0.6853	RLP	0.7135	L-Lectin-RLP	0.3661	(L-Lectin-RLP)	0.9882	Glycosyl hydrolases family 17 (Glycosyl-hydrolases-RLP)
AT1G16022.1	Y	Y	RLP	0.8955	RLP	0.9902	L-Lectin-RLP	0.3341	(L-Lectin-RLP)	0.9875	Subfamily not named and unknown function
AT3G26110.1	Y	Y	RLP	0.8426	RLP	0.9901	L-Lectin-RLP	0.3501	(L-Lectin-RLP)	0.9872	Subfamily not named and unknown function
AT1G70170.1	Y	Y	RLP	0.9961	RLP	0.9912	L-Lectin-RLP	0.4999	(L-Lectin-RLP)	0.987	metalloendoproteinase 1 precursor. putative. expressed
AT4G31840.1	Y	Y	RLP	0.7364	RLP	0.9909	L-Lectin-RLP	0.3835	(L-Lectin-RLP)	0.9867	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT5G40620.1	Y	Y	RLP	0.9967	RLP	0.9901	L-Lectin-RLP	0.4836	(L-Lectin-RLP)	0.9866	Subfamily not named and unknown function
AT1G05140.1	N	Y	RLP	0.6822	RLP	0.7146	L-Lectin-RLP	0.3668	(L-Lectin-RLP)	0.8001	peptidase M50 family protein. putative. expressed
ATMG00410.1	N	Y	RLP	0.6299	RLP	0.9898	L-Lectin-RLP	0.3998	(L-Lectin-RLP)	0.7992	ATP synthase. A subunit family protein. putative. expressed
AT2G07741.1	N	Y	RLP	0.6325	RLP	0.9911	L-Lectin-RLP	0.3999	(L-Lectin-RLP)	0.7989	ATP synthase. A subunit family protein. putative. expressed
AT2G32480.1	N	Y	RLP	0.6305	RLP	0.9901	L-Lectin-RLP	0.4333	(L-Lectin-RLP)	0.7988	peptidase M50 family protein. putative. expressed
AT2G24945.1	N	Y	RLP	0.6837	RLP	0.9909	L-Lectin-RLP	0.3664	(L-Lectin-RLP)	0.7982	Subfamily not named and unknown function
AT5G37480.1	N	Y	RLP	0.7367	RLP	0.9911	L-Lectin-RLP	0.3835	(L-Lectin-RLP)	0.7933	Subfamily not named and unknown function
AT1G25570.1	Y	Y	RLP	0.9966	RLP	0.9902	Malectin-RLP	0.4674	(Malectin-RLP)	0.9884	Subfamily not named and unknown function
AT4G16120.1	Y	Y	RLP	0.9966	RLP	0.9915	Malectin-RLP	0.3004	(Malectin-RLP)	0.988	COBRA-like protein 7 precursor. putative. expressed
AT4G12420.1	Y	Y	RLP	0.6821	RLP	0.9895	Malectin-RLP	0.4832	(Malectin-RLP)	0.9877	monocopper oxidase. putative. expressed
AT1G28340.1	Y	Y	RLP	0.8425	RLP	0.9905	Malectin-RLP	0.4502	(Malectin-RLP)	0.9875	**Malectin-RLP**
AT3G46270.1	Y	Y	RLP	0.9964	RLP	0.9892	Malectin-RLP	0.5168	(Malectin-RLP)	0.9873	Subfamily not named and unknown function
AT4G18340.1	Y	Y	RLP	0.7893	RLP	0.9907	Malectin-RLP	0.333	(Malectin-RLP)	0.987	glycosyl hydrolases family 17. putative. expressed (Glycosyl-hydrolases-RLP)
AT4G25240.1	Y	Y	RLP	0.7897	RLP	0.9903	Malectin-RLP	0.3995	(Malectin-RLP)	0.9867	monocopper oxidase. putative. expressed
AT5G51480.1	Y	Y	RLP	0.9963	RLP	0.9902	Malectin-RLP	0.4167	(Malectin-RLP)	0.9866	monocopper oxidase. putative. expressed
AT3G46280.1	Y	Y	RLP	0.9966	RLP	0.9911	Malectin-RLP	0.4335	(Malectin-RLP)	0.9853	Subfamily not named and unknown function
AT5G49150.1	N	Y	RLP	0.8938	RLP	0.9914	Malectin-RLP	0.3336	(Malectin-RLP)	0.7986	GEX2. putative. expressed
AT2G04060.1	N	Y	RLP	0.6831	RLP	0.9898	Malectin-RLP	0.4003	(Malectin-RLP)	0.7933	beta-galactosidase precursor. putative. expressed (Glycosyl-hydrolases-RLP)
AT1G24485.1	Y	Y	RLP	0.9963	RLP	0.9914	Unknown	0.2172	(Malectin-RLP)	0.7941	**Malectin-RLP**

AT1G10380.1	Y	Y	RLP	0.9964	RLP	0.9897	Thaumatin-RLP	0.4165	(Thaumatin-RLP)	0.988	**Thaumatin-RLP**
AT4G38660.1	Y	Y	RLP	0.8941	RLP	0.9906	Thaumatin-RLP	0.634	(Thaumatin-RLP)	0.9876	**Thaumatin-RLP**
AT4G36010.1	Y	Y	RLP	0.84	RLP	0.9902	Thaumatin-RLP	0.7497	(Thaumatin-RLP)	0.9875	**Thaumatin-RLP**
AT1G75800.1	Y	Y	RLP	0.9964	RLP	0.9893	Thaumatin-RLP	0.8491	(Thaumatin-RLP)	0.9871	**Thaumatin-RLP**
AT1G04520.1	Y	Y	RLP	0.8433	RLP	0.9905	Thaumatin-RLP	0.3664	(Thaumatin-RLP)	0.987	**Thaumatin-RLP**
AT1G77700.1	N	Y	RLP	0.9966	RLP	0.9904	Thaumatin-RLP	0.5666	(Thaumatin-RLP)	0.7974	**Thaumatin-RLP**
AT2G20700.1	Y	Y	RLP	0.9967	RLP	0.9899	Salt-stress-response/ antifungal-RLP	0.3336	(Salt-stress-response/ antifungal-RLP)	0.9856	Subfamily not named and unknown function
AT3G60720.1	Y	Y	RLP	0.997	RLP	0.9908	Unknown	0.1832	(Salt-stress-response/ antifungal-RLP)	0.7982	**Salt stress response/antifungal-RLP**
AT2G33330.1	Y	Y	RLP	0.7908	RLP	0.9898	Unknown	0.2498	(Salt-stress-response/ antifungal-RLP)	0.7989	**Salt stress response/antifungal-RLP**
AT1G61750.1	Y	Y	RLP	0.9963	RLP	0.9901	Salt-stress-response/ antifungal-RLP	0.3334	(Salt-stress-response/ antifungal-RLP)	0.9862	**Salt stress response/antifungal-RLP**
AT5G53110.1	Y	Y	RLP	0.9957	RLP	0.9907	WAK-RLP	0.3333	(WAK-RLP)	0.989	**WAK-RLP**
AT4G14746.1	Y	Y	RLP	0.8375	RLP	0.991	WAK-RLP	0.3176	(WAK-RLP)	0.9885	Subfamily not named and unknown function
AT2G30290.1	Y	Y	RLP	0.6813	RLP	0.9904	WAK-RLP	0.3164	(WAK-RLP)	0.9879	vacuolar-sorting receptor precursor. putative. expressed
AT1G11915.1	Y	Y	RLP	0.9966	RLP	0.9913	WAK-RLP	0.5163	(WAK-RLP)	0.9877	**WAK-RLP**
AT2G46494.1	Y	Y	RLP	0.9964	RLP	0.9908	WAK-RLP	0.3334	(WAK-RLP)	0.9874	zinc finger. C3HC4 type domain containing protein. expressed
AT5G37660.1	Y	Y	RLP	0.9963	RLP	0.9895	WAK-RLP	0.317	(WAK-RLP)	0.987	Salt stress response/antifungal-RLP
AT1G02300.1	Y	Y	RLP	0.6844	RLP	0.8582	WAK-RLP	0.2334	(WAK-RLP)	0.8005	Papain family cysteine protease domain containing protein. expressed
AT1G66940.1	Y	Y	RLP	0.9962	RLP	0.9898	WAK-RLP	0.2665	(WAK-RLP)	0.7981	**WAK-RLP**

AT1G74045.1	N	Y	RLP	0.8418	RLP	0.9915	WAK-RLP	0.4006	(WAK-RLP)	0.7964	tetraspanin family protein. putative. expressed
AT1G70690.1	Y	Y	RLP	0.9964	RLP	0.9906	Unknown	0.2993	(WAK-RLP)	0.7949	Salt stress response/antifungal-RLP
AT5G43980.1	Y	Y	RLP	0.9965	RLP	0.9902	Unknown	0.2672	(WAK-RLP)	0.7911	Salt stress response/antifungal-RLP
AT2G46495.1	Y	Y	RLP	0.8945	RLP	0.9908	WAK-RLP	0.3003	(WAK-RLP)	0.9859	**WAK-RLP**
AT2G46495.1	Y	Y	RLP	0.8945	RLP	0.9908	Unknown	0.3003	(WAK-RLP)	0.9859	**WAK-RLP**
AT5G03700.1	Y	Y	RLP	0.7348	RLP	0.8555	S-domain-RLP	0.4498	(S-domain-RLP)	0.9878	**(S-domain-RLP)**
AT1G46840.1	N	Y	RLP	0.6293	RLP	0.9913	S-domain-RLP	0.4168	(S-domain-RLP)	0.7988	OsFBO21 - F-box and other domain containing protein. expressed
AT1G48940.1	Y	Y	RLP	0.8424	RLP	0.9902	Other-RLP	0.5169	(Other-RLP)	0.9888	**plastocyanin-like domain containing protein. putative. expressed (Plastocyanin-like-RLK)**
AT1G69980.1	Y	Y	RLP	0.684	RLP	0.992	Other-RLP	0.6165	(Other-RLP)	0.9884	Subfamily not named and unknown function
AT5G14030.1	Y	Y	RLP	0.6827	RLP	0.9902	Other-RLP	0.5163	(Other-RLP)	0.9883	translocon-associated protein beta domain containing protein. expressed
AT3G51580.1	Y	Y	RLP	0.7894	RLP	0.9908	Other-RLP	0.3504	(Other-RLP)	0.9882	Subfamily not named and unknown function
AT5G66160.1	Y	Y	RLP	0.6327	RLP	0.9913	Other-RLP	0.3168	(Other-RLP)	0.9877	**RING finger protein 13 / Other-RLK (Ring finger-RLK)**
AT3G51710.1	Y	Y	RLP	0.9453	RLP	0.9913	Other-RLP	0.4495	(Other-RLP)	0.9876	D-mannose binding lectin protein with Apple-like carbohydrate-binding domain (B_Lectin-RLP)
AT2G44290.1	Y	Y	RLP	0.7367	RLP	0.9896	Other-RLP	0.3168	(Other-RLP)	0.9875	**LTPL82 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK**
AT1G35630.1	Y	Y	RLP	0.6857	RLP	0.9908	Other-RLP	0.4832	(Other-RLP)	0.9874	**RING finger protein 13. putative. expressed (Ring finger-RLK)**
AT5G66820.1	Y	Y	RLP	0.7932	RLP	0.9904	Other-RLP	0.4829	(Other-RLP)	0.9874	Subfamily not named and unknown function
AT3G53490.1	Y	Y	RLP	0.63	RLP	0.991	Other-RLP	0.3332	(Other-RLP)	0.9874	Subfamily not named and unknown function
AT4G26690.1	Y	Y	RLP	0.9479	RLP	0.9896	Other-RLP	0.3832	(Other-RLP)	0.9873	**glycerophosphoryl diester phosphodiesterase family protein. putative. expressed**
AT5G16660.1	Y	Y	RLP	0.634	RLP	0.9915	Other-RLP	0.3831	(Other-RLP)	0.9872	Subfamily not named and unknown function
AT3G20520.1	Y	Y	RLP	0.8425	RLP	0.9913	Other-RLP	0.351	(Other-RLP)	0.9872	**glycerophosphoryl diester phosphodiesterase family protein. putative. expressed**
AT2G26600.1	Y	Y	RLP	0.6849	RLP	0.7123	Other-RLP	0.3501	(Other-RLP)	0.9872	**glycosyl hydrolases family 17. putative. expressed (Glycosyl-hydrolases-RLP)**
AT2G32670.1	Y	Y	RLP	0.7367	RLP	0.8563	Other-RLP	0.416	(Other-RLP)	0.987	vesicle-associated membrane protein. putative. expressed

AT3G50050.1	Y	Y	RLP	0.8937	RLP	0.8558	Other-RLP	0.3665	(Other-RLP)	0.9867	eukaryotic aspartyl protease domain containing protein. expressed
AT5G42370.1	Y	Y	RLP	0.6338	RLP	0.9905	Other-RLP	0.483	(Other-RLP)	0.9866	Subfamily not named and unknown function
AT1G59970.1	Y	Y	RLP	0.6337	RLP	0.9911	Other-RLP	0.3997	(Other-RLP)	0.9866	metalloendoproteinase 1 precursor. putative. expressed
AT4G23720.1	Y	Y	RLP	0.6818	RLP	0.9901	Other-RLP	0.3495	(Other-RLP)	0.9866	Subfamily not named and unknown function
AT1G66970.1	Y	Y	RLP	0.8952	RLP	0.9901	Other-RLP	0.3996	(Other-RLP)	0.9865	**glycerophosphoryl diester phosphodiesterase family protein. putative. expressed**
AT5G58050.1	Y	Y	RLP	0.6339	RLP	0.9911	Other-RLP	0.3171	(Other-RLP)	0.9859	**glycerophosphoryl diester phosphodiesterase family protein. putative. expressed**
AT5G55480.1	Y	Y	RLP	0.7973	RLP	0.8751	Other-RLP	0.352	(Other-RLP)	0.9858	**glycerophosphoryl diester phosphodiesterase family protein. putative. expressed**
AT5G46850.1	N	Y	RLP	0.6337	RLP	0.9898	Other-RLP	0.3503	(Other-RLP)	0.8033	Subfamily not named and unknown function
AT5G42860.1	N	Y	RLP	0.8408	RLP	0.9911	Other-RLP	0.4662	(Other-RLP)	0.8032	Subfamily not named and unknown function
AT5G54170.1	N	Y	RLP	0.6817	RLP	0.9901	Other-RLP	0.3671	(Other-RLP)	0.802	membrane related protein CP5. putative. expressed
AT1G14490.1	N	Y	RLP	0.8419	RLP	0.991	Other-RLP	0.4336	(Other-RLP)	0.8009	Subfamily not named and unknown function
AT1G72820.1	N	Y	RLP	0.6787	RLP	0.7161	Other-RLP	0.3667	(Other-RLP)	0.8008	mitochondrial carrier protein. putative. expressed
ATCG00540.1	N	Y	RLP	0.6866	RLP	0.9906	Other-RLP	0.5661	(Other-RLP)	0.8002	apocytochrome f precursor. putative. expressed
AT4G17905.1	N	Y	RLP	0.6841	RLP	0.8584	Other-RLP	0.3333	(Other-RLP)	0.8	**RING-H2 finger protein ATL5G / Other-RLK (Ring finger-RLK)**
AT5G47180.1	N	Y	RLP	0.6861	RLP	0.9895	Other-RLP	0.4828	(Other-RLP)	0.7997	MSP domain containing protein. expressed
AT5G41530.1	N	Y	RLP	0.6812	RLP	0.9903	Other-RLP	0.3337	(Other-RLP)	0.7996	Subfamily not named and unknown function
AT5G63780.1	N	Y	RLP	0.8449	RLP	0.9903	Other-RLP	0.417	(Other-RLP)	0.7992	zinc finger. C3HC4 type domain containing protein. expressed
AT4G25030.1	N	Y	RLP	0.6286	RLP	0.992	Other-RLP	0.4671	(Other-RLP)	0.7991	Subfamily not named and unknown function
AT2G20590.1	N	Y	RLP	0.6826	RLP	0.9912	Other-RLP	0.5665	(Other-RLP)	0.7989	reticulon domain containing protein. putative. expressed
AT3G20270.1	N	Y	RLP	0.7896	RLP	0.8535	Other-RLP	0.3163	(Other-RLP)	0.7989	BPI/LBP family protein At3g20270 precursor. putative. expressed
AT1G74730.1	N	Y	RLP	0.631	RLP	0.9903	Other-RLP	0.3012	(Other-RLP)	0.7989	Subfamily not named and unknown function
AT2G30505.1	N	Y	RLP	0.7873	RLP	0.9898	Other-RLP	0.4996	(Other-RLP)	0.7988	Subfamily not named and unknown function
AT2G33110.1	N	Y	RLP	0.7364	RLP	0.9917	Other-RLP	0.4005	(Other-RLP)	0.7986	vesicle-associated membrane protein. putative. expressed
AT4G22890.1	N	Y	RLP	0.7345	RLP	0.9915	Other-RLP	0.3502	(Other-RLP)	0.7984	Subfamily not named and unknown function
AT1G78880.1	N	Y	RLP	0.8948	RLP	0.9901	Other-RLP	0.4337	(Other-RLP)	0.7977	Subfamily not named and unknown function

AT5G59400.1	N	Y	RLP	0.7376	RLP	0.9901	Other-RLP	0.5168	(Other-RLP)	0.7974	Subfamily not named and unknown function
AT5G63050.1	N	Y	RLP	0.7383	RLP	0.9903	Other-RLP	0.4166	(Other-RLP)	0.7974	Subfamily not named and unknown function
AT1G67540.1	N	Y	RLP	0.6852	RLP	0.9919	Other-RLP	0.3664	(Other-RLP)	0.7973	Subfamily not named and unknown function
AT5G52980.1	N	Y	RLP	0.8419	RLP	0.9918	Other-RLP	0.3996	(Other-RLP)	0.797	Subfamily not named and unknown function
AT1G76490.1	N	Y	RLP	0.6824	RLP	0.9889	Other-RLP	0.5165	(Other-RLP)	0.7963	3-hydroxy-3-methylglutaryl-coenzyme A reductase. putative. expressed
AT2G34380.1	N	Y	RLP	0.6318	RLP	0.9913	Other-RLP	0.3669	(Other-RLP)	0.7951	Subfamily not named and unknown function
AT5G48830.1	N	Y	RLP	0.6317	RLP	0.9908	Other-RLP	0.3667	(Other-RLP)	0.7922	Subfamily not named and unknown function
AT1G48510.1	N	Y	RLP	0.6308	RLP	0.9906	Other-RLP	0.5332	(Other-RLP)	0.7916	SURF1. putative. expressed
AT5G64930.1	N	Y	RLP	0.841	RLP	0.9906	Other-RLP	0.4667	(Other-RLP)	0.7913	Subfamily not named and unknown function
AT4G35170.1	N	Y	RLP	0.7899	RLP	0.9912	Other-RLP	0.3162	(Other-RLP)	0.7986	Subfamily not named and unknown function
AT3G27410.1	Y	Y	RLP	0.996	RLP	0.9911	Ethylene-responsive-RLP	0.3665	(Ethylene-responsive-RLP)	0.9874	Subfamily not named and unknown function
AT2G28440.1	Y	Y	RLP	0.8409	RLP	0.9912	Ethylene-responsive-RLP	0.466	(Ethylene-responsive-RLP)	0.9869	Subfamily not named and unknown function
AT1G30515.1	N	Y	RLP	0.6846	RLP	0.9906	Ethylene-responsive-RLP	0.3163	(Ethylene-responsive-RLP)	0.7923	Subfamily not named and unknown function
AT3G01070.1	Y	Y	RLP	0.6867	RLP	0.9907	Glycosyl-hydrolases-RLP	0.3661	(Glycosyl-hydrolases-RLP)	0.9874	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT5G14345.1	Y	Y	RLP	0.735	RLP	0.99	Glycosyl-hydrolases-RLP	0.3342	(Glycosyl-hydrolases-RLP)	0.9873	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT2G44520.1	N	Y	RLP	0.6813	RLP	0.9901	Glycosyl-hydrolases-RLP	0.4503	(Glycosyl-hydrolases-RLP)	0.8024	prenyltransferase. putative. expressed
AT1G35880.1	N	Y	RLP	0.844	RLP	0.9909	Glycosyl-hydrolases-RLP	0.3334	(Glycosyl-hydrolases-RLP)	0.8005	Subfamily not named and unknown function
AT5G66450.1	N	Y	RLP	0.7877	RLP	0.9904	Glycosyl-hydrolases-RLP	0.3835	(Glycosyl-hydrolases-RLP)	0.7993	Subfamily not named and unknown function

AT5G57345.1	N	Y	RLP	0.7378	RLP	0.9897	Glycosyl-hydrolases-RLP	0.3327	(Glycosyl-hydrolases-RLP)	0.7973	Subfamily not named and unknown function
AT5G19750.1	N	Y	RLP	0.682	RLP	0.99	Glycosyl-hydrolases-RLP	0.3833	(Glycosyl-hydrolases-RLP)	0.7963	Mpv17 / PMP22 family domain containing protein. expressed
AT2G39805.1	N	Y	RLP	0.8952	RLP	0.9907	Glycosyl-hydrolases-RLP	0.4001	(Glycosyl-hydrolases-RLP)	0.7949	Yip1 domain containing protein. expressed
AT2G42390.1	Y	Y	RLP	0.6859	RLP	0.992	PAN-RLP	0.334	(PAN-RLP)	0.9879	glucosidase II beta subunit-like domain containing protein. expressed
AT3G03726.1	N	Y	RLP	0.79	RLP	0.9896	PAN-RLP	0.3661	(PAN-RLP)	0.7988	Subfamily not named and unknown function
		Y									
AT3G27200.1	Y	Y	RLP	0.787	RLP	0.9903	RCC1-RLP	0.4167	(RCC1-RLP)	0.989	Subfamily not named and unknown function
AT4G01575.1	Y	Y	RLP	0.8394	RLP	0.991	RCC1-RLP	0.383	(RCC1-RLP)	0.9871	Subfamily not named and unknown function
AT2G01660.1	Y	Y	RLP	0.9969	RLP	0.9906	RCC1-RLP	0.3166	(RCC1-RLP)	0.9869	Salt stress response/antifungal-RLP
AT2G27389.1	Y	Y	RLP	0.7899	RLP	0.9915	RCC1-RLP	0.45	(RCC1-RLP)	0.9867	Subfamily not named and unknown function
AT2G40316.1	Y	Y	RLP	0.9967	RLP	0.9907	RCC1-RLP	0.4007	(RCC1-RLP)	0.9866	Subfamily not named and unknown function
AT1G56200.1	N	Y	RLP	0.8952	RLP	0.9903	RCC1-RLP	0.3168	(RCC1-RLP)	0.8009	Subfamily not named and unknown function
AT1G53640.1	N	Y	RLP	0.6809	RLP	0.9917	RCC1-RLP	0.3501	(RCC1-RLP)	0.7979	Subfamily not named and unknown function
AT3G46240.1	N	Y	RLP	0.7885	RLP	0.9894	RCC1-RLP	0.3338	(RCC1-RLP)	0.7974	Malectin-RLP
AT3G08600.1	Y	Y	RLP	0.7903	RLP	0.9914	Unknown-RLP	0.6003	(Unknown-RLP)	0.9887	**Subfamily not named and unknown function**
AT5G65390.1	Y	Y	RLP	0.9963	RLP	0.9904	Unknown-RLP	0.3665	(Unknown-RLP)	0.9883	**Subfamily not named and unknown function**
AT3G03860.1	Y	Y	RLP	0.686	RLP	0.9906	Unknown-RLP	0.3338	(Unknown-RLP)	0.9883	OsAPRL5 adenosine 5\'-phosphosulfate reductase-like OsAPRL5. expressed
AT4G08670.1	Y	Y	RLP	0.9965	RLP	0.9911	Unknown-RLP	0.3329	(Unknown-RLP)	0.9881	LTPL66 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT4G11950.1	Y	Y	RLP	0.7883	RLP	0.9909	Unknown-RLP	0.4833	(Unknown-RLP)	0.9879	**Subfamily not named and unknown function**
AT1G11130.1	Y	Y	RLP	0.6837	RLP	0.7131	Unknown-RLP	0.3502	(Unknown-RLP)	0.9878	Leucine-rich repeat protein kinase family protein
AT1G23040.1	Y	Y	RLP	0.9485	RLP	0.8575	Unknown-RLP	0.5163	(Unknown-RLP)	0.9877	**Subfamily not named and unknown function**

AT2G47930.1	Y	Y	RLP	0.9964	RLP	0.9896	Unknown-RLP	0.4668	(Unknown-RLP)	0.9877	**Subfamily not named and unknown function**
AT1G35230.1	Y	Y	RLP	0.9961	RLP	0.9904	Unknown-RLP	0.5497	(Unknown-RLP)	0.9876	**Subfamily not named and unknown function**
AT3G27416.1	Y	Y	RLP	0.9966	RLP	0.9909	Unknown-RLP	0.4169	(Unknown-RLP)	0.9875	**Subfamily not named and unknown function**
AT5G10430.1	Y	Y	RLP	0.9452	RLP	0.9907	Unknown-RLP	0.534	(Unknown-RLP)	0.9874	**Subfamily not named and unknown function**
AT5G53870.1	Y	Y	RLP	0.9965	RLP	0.9911	Unknown-RLP	0.5164	(Unknown-RLP)	0.9874	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT4G09560.1	Y	Y	RLP	0.8934	RLP	0.991	Unknown-RLP	0.3335	(Unknown-RLP)	0.9872	RING finger protein 13 / Other-RLK (Ring finger-RLK)
AT2G23130.1	Y	Y	RLP	0.996	RLP	0.991	Unknown-RLP	0.6997	(Unknown-RLP)	0.9871	**Subfamily not named and unknown function**
AT4G27520.1	Y	Y	RLP	0.9466	RLP	0.99	Unknown-RLP	0.433	(Unknown-RLP)	0.9871	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT5G60650.1	Y	Y	RLP	0.8417	RLP	0.856	Unknown-RLP	0.3497	(Unknown-RLP)	0.9871	**Subfamily not named and unknown function**
AT4G34190.1	Y	Y	RLP	0.7873	RLP	0.9905	Unknown-RLP	0.4495	(Unknown-RLP)	0.987	**Subfamily not named and unknown function**
AT3G60280.1	Y	Y	RLP	0.8962	RLP	0.9911	Unknown-RLP	0.3169	(Unknown-RLP)	0.9868	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT1G02405.1	Y	Y	RLP	0.9961	RLP	0.9903	Unknown-RLP	0.6669	(Unknown-RLP)	0.9867	**Proline-rich family protein/Subfamily not named and unknown function**
AT1G36150.1	Y	Y	RLP	0.9465	RLP	0.9902	Unknown-RLP	0.4164	(Unknown-RLP)	0.9867	LTPL69 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT5G11990.1	Y	Y	RLP	0.9966	RLP	0.9914	Unknown-RLP	0.5335	(Unknown-RLP)	0.9866	**Proline-rich family protein/Subfamily not named and unknown function**
AT4G28100.1	Y	Y	RLP	0.8927	RLP	0.9907	Unknown-RLP	0.3832	(Unknown-RLP)	0.9865	**Subfamily not named and unknown function**
AT4G09030.1	Y	Y	RLP	0.9965	RLP	0.9908	Unknown-RLP	0.6673	(Unknown-RLP)	0.9864	**Subfamily not named and unknown function**
AT5G18690.1	Y	Y	RLP	0.9963	RLP	0.9902	Unknown-RLP	0.3503	(Unknown-RLP)	0.9864	**Subfamily not named and unknown function**
AT3G45275.1	Y	Y	RLP	0.8421	RLP	0.9907	Unknown-RLP	0.4164	(Unknown-RLP)	0.9862	**Subfamily not named and unknown function**
AT2G22470.1	Y	Y	RLP	0.8958	RLP	0.9906	Unknown-RLP	0.4672	(Unknown-RLP)	0.986	**Subfamily not named and unknown function**
AT1G72600.1	Y	Y	RLP	0.8427	RLP	0.9914	Unknown-RLP	0.4997	(Unknown-RLP)	0.9858	**Subfamily not named and unknown function**
AT5G14380.1	Y	Y	RLP	0.8933	RLP	0.9906	Unknown-RLP	0.3833	(Unknown-RLP)	0.9858	**Subfamily not named and unknown function**
AT3G45230.1	Y	Y	RLP	0.8437	RLP	0.9914	Unknown-RLP	0.5834	(Unknown-RLP)	0.9855	**Subfamily not named and unknown function**

AT3G29270.1	N	Y	RLP	0.7359	RLP	0.9904	Unknown-RLP	0.35	(Unknown-RLP)	0.8032	RING zinc finger protein-like / Other-RLK (Ring finger-RLK)
AT2G07678.1	N	Y	RLP	0.7893	RLP	0.8558	Unknown-RLP	0.3661	(Unknown-RLP)	0.7997	**Subfamily not named and unknown function**
AT2G01590.1	N	Y	RLP	0.6804	RLP	0.9905	Unknown-RLP	0.4832	(Unknown-RLP)	0.7993	**Subfamily not named and unknown function**
AT5G06660.1	N	Y	RLP	0.736	RLP	0.9907	Unknown-RLP	0.466	(Unknown-RLP)	0.799	**Subfamily not named and unknown function**
AT1G14345.1	N	Y	RLP	0.6346	RLP	0.9902	Unknown-RLP	0.5664	(Unknown-RLP)	0.7989	**Subfamily not named and unknown function**
AT5G63040.1	N	Y	RLP	0.7899	RLP	0.9919	Unknown-RLP	0.3167	(Unknown-RLP)	0.7985	**Subfamily not named and unknown function**
AT2G20230.1	N	Y	RLP	0.7363	RLP	0.9905	Unknown-RLP	0.3332	(Unknown-RLP)	0.7982	**Subfamily not named and unknown function**
AT4G35080.1	N	Y	RLP	0.6868	RLP	0.9911	Unknown-RLP	0.3838	(Unknown-RLP)	0.7979	high-affinity nickel-transport family protein. putative. expressed
AT4G30260.1	N	Y	RLP	0.7871	RLP	0.9912	Unknown-RLP	0.3666	(Unknown-RLP)	0.7979	Yip1 domain containing protein. expressed
AT3G56010.1	N	Y	RLP	0.7366	RLP	0.99	Unknown-RLP	0.3169	(Unknown-RLP)	0.7979	**Subfamily not named and unknown function**
AT3G12030.1	N	Y	RLP	0.7916	RLP	0.9905	Unknown-RLP	0.4002	(Unknown-RLP)	0.7978	**Subfamily not named and unknown function**
AT4G28770.1	N	Y	RLP	0.9479	RLP	0.9907	Unknown-RLP	0.3338	(Unknown-RLP)	0.7978	**Subfamily not named and unknown function**
AT4G14690.1	N	Y	RLP	0.7387	RLP	0.9913	Unknown-RLP	0.317	(Unknown-RLP)	0.7977	early light-induced protein. chloroplast precursor. putative. expressed
AT3G26350.1	N	Y	RLP	0.6292	RLP	0.9915	Unknown-RLP	0.5831	(Unknown-RLP)	0.7976	harpin-induced protein 1 domain containing protein. expressed
AT4G24460.1	N	Y	RLP	0.7368	RLP	0.9902	Unknown-RLP	0.3169	(Unknown-RLP)	0.7974	**Subfamily not named and unknown function**
AT2G25169.1	N	Y	RLP	0.8403	RLP	0.9905	Unknown-RLP	0.3667	(Unknown-RLP)	0.7973	**Subfamily not named and unknown function**
AT1G44890.1	N	Y	RLP	0.6841	RLP	0.8546	Unknown-RLP	0.3498	(Unknown-RLP)	0.7971	**Subfamily not named and unknown function**
AT3G49840.1	N	Y	RLP	0.7341	RLP	0.991	Unknown-RLP	0.3502	(Unknown-RLP)	0.7967	**Subfamily not named and unknown function**
AT2G38360.1	N	Y	RLP	0.629	RLP	0.9902	Unknown-RLP	0.3174	(Unknown-RLP)	0.7967	prenylated rab acceptor. putative. expressed
AT1G29390.1	N	Y	RLP	0.8953	RLP	0.9913	Unknown-RLP	0.4835	(Unknown-RLP)	0.7966	cold acclimation protein WCOR413. putative. expressed
AT1G54215.1	N	Y	RLP	0.6825	RLP	0.9908	Unknown-RLP	0.6666	(Unknown-RLP)	0.7963	**Subfamily not named and unknown function**
AT2G16800.1	N	Y	RLP	0.8426	RLP	0.9904	Unknown-RLP	0.4	(Unknown-RLP)	0.796	high-affinity nickel-transport family protein. putative. expressed

ATMG00920.1	N	Y	RLP	0.7872	RLP	0.858	Unknown-RLP	0.3839	(Unknown-RLP)	0.7959	**Subfamily not named and unknown function**
AT4G32490.1	N	Y	RLP	0.7914	RLP	0.9892	Unknown-RLP	0.3164	(Unknown-RLP)	0.7959	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT4G32600.1	N	Y	RLP	0.7913	RLP	0.9907	Unknown-RLP	0.4002	(Unknown-RLP)	0.7958	zinc finger family protein. putative. expressed
AT5G01080.1	N	Y	RLP	0.6836	RLP	0.9903	Unknown-RLP	0.3334	(Unknown-RLP)	0.7957	**Subfamily not named and unknown function**
AT3G01345.1	N	Y	RLP	0.8439	RLP	0.9904	Unknown-RLP	0.3499	(Unknown-RLP)	0.7951	**Subfamily not named and unknown function**
AT4G04870.1	N	Y	RLP	0.6829	RLP	0.8572	Unknown-RLP	0.3663	(Unknown-RLP)	0.7941	CDP-alcohol phosphatidyltransferase. putative. expressed
AT4G03298.1	N	Y	RLP	0.7313	RLP	0.9896	Unknown-RLP	0.3998	(Unknown-RLP)	0.794	**Subfamily not named and unknown function**
AT4G08874.1	N	Y	RLP	0.6312	RLP	0.991	Unknown-RLP	0.4	(Unknown-RLP)	0.7932	**Subfamily not named and unknown function**
AT4G21740.1	N	Y	RLP	0.9963	RLP	0.9922	Unknown-RLP	0.3004	(Unknown-RLP)	0.7929	**Subfamily not named and unknown function**
AT5G11280.1	N	Y	RLP	0.6307	RLP	0.9893	Unknown-RLP	0.5336	(Unknown-RLP)	0.7928	**Subfamily not named and unknown function**
AT1G16860.1	N	Y	RLP	0.8935	RLP	0.9908	Unknown-RLP	0.3663	(Unknown-RLP)	0.7928	**Subfamily not named and unknown function**
AT1G80200.1	N	Y	RLP	0.7877	RLP	0.9907	Unknown-RLP	0.416	(Unknown-RLP)	0.792	**Subfamily not named and unknown function**
c	Y	Y	RLP	0.8937	RLP	0.9908	Unknown	0.5159	(Undefined)	0.9891	Subfamily not named and unknown function
AT3G16670.1	Y	Y	RLP	0.8393	RLP	0.9893	Unknown	0.283	(Undefined)	0.7975	Subfamily not named and unknown function
AT2G16230.1	Y	Y	RLP	0.6837	RLP	0.9915	Unknown	0.2834	(Undefined)	0.7973	glucan endo-1.3-beta-glucosidase precursor. putative. expressed
AT3G06035.1	Y	Y	RLP	0.9967	RLP	0.9919	Unknown	0.2661	(Undefined)	0.7959	Subfamily not named and unknown function
AT1G62790.1	Y	Y	RLP	0.738	RLP	0.9912	Unknown	0.2662	(Undefined)	0.7921	LTPL85 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT5G50050.1	Y	Y	RLP	0.6851	RLP	0.9904	Unknown	0.2329	(Undefined)	0.8001	Plant invertase/pectin methylesterase inhibitor superfamily protein
AT1G43090.1	Y	Y	RLP	0.7367	RLP	0.9903	Unknown	0.4498	(Undefined)	0.9885	polygalacturonase. putative. expressed
AT4G26466.1	Y	Y	RLP	0.8421	RLP	0.9897	Unknown	0.3165	(Undefined)	0.9884	Subfamily not named and unknown function
AT1G71696.1	Y	Y	RLP	0.6811	RLP	0.9899	Unknown	0.3001	(Undefined)	0.9884	Subfamily not named and unknown function
AT4G36945.1	Y	Y	RLP	0.9966	RLP	0.9912	Unknown	0.3836	(Undefined)	0.988	Subfamily not named and unknown function
AT2G01630.1	Y	Y	RLP	0.6316	RLP	0.9903	Unknown	0.333	(Undefined)	0.9879	glucan endo-1.3-beta-glucosidase precursor. putative.

											expressed
AT2G19060.1	Y	Y	RLP	0.682	RLP	0.9896	Unknown	0.3672	(Undefined)	0.9878	GDSL-like lipase/acylhydrolase. putative. expressed
AT5G49270.1	Y	Y	RLP	0.7377	RLP	0.9918	Unknown	0.3502	(Undefined)	0.9877	COBRA-like protein 7 precursor. putative. expressed
AT5G48750.1	Y	Y	RLP	0.7349	RLP	0.9901	Unknown	0.3333	(Undefined)	0.9875	Subfamily not named and unknown function
AT3G18590.1	Y	Y	RLP	0.8457	RLP	0.9908	Unknown	0.3167	(Undefined)	0.9875	plastocyanin-like domain containing protein / Other-RLP associated with Probable-lipid-transfer-RLK
AT3G15720.1	Y	Y	RLP	0.688	RLP	0.991	Unknown	0.4662	(Undefined)	0.9874	polygalacturonase. putative. expressed
AT2G15770.1	Y	Y	RLP	0.9455	RLP	0.9905	Unknown	0.3167	(Undefined)	0.9874	plastocyanin-like domain containing protein / Other-RLP associated with Probable-lipid-transfer-RLK
AT1G71980.1	Y	Y	RLP	0.7903	RLP	0.9901	Unknown	0.3	(Undefined)	0.9874	RING finger protein 13. putative / Other-RLK (Ring finger-RLK)
AT3G13560.1	Y	Y	RLP	0.6873	RLP	0.9904	Unknown	0.3502	(Undefined)	0.9873	glucan endo-1.3-beta-glucosidase precursor. putative. expressed
AT1G74790.1	Y	Y	RLP	0.9469	RLP	0.9909	Unknown	0.333	(Undefined)	0.9873	expressed protein
AT5G42100.1	Y	Y	RLP	0.7381	RLP	0.9898	Unknown	0.3831	(Undefined)	0.9872	glycosyl hydrolases family 17. putative. expressed
AT1G43100.1	Y	Y	RLP	0.6837	RLP	0.9898	Unknown	0.5506	(Undefined)	0.987	polygalacturonase. putative. expressed
AT4G36440.1	Y	Y	RLP	0.7391	RLP	0.854	Unknown	0.5169	(Undefined)	0.987	Subfamily not named and unknown function
AT4G20790.1	Y	Y	RLP	0.9473	RLP	0.7125	Unknown	0.4673	(Undefined)	0.9868	Leucine-rich repeat protein kinase family protein
AT1G09790.1	Y	Y	RLP	0.7891	RLP	0.8569	Unknown	0.3668	(Undefined)	0.9868	COBRA. putative. expressed
AT2G17760.1	Y	Y	RLP	0.7893	RLP	0.8587	Unknown	0.5002	(Undefined)	0.9866	eukaryotic aspartyl protease domain containing protein. expressed
AT1G65720.1	Y	Y	RLP	0.8959	RLP	0.9899	Unknown	0.3163	(Undefined)	0.9866	Subfamily not named and unknown function
AT3G13410.1	Y	Y	RLP	0.789	RLP	0.9903	Unknown	0.3501	(Undefined)	0.9865	Subfamily not named and unknown function
AT1G22670.1	Y	Y	RLP	0.8951	RLP	0.9899	Unknown	0.3498	(Undefined)	0.9865	RING finger protein 13. putative / Other-RLK (Ring finger-RLK)
AT4G33490.1	Y	Y	RLP	0.6828	RLP	0.8551	Unknown	0.4167	(Undefined)	0.9864	eukaryotic aspartyl protease domain containing protein. expressed
AT2G25410.1	Y	Y	RLP	0.9959	RLP	0.9908	Unknown	0.4165	(Undefined)	0.9864	RING-H2 finger protein ATL2M / Other-RLK (Ring finger-RLK)
AT5G36001.1	Y	Y	RLP	0.9964	RLP	0.9912	Unknown	0.3667	(Undefined)	0.9862	zinc finger. C3HC4 type domain containing protein. expressed
AT3G28720.1	Y	Y	RLP	0.6305	RLP	0.9909	Unknown	0.4169	(Undefined)	0.9854	Subfamily not named and unknown function
AT4G34480.1	Y	Y	RLP	0.6837	RLP	0.9906	Unknown	0.267	(Undefined)	0.8062	glucan endo-1.3-beta-glucosidase precursor. putative. expressed
AT3G52640.1	Y	Y	RLP	0.7366	RLP	0.9907	Unknown	0.2998	(Undefined)	0.8046	nicalin. putative. expressed

AT2G38195.1	N	Y	RLP	0.7353	RLP	0.8595	Unknown	0.3169	(Undefined)	0.8023	Subfamily not named and unknown function
AT3G29810.1	Y	Y	RLP	0.7372	RLP	0.9904	Unknown	0.2343	(Undefined)	0.8016	COBRA. putative. expressed
AT4G16140.1	Y	Y	RLP	0.8414	RLP	0.9907	Unknown	0.2003	(Undefined)	0.8015	Subfamily not named and unknown function
AT1G26510.1	N	Y	RLP	0.6856	RLP	0.9913	Unknown	0.4834	(Undefined)	0.8004	OsFBX440 - F-box domain containing protein. expressed
AT2G11005.1	N	Y	RLP	0.8923	RLP	0.9917	Unknown	0.3168	(Undefined)	0.8004	Subfamily not named and unknown function
AT2G15910.1	N	Y	RLP	0.8941	RLP	0.9905	Unknown	0.3001	(Undefined)	0.7996	expp1 protein precursor. putative. expressed
AT2G36100.1	N	Y	RLP	0.6828	RLP	0.9891	Unknown	0.3501	(Undefined)	0.7992	Subfamily not named and unknown function
AT3G51040.1	N	Y	RLP	0.6309	RLP	0.9907	Unknown	0.3006	(Undefined)	0.7974	green ripe-like. putative. expressed
AT5G40990.1	N	Y	RLP	0.7878	RLP	0.9907	Unknown	0.3001	(Undefined)	0.7972	GDSL-like lipase/acylhydrolase. putative. expressed
AT1G69330.1	N	Y	RLP	0.9961	RLP	0.9899	Unknown	0.3661	(Undefined)	0.7969	RING zinc finger protein-like / Other-RLK (Ring finger-RLK)
AT3G15800.1	Y	Y	RLP	0.6867	RLP	0.9899	Unknown	0.25	(Undefined)	0.7969	glycosyl hydrolases family 17. putative. expressed
AT3G20610.1	N	Y	RLP	0.6334	RLP	0.9904	Unknown	0.3662	(Undefined)	0.7967	non-race specific disease resistance protein/Subfamily not named and unknown function
AT1G80400.1	N	Y	RLP	0.8944	RLP	0.9909	Unknown	0.3993	(Undefined)	0.7965	zinc finger family protein. putative. expressed
AT2G36670.1	Y	Y	RLP	0.63	RLP	0.9909	Unknown	0.2999	(Undefined)	0.7964	Xylanase inhibitor C-terminal
AT4G22650.1	Y	Y	RLP	0.6828	RLP	0.9905	Unknown	0.2334	(Undefined)	0.7953	Subfamily not named and unknown function
AT4G39840.1	Y	Y	RLP	0.9478	RLP	0.9898	Unknown	0.2499	(Undefined)	0.7949	Subfamily not named and unknown function
AT4G22630.1	Y	Y	RLP	0.946	RLP	0.9906	Unknown	0.2501	(Undefined)	0.7947	Subfamily not named and unknown function
AT5G46370.1	N	Y	RLP	0.6819	RLP	0.8558	Unknown	0.3661	(Undefined)	0.7943	potassium channel protein. putative. expressed
AT3G62280.1	Y	Y	RLP	0.8399	RLP	0.9898	Unknown	0.25	(Undefined)	0.7937	GDSL-like lipase/acylhydrolase. putative. expressed
AT1G16360.1	N	Y	RLP	0.8429	RLP	0.9896	Unknown	0.3497	(Undefined)	0.7916	cell cycle control protein. putative. expressed
AT2G30933.1	Y	Y	RLP	0.6858	RLP	0.9916	Unknown	0.4672	(Undefined)	0.9867	X8 domain containing protein. expressed
AT2G44790.1	Y	Y	RLP	0.6334	RLP	0.992	Unknown	0.2001	(Undefined)	0.8006	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT2G03505.1	Y	Y	RLP	0.7906	RLP	0.9895	Unknown	0.3	(Undefined)	0.7991	X8 domain containing protein. expressed
AT1G21090.1	Y	Y	RLP	0.9466	RLP	0.9906	Unknown	0.2843	(Undefined)	0.7994	Subfamily not named and unknown function
AT5G49280.1	Y	Y	RLP	0.9964	RLP	0.9909	Unknown	0.2333	(Undefined)	0.7974	Subfamily not named and unknown function
AT1G07690.1	Y	Y	RLP	0.8407	RLP	0.9899	Unknown	0.2338	(Undefined)	0.7954	Subfamily not named and unknown function

AT4G01140.1	Y	Y	RLP	0.896	RLP	0.9903	Unknown	0.2833	(Undefined)	0.7976	Subfamily not named and unknown function
AT5G07190.1	Y	Y	RLP	0.8412	RLP	0.9905	Unknown	0.2502	(Undefined)	0.7962	embryo-specific 3. putative. expressed
AT5G60630.1	Y	Y	RLP	0.6846	RLP	0.9909	Unknown	0.2332	(Undefined)	0.7981	Subfamily not named and unknown function
AT1G71110.1	Y	Y	RLP	0.7383	RLP	0.9913	Unknown	0.2667	(Undefined)	0.7978	Subfamily not named and unknown function
AT2G12400.1	Y	Y	RLP	0.7394	RLP	0.9908	Unknown	0.2333	(Undefined)	0.7977	Subfamily not named and unknown function
AT2G25270.1	Y	Y	RLP	0.7904	RLP	0.9908	Unknown	0.2669	(Undefined)	0.7985	Subfamily not named and unknown function
AT4G31370.1	Y	Y	RLP	0.893	RLP	0.9918	Unknown	0.2666	(Undefined)	0.8001	fasciclin-like arabinogalactan precursor protein / Other-RLP associated with Plastocyanin-like-RLK
AT3G58100.1	Y	Y	RLP	0.6331	RLP	0.9913	Unknown	0.25	(Undefined)	0.8008	X8 domain containing protein. expressed
AT3G51510.1	Y	Y	RLP	0.6313	RLP	0.9897	Unknown	0.2669	(Undefined)	0.8007	Subfamily not named and unknown function
AT5G27830.1	Y	Y	RLP	0.7906	RLP	0.9911	Unknown	0.2835	(Undefined)	0.8005	CRP12 - Cysteine-rich family protein precursor. expressed
AT5G07475.1	Y	Y	RLP	0.8955	RLP	0.9905	Unknown	0.2008	(Undefined)	0.8031	plastocyanin-like domain containing protein / Other-RLP associated with Plastocyanin-like-RLK
AT3G43720.1	Y	Y	RLP	0.7375	RLP	0.9911	Unknown	0.2992	(Undefined)	0.7985	LTPL67 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT2G13820.1	Y	Y	RLP	0.8407	RLP	0.9908	Unknown	0.2335	(Undefined)	0.8024	LTPL69 - Protease inhibitor / Other-RLP associated with Probable-lipid-transfer-RLK
AT1G09176.1	Y	Y	RLP	0.8432	RLP	0.9911	Unknown	0.2666	(Undefined)	0.801	Subfamily not named and unknown function
AT4G22900.1	Y	Y	RLP	0.842	RLP	0.991	Unknown	0.2839	(Undefined)	0.8026	Subfamily not named and unknown function
AT1G11362.1	Y	Y	RLP	0.8391	RLP	0.9911	Unknown	0.2502	(Undefined)	0.8002	Subfamily not named and unknown function
AT5G19250.1	Y	Y	RLP	0.9963	RLP	0.9908	Unknown	0.2501	(Undefined)	0.7944	Subfamily not named and unknown function

Table S3. Enriched GO terms in three categories, Biological Process, Molecular Function, or Cellular Component ontology

Celular component						
GOCCID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
GO:0046658	0.00	48.75	0.207	8	68	anchored component of plasma membrane
GO:0044459	0.00	13.09	0.699	8	230	plasma membrane part
GO:0005886	0.00	3.68	8.324	23	3032	plasma membrane
GO:0048046	0.00	7.77	0.997	7	328	apoplast
GO:0005794	0.00	4.43	2.821	11	928	Golgi apparatus
GO:0009505	0.00	7.89	0.833	6	274	plant-type cell wall
GO:0009506	0.00	4.33	2.590	10	852	plasmodesma
GO:0030054	0.00	4.32	2.596	10	854	cell junction
GO:0030312	0.00	4.27	1.782	7	586	external encapsulating structure
GO:0005783	0.01	3.87	1.660	6	546	endoplasmic reticulum
GO:0030173	0.01	110.75	0.012	1	4	integral component of Golgi membrane
GO:0031225	0.01	6.44	0.491	3	180	anchored component of membrane
GO:0005773	0.02	2.87	2.605	7	857	vacuole
GO:0005737	0.04	1.56	36.641	45	12052	cytoplasm
GO:0008180	0.04	27.68	0.040	1	13	COP9 signalosome
GO:0030176	0.04	27.68	0.040	1	13	integral component of endoplasmic reticulum membrane
GO:0098791	0.04	4.04	0.772	3	254	Golgi subcompartment
GO:0098588	0.04	2.74	1.915	5	630	bounding membrane of organelle
GO:0005623	0.05	2.23	65.718	71	21616	cell
GO:0055028	0.05	22.14	0.049	1	16	cortical microtubule
Molecular function						

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0008889	1.23E-12	301.83	0.041	6	13	glycerophosphodiester phosphodiesterase activity
GO:0042578	4.89E-09	13.25	1.018	11	326	phosphoric ester hydrolase activity
GO:0045309	1.03E-05	91.30	0.044	3	14	protein phosphorylated amino acid binding
GO:1901981	0.001	54.89	0.044	2	14	phosphatidylinositol phosphate binding
GO:0016301	0.001	3.25	3.870	11	1259	kinase activity
GO:0004127	0.003	Inf	0.003	1	1	cytidylate kinase activity
GO:0004140	0.003	Inf	0.003	1	1	dephospho-CoA kinase activity
GO:0005545	0.003	Inf	0.003	1	1	1-phosphatidylinositol binding
GO:0031210	0.003	Inf	0.003	1	1	phosphatidylcholine binding
GO:0042973	0.003	Inf	0.003	1	1	glucan endo-1.3-beta-D-glucosidase activity
GO:0016787	0.004	2.445	7.145	15	2288	hydrolase activity
GO:0003993	0.008	16.444	0.131	2	42	acid phosphatase activity
GO:0008017	0.008	16.042	0.134	2	43	microtubule binding
GO:0009041	0.009	162.135	0.009	1	3	uridylate kinase activity
GO:0046577	0.012	108.085	0.012	1	4	long-chain-alcohol oxidase activity
GO:0030554	0.014	3.780	1.415	5	453	adenyl nucleotide binding
GO:0005460	0.016	81.060	0.016	1	5	UDP-glucose transmembrane transporter activity
GO:0005546	0.016	81.060	0.016	1	5	phosphatidylinositol-4.5-bisphosphate binding
GO:0050378	0.016	81.060	0.016	1	5	UDP-glucuronate 4-epimerase activity
GO:0000062	0.019	64.844	0.019	1	6	fatty-acyl-CoA binding
GO:0004605	0.025	46.313	0.025	1	8	phosphatidate cytidylyltransferase activity
GO:0008195	0.028	40.522	0.028	1	9	phosphatidate phosphatase activity
GO:0005543	0.029	7.848	0.267	2	87	phospholipid binding
GO:0004630	0.037	29.466	0.037	1	12	phospholipase D activity
GO:0032555	0.041	2.782	1.902	5	609	purine ribonucleotide binding

GO:0005338	0.043	24.930	0.044	1	14	nucleotide-sugar transmembrane transporter activity
GO:0008092	0.044	6.245	0.334	2	107	cytoskeletal protein binding
GO:0097367	0.045	2.708	1.952	5	625	carbohydrate derivative binding
GO:0000166	0.048	2.419	2.636	6	844	nucleotide binding
GO:0005524	0.049	3.050	1.377	4	441	ATP binding

Biological process

GOBPID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
GO:0006071	2.05E-11	162.45	0.06	6	18	glycerol metabolic process
GO:0019751	1.65E-10	105.43	0.08	6	25	polyol metabolic process
GO:0010015	3.62E-08	9.44	1.54	12	458	root morphogenesis
GO:0048765	1.19E-07	10.91	1.08	10	323	root hair cell differentiation
GO:0048469	1.23E-07	10.88	1.09	10	324	cell maturation
GO:0010054	2.78E-07	9.91	1.19	10	354	trichoblast differentiation
GO:0022622	9.83E-07	6.84	2.09	12	623	root system development
GO:0044262	1.94E-06	6.39	2.23	12	665	cellular carbohydrate metabolic process
GO:0071695	6.44E-06	6.88	1.68	10	502	anatomical structure maturation
GO:0010442	1.11E-05	Inf	0.01	2	2	guard cell morphogenesis
GO:0048589	1.21E-05	5.29	2.67	12	808	developmental growth
GO:0009832	2.35E-05	11.73	0.57	6	171	plant-type cell wall biogenesis
GO:0090627	2.45E-05	7.61	1.19	8	366	plant epidermal cell differentiation
GO:0090558	3.69E-05	8.57	0.92	7	286	plant epidermis development
GO:1901615	4.17E-05	5.00	2.55	11	759	organic hydroxy compound metabolic process
GO:0031110	0.000	152.96	0.02	2	6	regulation of microtubule polymerization or depolymerization
GO:0032886	0.000	122.36	0.02	2	7	regulation of microtubule-based process
GO:0048767	0.000	9.80	0.56	5	172	root hair elongation

GO:0010052	0.000	87.39	0.03	2	9	guard cell differentiation
GO:0000902	0.001	4.33	2.33	9	695	cell morphogenesis
GO:0099402	0.001	3.49	3.60	11	1119	plant organ development
GO:0048768	0.001	43.68	0.05	2	16	root hair cell tip growth
GO:0006073	0.001	8.95	0.48	4	149	cellular glucan metabolic process
GO:0090066	0.002	7.68	0.56	4	167	regulation of anatomical structure size
GO:0016049	0.003	3.63	2.42	8	722	cell growth
GO:0052541	0.003	27.79	0.08	2	24	plant-type cell wall cellulose metabolic process
GO:0009267	0.003	5.46	0.99	5	294	cellular response to starvation
GO:0009605	0.003	2.49	6.95	15	2072	response to external stimulus
GO:0080184	0.003	Inf	0.00	1	1	response to phenylpropanoid
GO:0006650	0.004	6.87	0.62	4	186	glycerophospholipid metabolic process
GO:0007275	0.004	2.17	11.57	21	3451	multicellular organism development
GO:0045017	0.004	6.68	0.64	4	191	glycerolipid biosynthetic process
GO:0030243	0.005	9.64	0.33	3	102	cellulose metabolic process
GO:0009932	0.005	4.97	1.08	5	332	cell tip growth
GO:0006661	0.005	9.18	0.35	3	104	phosphatidylinositol biosynthetic process
GO:0031667	0.005	4.77	1.12	5	335	response to nutrient levels
GO:0009816	0.006	18.51	0.12	2	35	defense response to bacterium. incompatible interaction
GO:0031668	0.006	4.58	1.17	5	349	cellular response to extracellular stimulus
GO:0031115	0.007	301.52	0.01	1	2	negative regulation of microtubule polymerization
GO:0009653	0.008	2.53	4.85	11	1512	anatomical structure morphogenesis
GO:0019637	0.008	2.80	3.53	9	1078	organophosphate metabolic process
GO:0052546	0.008	15.66	0.14	2	41	cell wall pectin metabolic process
GO:0044281	0.009	2.11	9.27	17	2765	small molecule metabolic process
GO:0048527	0.009	7.53	0.42	3	126	lateral root development

GO:0008654	0.010	4.11	1.30	5	387	phospholipid biosynthetic process
GO:0009173	0.010	150.75	0.01	1	3	pyrimidine ribonucleoside monophosphate metabolic process
GO:0000272	0.011	13.27	0.16	2	48	polysaccharide catabolic process
GO:0051493	0.013	104.51	0.01	1	4	regulation of cytoskeleton organization
GO:0010971	0.013	100.50	0.01	1	4	positive regulation of G2/M transition of mitotic cell cycle
GO:0030002	0.013	100.50	0.01	1	4	cellular anion homeostasis
GO:0030643	0.013	100.50	0.01	1	4	cellular phosphate ion homeostasis
GO:0080186	0.013	100.50	0.01	1	4	developmental vegetative growth
GO:1901989	0.013	100.50	0.01	1	4	positive regulation of cell cycle phase transition
GO:0046467	0.014	6.29	0.50	3	150	membrane lipid biosynthetic process
GO:0010393	0.015	11.30	0.19	2	56	galacturonan metabolic process
GO:0008361	0.016	11.10	0.19	2	57	regulation of cell size
GO:0044255	0.016	2.48	3.95	9	1177	cellular lipid metabolic process
GO:0000338	0.017	75.37	0.02	1	5	protein deneddylation
GO:0016126	0.018	5.74	0.55	3	164	sterol biosynthetic process
GO:0090626	0.018	4.23	1.00	4	298	plant epidermis morphogenesis
GO:0071555	0.018	5.67	0.56	3	179	cell wall organization
GO:0010026	0.019	5.57	0.57	3	169	trichome differentiation
GO:0010075	0.020	5.50	0.57	3	171	regulation of meristem growth
GO:0010387	0.020	60.29	0.02	1	6	COP9 signalosome assembly
GO:0045931	0.020	60.29	0.02	1	6	positive regulation of mitotic cell cycle
GO:0009664	0.021	4.04	1.04	4	320	plant-type cell wall organization
GO:0010044	0.023	50.24	0.02	1	7	response to aluminum ion
GO:0032271	0.026	44.13	0.03	1	8	regulation of protein polymerization
GO:0015937	0.027	43.06	0.03	1	8	coenzyme A biosynthetic process
GO:0006950	0.027	1.76	12.99	20	3873	response to stress

GO:1902903	0.029	38.61	0.03	1	9	regulation of supramolecular fiber organization
GO:0031333	0.030	37.68	0.03	1	9	negative regulation of protein complex assembly
GO:0090696	0.035	3.41	1.23	4	367	post-embryonic plant organ development
GO:0006796	0.035	2.00	6.01	11	1836	phosphate-containing compound metabolic process
GO:0043647	0.036	7.00	0.30	2	89	inositol phosphate metabolic process
GO:0016043	0.036	1.78	10.01	16	2986	cellular component organization
GO:0051494	0.036	30.14	0.04	1	11	negative regulation of cytoskeleton organization
GO:0007155	0.038	6.77	0.31	2	92	cell adhesion
GO:0010035	0.038	2.11	4.58	9	1367	response to inorganic substance
GO:0010374	0.039	4.21	0.74	3	222	stomatal complex development
GO:0033875	0.040	6.62	0.32	2	94	ribonucleoside bisphosphate metabolic process
GO:0034032	0.040	6.62	0.32	2	94	purine nucleoside bisphosphate metabolic process
GO:0000271	0.042	2.76	1.91	5	570	polysaccharide biosynthetic process
GO:0005982	0.044	4.01	0.78	3	233	starch metabolic process
GO:0019375	0.044	6.28	0.33	2	99	galactolipid biosynthetic process
GO:0055083	0.046	23.18	0.05	1	14	monovalent inorganic anion homeostasis
GO:0072506	0.046	23.18	0.05	1	14	trivalent inorganic anion homeostasis
GO:0009926	0.046	6.09	0.34	2	102	auxin polar transport
GO:0045010	0.047	6.03	0.35	2	103	actin nucleation
GO:0032273	0.048	5.97	0.35	2	104	positive regulation of protein polymerization
GO:0008202	0.049	3.82	0.82	3	244	steroid metabolic process
GO:0072505	0.049	21.52	0.05	1	15	divalent inorganic anion homeostasis
GO:0044089	0.049	5.85	0.36	2	106	positive regulation of cellular component biogenesis
GO:0051495	0.049	5.85	0.36	2	106	positive regulation of cytoskeleton organization
GO:1902905	0.049	5.85	0.36	2	106	positive regulation of supramolecular fiber organization