Outils et concepts pour l'annotation de familles de gènes complexes : le cas des récepteurs à LRR chez le riz

Celine Gottin

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE MONTPELLIER SUPAGRO

En Génétique et Génomique

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Portée par

L'Unité de recherche AGAP

Outils et concepts pour l'annotation de familles de gènes complexes : le cas des récepteurs à LRR chez le riz

Présentée par Céline GOTTIN
Le 26 novembre 2021

Sous la direction de Vincent RANWEZ

Devant le jury composé de

M. Sébastien AUBOURG, DR, INRAe	Rapporteur
M. Benoît LEFEBVRE, DR, INRAe	Rapporteur
Mme Valérie GEFFROY, DR, INRAe	Examinatrice
M. François SABOT, DR, IRD	Examinateur, Président du jury
M. Matthieu CONTE, PhD., Syngenta	Examinateur
M. Vincent RANWEZ, Pr, Institut Agro	Directeur de thèse
Mme Nathalie CHANTRET, CR, INRAe	Co-encadrante
Mme Anne DIEVART, CR, CIRAD	Co-encadrante

UNIVERSITÉ DE MONTPELLIER

Montpellier SupAgro

Remerciements

Voilà donc venu le temps des remerciements. Bien que cette section apparaisse en début de manuscrit, elle signe pour moi la fin de cette longue aventure. Et il n'y a pas d'aventure sans compagnons de route, alors je voudrais adresser à chacun mes remerciements.

Et pour commencer, je voudrais remercier mes encadrants Vincent, Nathalie et Anne. Toujours présents, motivants et prêts à prendre un café pour parler de tout, et parfois de la thèse, vous avez fait de ces trois années un moment bien plus *fun* que je ne l'aurais espéré. Merci de m'avoir tant appris et de m'avoir soutenu et encouragé quand les choses se sont un peu… compliquées (les boîtes bleues ont bien failli avoir ma santé mentale !). A vous trois, un immense merci.

Je souhaite aussi remercier l'équipe DAR de m'avoir accueilli au début de ma thèse. Merci à Christophe, Ian, Léo, Thibault, Remy et tous les autres pour votre bonne humeur contagieuse.

Merci à Marilyne qui a développé et mis en place le site Geloc, offrant une fin parfaite à un projet fou. Merci également à l'équipe bio-info du CIRAD pour son aide en toutes situations.

Merci à mes compagnons de bureau à ARCAD, Pauline, Yacine, Michel, Germain et Josselin. C'était un plaisir de vous tirer de votre travail pour vous raconter des histoires souvent inutiles. Je voudrais aussi remercier les membres de l'équipe Ge2pop qui m'ont accueilli et supporté pendant deux ans. Merci à Morgane, Sylvain, Johanna, Audrey, Muriel, Hélène, Joëlle, et bien d'autres pour ces discussions et ces rires au détour d'un couloir ou autour d'un café.

Merci également aux doctorants d'AGAP, présents et anciens, pour la bonne entente, les cafés au bâtiment 3 et les Midi Doc ! Merci donc à Benjamin, Clara, Kelly, Aurélie, Dylan, mais aussi à Léo, Thibault et Ian que j'ai déjà cité, et à bien d'autres !

Merci aux amis, Colin, Martin, Jade, Nicolas, Pascal, Sophie, Camille, Louis, Simon, Charlotte, toujours partant pour boire un verre, faire des jeux de sociétés ou traverser la France pour une bonne soirée !

Je remercie enfin ma famille, et tout particulièrement mes parents. Merci de m'avoir encouragé et de m'accorder votre fierté en toutes situations qui est pour moi la plus grande des motivations.

On dit souvent qu'on garde le meilleur pour la fin, je ne dirais pas le contraire. Nicolas, merci de me soutenir et de me supporter jour après jour et en particulier pendant ces dernières semaines pour le moins intenses. Merci d'être là, merci d'être toi.

Résumé

Les maladies des plantes causées par des agents pathogènes constituent une menace pour la sécurité alimentaire mondiale. Les gènes impliqués dans les mécanismes de résistance aux pathogènes sont des éléments clés permettant un contrôle efficace des maladies tout en limitant l'utilisation d'intrants chimiques. Les réactions immunitaires chez les plantes peuvent être initiées par des récepteurs portant un domaine répété riche en leucine (*Leucine-rich repeats,* LRR). Les gènes codant pour ces récepteurs appartiennent à différentes familles, dont les trois principales sont les *LRR-Receptor-Like Kinase* (LRR-RLK), les *LRR-Receptor-Like Protein* (LRR-RLP) et les *Nucleotide binding site LRR Receptor* (NLR). Etudier ces gènes et leur évolution est essentiel non seulement pour comprendre comment les résistances émergent, disparaissent ou se maintiennent chez les plantes, mais aussi pour mettre en place de nouvelles stratégies de sélection variétale. Malgré l'abondance des données génomiques, l'étude de ces récepteurs à LRR reste difficile car ces familles de gènes sont complexes. Premièrement, une grande partie des gènes de ces familles se dupliquent fréquemment et sont donc présents en de multiples copies dans les génomes. Ces copies accumulent des mutations dont certaines peuvent être non-sens, i.e. impacter la structure et/ou la fonction initiale du gène. Deuxièmement, les protéines de ces trois familles partagent un domaine composé de motifs LRR répétés jusqu'à plus de 30 fois, essentiel pour la spécificité du récepteur. Dans ce contexte particulier, les annotations structurales des gènes et celles des motifs LRR dans les séquences protéiques obtenues par les outils génériques contiennent beaucoup d'erreurs.

Au cours de cette thèse, je me suis intéressée tout d'abord à la détection et l'annotation des motifs LRR dans les protéomes de plantes en développant un outil dédié, LRRprofiler. En se basant sur un ensemble de profils HMM : LRRprofiler fournit une annotation complète, reproductible et comparable des protéines LRR pour les trois familles d'intérêt (LRR-RLK, LRR-RLP et NLR). Ensuite, je me suis intéressée aux biais des méthodes automatiques pour l'annotation structurale (intron-exon) des récepteurs LRR en comparant trois annotations publiques disponibles pour le riz (*Oryza sativa*) Nipponbare. L'identification d'erreurs récurrentes pour ces trois familles (fusion de gènes, gènes tronqués, gènes non-identifiés, etc.) m'a amenée à proposer une annotation expertisée manuellement pour les trois familles. La stratégie d'annotation proposée consiste à identifier tous les gènes, même ceux dont la structure est impactée par des mutations non-sens. Les modèles sont alors qualifiés de 'canoniques' ou 'non-canoniques' en fonction de la présence ou non de ces mutations non-sens. Afin de limiter l'intervention manuelle pour l'annotation d'autres génomes, un pipeline de transfert des annotations a été développé : LRRtransfer. Ce pipeline a permis d'annoter les récepteurs à LRR d'un autre cultivar de riz, KitaakeX, et de comparer les répertoires entre ces deux génotypes proches. L'ensemble des outils développés et des données générées au cours de cette thèse sont disponibles librement sous des formats standards et facilement réutilisables. Enfin, le site web 'Geloc' (https://rice-genome-hub.southgreen.fr/content/geloc) a été développé pour permettre la visualisation, l'exploration et la comparaison des annotations des récepteurs LRR chez le riz.

Abstract

Plant diseases caused by pathogens are a threat to global food security. Genes involved in pathogen resistance mechanisms are key elements for effective disease control while limiting the use of chemical inputs. Plant immune responses can be initiated by receptors carrying a leucine-rich repeat (LRR) domain. The genes encoding these receptors belong to different families, the three main ones being LRR-Receptor-Like Kinase (LRR-RLK), LRR-Receptor-Like Protein (LRR-RLP) and Nucleotide binding site LRR Receptor (NLR). Studying LRR-containing genes and their evolution is essential to understand how new resistances emerge, disappear or are maintained in plants, as well as to develop new breeding strategies. Despite the abundance of genomic data, the study of these receptors remains difficult because these gene families are particularly complex. First, many of these genes are frequently duplicated and are therefore present in multiple copies. These copies can accumulate mutations, some of which may be nonsense, i.e. they can impact the initial structure and/or function of the gene. Secondly, the proteins of these three families share a domain composed of up to 30 repeated LRR motifs, essential for receptor specificity. In this particular context, the structural annotations of the genes and of the LRR motifs in the protein sequences obtained by generic tools contain many errors.

During this thesis, I first focused on the detection and annotation of LRR motifs in plant proteomes by developing a dedicated tool named LRRprofiler. Based on a set of HMM profiles, the LRRprofiler pipeline provides a complete, reproducible and similar annotation of LRR proteins for the three families of interest (LRR-RLK, LRR-RLP and NLR). Next, I addressed the biases of automatic methods for structural (intron-exon) annotation of LRR containing receptors by comparing three publicly available annotations for rice (*Oryza sativa*) Nipponbare. The identification of recurrent errors for these three families (gene fusion, truncated genes, unidentified genes, etc.) led me to propose a manually expert annotation for these three families. The proposed annotation strategy consists in identifying all genes, even those whose structure contain nonsense mutations. The gene models are then qualified as 'canonical' or 'non-canonical' depending on the presence or absence of these nonsense mutations. In order to limit the manual intervention for the annotation of other genomes, an annotation transfer pipeline named LRRtransfer has been developed. This pipeline was used to annotate LRR receptors for another rice cultivar, KitaakeX, and to compare the gene repertoires between these two closely related genotypes. All the tools developed and the data generated during this thesis are freely available in standard and easily reusable formats. Finally, the 'Geloc' website (https://rice-genome-hub.southgreen.fr/content/geloc) has been developed to allow visualization, exploration and comparison of these LRR annotations in rice.

Sommaire

Liste des figures

Liste des tableaux

Abréviations

ADN : Acide DésoxyriboNucléique

ARN : Acide RiboNucléique

ARNm : ARN messager

BAK1 : *Brassinosteroid insensitive 1-Associated Kinase*

BLAST : *Basic Local Alignment Search Tool*

BRI1 : *Brassinosteroid Insensitive 1*

CC : *Coil coiled*

CDS : *Coding DNA Sequence*

CGIAR : *Consultative Group on International Agricultural Research*

CIRAD : Centre de coopération Internationale en Recherche Agronomique pour le Développement

CNL : CC-*containing* NBS-LRR

DAMP : *Damage-Associated Molecular Pattern*

EGF : *Epidermal Growth Factor*

EST : *Expressed Sequence Tag*

ETI : *Effector Triggered Immunity*

FAO : *Food and Agriculture Organization of the United Nations*

FBL : *F-box and LRR containing protein*

FLS2 : *Flagellin-Sensing 2*

GFF : *General Feature Format*

HCS : *Highly Conserved Segment*

HMM : *Hidden Markov Model*

ID : *Integrated Domains* ou *Integrated Decoy*

I-*O*MAP : *International Oryza Map Alignment Project*

IRAK : *Interleukine-1 Receptor-Associated Kinase*

IRGSP : *International Rice Genome Sequencing Project*

LRR : *Leucine-Rich Repeat*

LRX : *Extensin-like LRR*

LysM : *Lysin Motif domain*

MAMP : *Microbe-Associated Molecular Pattern*

MSU : *Michigan State University*

NACHT : *NAIP, CIITA, HET-E and TEP1 homologous domain*

NB-ARC : *Nucleotide-Binding domain homolog to APAF-1, R proteins and CED-4*

NBS : *Nucleotide-Binding Site*

NCBI : *National Center for Biotechnology Information*

NLP : *Necrosis and ethylene-inducing peptide 1-Like Proteins*

NLR : *Nucleotide-binding site Leucine-rich Repeat* (aussi appelé NBS-LRR)

ONU : Organisation des Nations Unies

PAMP : *Pathogen-Associated Molecular Pattern*

PGIP : *PolyGalacturonase-Inhibiting Protein*

PIRL : *Plant Intracellular Ras group-related LRR*

PPR : *Pentatricopeptide Repeat*

PRR : *Pattern Recognition Receptors*

PSSM : *Position Scoring Matrix*

PTI : *PAMP-Triggered Immunity*

RAP-DB : *Rice Annotation Project DataBase*

R genes : *Resistance genes*

RI-like : *Ribonuclease Inibitor like*

RLK : *Receptor-Like Kinase*

RLP : *Receptor-Like Protein*

RNL : *RPW8-containing NBS-LRR*

RPW8 : *Resistance to Powdery Mildew 8*

SERK3 : *Somatic Embryogenesis Receptor-like Kinase 3*

SOBIR1 : *Suppressor of BAK1-Interacting Receptor kinase 1*

STAND : *Signal Transduction ATPases with Numerous Domains*

TE : *Transposable Element*

TIR : *Toll-Interleukin Receptor*

TLR : *Toll-Like Receptors*

TM : *Transmembrane domain*

TNL : *TIR-containing NBS-LRR*

TpLRR : *Treponema Pallidum LRR motif*

AGAP : Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales

UTR : *Untranslated Transcribed Region*

VM : *Virtual Machine*

VS : *Variable Segment*

WGS : *Whole Genome Sequencing*

Introduction

Introduction

Table des matières

L'agriculture fait face aujourd'hui à d'importants challenges d'ordre économiques, sociétaux et environnementaux. La production alimentaire doit répondre à une demande croissante de nourriture tout en préservant l'environnement, les ressources naturelles et la santé des populations dans un contexte de changement global du climat.

L'augmentation des demandes alimentaires a deux origines. La première est la lutte contre la faim et l'insécurité alimentaire. Cet enjeu sociétal majeur est au cœur de l'Objectif « Faim zéro », développé par l'ONU (Organisation des Nations Unies), pour endiguer la faim dans le monde à l'horizon 2030. En 2018, 8,9 % de la population mondiale était déclarée sous-alimentée. Ces valeurs marquent la première année de progression de la faim, de 0,2 %, après plus d'une décennie de recul constant (FAO et al., 2020). Selon les projections de l'Organisation des Nations Unies pour l'alimentation et l'agriculture (*Food and Agriculture Organisation,* FAO) à partir des données de 2019, la prévalence de la faim devrait poursuivre son avancée pour toucher 9,8 % de la population mondiale en 2030 (FAO et al., 2020). Dernièrement, la pandémie de COVID-19 a été un facteur supplémentaire important de la progression de l'insécurité alimentaire, même pour les pays occidentaux. Bien que les conséquences exactes de cette pandémie soient difficiles à évaluer aujourd'hui, elles constitueront vraisemblablement un frein aux objectifs de l'ONU pour 2030.

La seconde origine de l'augmentation des demandes alimentaires est la croissance démographique. Selon les prévisions de l'ONU, la population mondiale devrait atteindre 9,7 milliards de personnes en 2050, représentant une croissance de plus de 24 % par rapport à 2020 (7,8 milliards de personnes). Une étude récente, publiée dans le journal The Lancet (Vollset et al., 2020), intégrant différents facteurs tels que l'éducation des filles, l'utilisation de la contraception et l'évolution de la mortalité, prévoit un pic de population pour le milieu du siècle (2046-2064), suivie d'une décroissance progressive. En fonction des modèles, ce pic irait de 8,8 à 9,7 milliards de personnes.

Ainsi, pour poursuivre la lutte contre l'insécurité alimentaire et pour faire face à la croissance démographique, il est nécessaire d'augmenter les productions agricoles. A partir des années 1960, d'importants moyens techniques et scientifiques mis en œuvre ont permis d'augmenter massivement les productions alimentaires mondiales (Pingali, 2012). Parmi ces progrès, désignés sous le nom de « Révolution Verte », on peut noter : la diffusion de variétés végétales à haut rendement issues de programmes de sélection ; l'utilisation d'intrants, et notamment

d'engrais chimiques et de produits phytosanitaires pour le contrôle des maladies ; la mécanisation du travail agricole ; et le développement de l'irrigation. Grâce à ces progrès, les productions mondiales globales ont augmenté de 208 % pour le blé, 109 % pour le riz et 157 % pour le maïs entre 1960 et 2000 (FAO, 2004; Pingali, 2012). Mais cette augmentation, résultant de l'industrialisation des productions agricoles, de l'extension des terres cultivables sur les plaines et les forêts, mais aussi de l'utilisation massive de pesticides et engrais chimiques, n'a pas été sans conséquence pour l'Homme et l'environnement. Ces modes de production, faisant pression sur les ressources naturelles, ne permettront pas de subvenir aux besoins alimentaires futurs de l'ensemble de la population. Ainsi, d'importants changements structurels dans les méthodes culturales sont nécessaires pour faire face aux besoins futurs tout en gérant au mieux les ressources.

A. Le riz, céréale clé de la sécurité alimentaire

Le riz (*Oryza sativa*) est la céréale la plus consommée par l'Homme dans le monde. Elle occupe et occupera une place centrale dans la lutte contre la faim. De nombreux projets internationaux ont pour objectifs d'adapter les cultures de riz aux conséquences environnementales du changement climatique, à l'émergence de nouvelles maladies et à l'augmentation de la population (*International Oryza Map Alignment Project* (I-*O*MAP) (Jacquemin et al., 2013), *CGIAR Research Program on Rice* (CRP RICE) (CGIAR, 2018), *Closing Rice Yield Gaps in Asia* (CORIGAP, https://corigap.irri.org/), *Hybrid Rice Development Consortium* (HRDC, https://hrdc.irri.org/), etc).

A.1. Production et consommation

En 2018, le riz représentait 8,2 % des productions totales et 26,2 % des productions céréalières au niveau mondial. Plus de 50 % de cette production est assurée par la Chine et l'Inde (Figure 1). C'est la deuxième céréale la plus produite dans le monde en 2020 (782 millions de tonnes), après le maïs (1 148 millions de tonnes) et avant le blé (734 millions de tonnes) (FAO, 2020). En revanche, le riz est la première céréale de consommation humaine avec une moyenne stable de 54,1 kg par habitant et par an (FAO, 2021). Cette consommation est très variable dans le monde. Par exemple, les français consomment en moyenne 7 kg de riz par habitant et par an. Mais le plus gros consommateur de riz en 2020 fut Myanmar (la Birmanie) avec en moyenne 183 kg par habitant (FAO, 2021). Pour de nombreux pays comme celui-ci,

notamment en Asie, en Afrique et en Amérique du Sud, il peut être difficile de se procurer de la viande, et des fruits et légumes frais pour des raisons économiques. La consommation de riz brut ou transformé constitue alors l'apport alimentaire principal et peut couvrir de 15 % à plus de 50 % des apports énergétiques journaliers des populations (Figure 2).

A.2. Types de rizicultures

Le riz est une plante d'origine aquatique. La riziculture est la culture céréalière la plus exigeante en eau. On peut distinguer quatre grands systèmes de rizicultures en fonction de leur hydrologie (GRiSP, 2013). Les cultures irriguées représentent environ 54 % des surfaces de riziculture et 75 % de la production mondiale de riz. C'est le mode de culture le plus présent parmi les rizicultures d'Asie. Il permet des rendements importants d'en moyenne 5,4 tonnes/hectare (t/ha). L'apport en eau contrôlé par irrigation rend cette culture moins sensible aux sécheresses mais la plus gourmande en eau ; on estime qu'elle recevrait 34 à 43 % des eaux d'irrigation dans le monde.

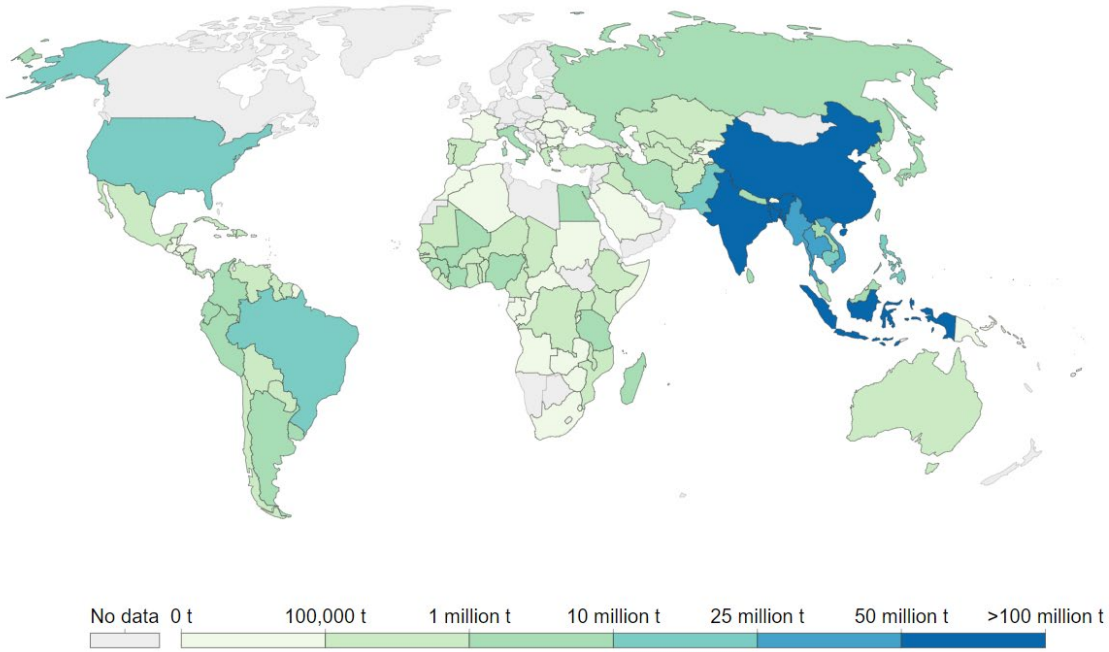


Figure 1 : Production mondiale de riz.
Représentation du niveau de production de riz en tonnes par pays. D'après les données de 2018 issues de la FAO (FAO, 2019) mises en forme par ourworldindata.org

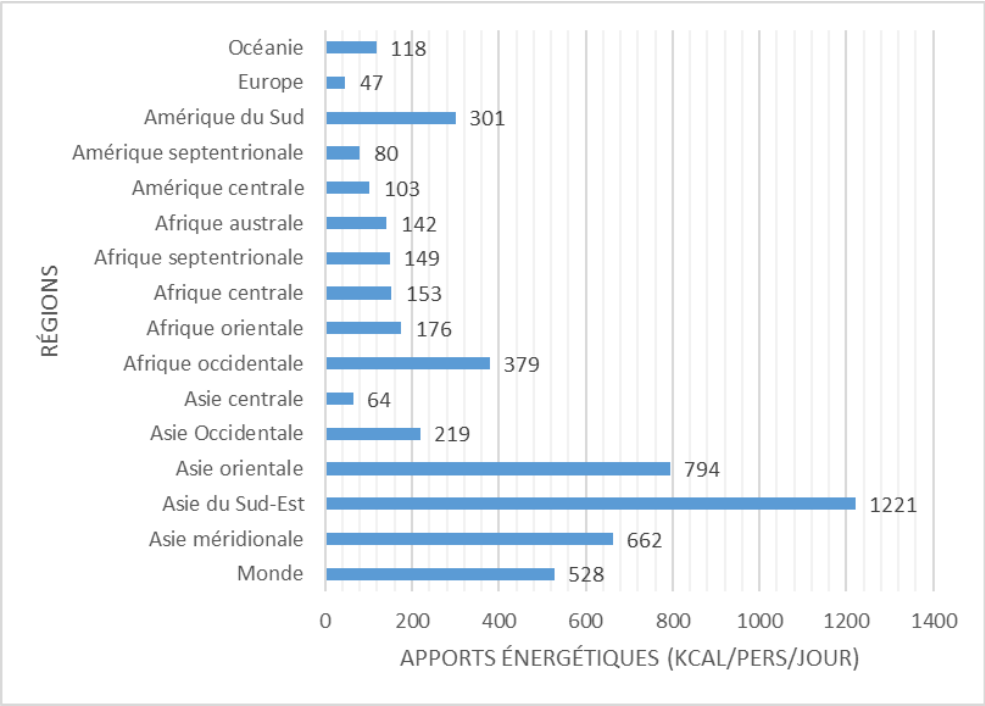


Figure 2 : Place du riz dans l'alimentation.
Représentation des apports énergétiques assurés par la consommation de riz brut et transformé en fonction des régions en 2018 (en kilocalorie par personne et par jour). L'apport énergétique standard est fixé à 1 800 et 2 400 kcal/jour pour une femme et un homme adulte respectivement. D'après les données de 2019 issues de la FAO (FAO, 2019).

Les cultures inondées concernent 30 % des surfaces cultivées pour 19 % de la production mondiale. Le riz est produit dans les mêmes conditions que pour les cultures irriguées, mais l'apport en eau est assuré par les pluies et il n'y pas ou peu de maîtrise des sorties d'eau. Ces conditions rendent les cultures aussi sensibles aux excès qu'aux déficits hydriques. De ce fait, il présente un rendement variable d'en moyenne 1 à 5 t/ha.

Les cultures pluviales concernent 9 % des surfaces de riziculture et environ 4 % de la production mondiale. Ce mode de culture est majoritaire en Afrique de l'Ouest et en Amérique Latine. L'apport en eau est exclusivement assuré par les pluies rendant ce type de cultures particulièrement sensible aux sécheresses. Environ 40 % des surfaces pluviales sont régulièrement touchées par des épisodes de sécheresse. Le rendement total est finalement assez faible avec 1 à 2,3 t/ha en moyenne.

Les cultures flottantes concernent 7 % de surfaces cultivées et 2 % de la production mondiale. Ce mode de culture présente le plus faible rendement avec en moyenne 1,5 t/ha. Il est

pratiqué sur des terres régulièrement inondées, comme au bord de mer en fonction des marées, ou inondées sur de longues périodes (de 10 jours à plusieurs mois).

A.3. Pertes agricoles affectant la production de riz

Tout au long de sa chaine de production, le riz est sujet à des pertes agricoles de différentes natures. Ces pertes ont un impact direct sur la sécurité économique des producteurs et la sécurité alimentaire des consommateurs, en particulier pour les populations d'Asie du Sud et du Sud-Est.

A.3.1. Facteurs abiotiques

La première source de perte agricole pour la riziculture est liée à des facteurs environnementaux. De par ses différents modes de cultures, le riz est une céréale sensible à la fois aux sécheresses et aux inondations. Le développement de variétés de riz adaptées à ces stress abiotiques est essentiel car ces évènements de sécheresses et d'inondations sont attendus de plus en plus fréquemment avec le dérèglement climatique en cours (FAO, 2017). Par ailleurs, l'élévation globale des températures a aussi un impact négatif sur les productions. Par exemple, les variations de rendement du riz ont pu être corrélées aux variations des températures. Ces rendements diminuent lorsque les températures augmentent du fait de l'augmentation d'un phénotype de stérilité (Oh-e et al., 2007; GRiSP, 2013).

A.3.2. Facteurs biotiques : les pathogènes du riz

Les pathogènes responsables de maladies chez le riz sont de 4 types : les parasites, les champignons (agents fongiques), les bactéries et les virus. Chaque maladie a des impacts variables sur les cultures et sur les rendements finaux (Tableau 1). L'impact des maladies fongiques semble particulièrement important avec notamment la pyriculariose (*Rice blast*) et la brûlure de la graine de riz (*Rice sheath blight*) responsables respectivement de 4,33% et 6,78% de perte globale d'après des estimations récentes (Savary et al., 2019). Les pertes moyennes de production globale de riz attribuables aux pathogènes ont été estimées à 30%, devant le blé et le maïs pour lesquelles elles sont d'environ 22%. Ces données sont stables depuis le début des années 2000 où les pertes pour le riz avaient été estimées autour de 27% (Oerke, 2006; Savary et al., 2019).

A.3.3. Interaction des facteurs abiotiques et biotiques

L'augmentation globale des températures liée au changement climatique a également été associée à une augmentation de l'incidence de certaines maladies. En effet, le réchauffement climatique favorise la propagation de certains pathogènes à de nouvelles régions (Bebber et al., 2013). Par exemple, il a été observé que la bactérie *Burkholderia glumae* (aussi appelée *Pseudomonas glumae*) responsable de la brûlure bactérienne des panicules (*bacterial panicle blight*) avait une incidence plus forte lors d'exposition à des températures élevées, avec un optimum à 35°C (Nandakumar et al., 2007; Nandakumar et al., 2009; Ham et al., 2011). Cette souche fut responsable d'épidémies très importantes dans les cultures de riz de Louisiane aux USA en 1995, 1998 et 2000 alors que ces années ont connu des températures nocturnes particulièrement importantes (Nandakumar et al., 2009). L'élévation des températures impacte également d'autres types de pathogènes comme les champignons. La décoloration du riz (*rice discolouration*) provoquée par plusieurs souches fongiques (*Curvularia lunata*, *Bipolaris oryzae*, *Pyricularia oryzae* par exemple) a une incidence plus forte sous des températures élevées (Baite et al., 2020).

A.3.4. Usage des pesticides

La protection des cultures contre les pathogènes, par l'usage de pesticides, est une des principales sources de pollution agricole. L'utilisation de pesticides a augmenté de 79% entre 1992 et 2019 dans le monde et de 85% en Asie (Figure 3A). Sur la même période, la superficie de terres cultivées en Asie n'a progressé que de 3% (Figure 3B). Il y a donc eu une augmentation massive de l'usage des pesticides depuis le début des années 90 (FAO, 2020).

Cette consommation de pesticides a été favorisée notamment par l'augmentation de la prévalence de certaines maladies comme conséquences du changement climatique, de la diminution de la diversité génétique des cultures par la sélection humaine et de pratiques culturales intensives permettant une propagation rapide des infections. Outre la pollution des sols et des eaux que cette utilisation massive de pesticides engendre, elle peut paradoxalement favoriser la survenue d'autres maladies en détruisant des prédateurs naturels de certains pathogènes. Par exemple, l'Asie du Sud-Est a connu, à la fin des années 80, une importante épidémie provoquée par le parasite *Nilaparvata lugens*. Cette importante

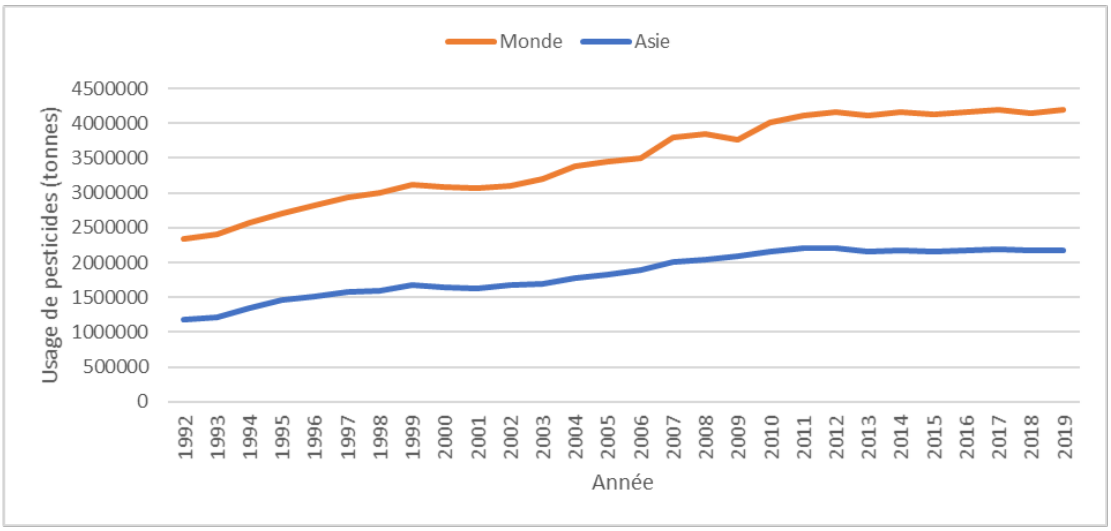
propagation du parasite a été associée à une surutilisation de pesticides éliminant ses ennemis naturels (Way and Heong, 1994).

Tableau 1 : Exemple de maladies affectant les cultures de riz et estimation du pourcentage de pertes qu'elles représentent.

Type	Maladie (EN)	Pathogène	Pertes (%)[a]
Maladies Fongiques	Pyriculariose (*Rice Blast*)	*Magnoporthe oryzae*	4.33
	Brûlure de la graine de riz (*Rice sheath blight*)	*Rhizoctonia solani*	6.78
	Taches brunes du riz (*Brown spot*)	*Cochliobolus miyabeanus*	3.77
	Faux charbon (*False smut*)	*Ustilaginoidea virens*	0.68
	Pourriture de la graine de riz (*Sheath rot*)	*Sarocladium oryzae*	0.40
Maladies Bactériennes	Brûlure bactérienne (*Bacterial blight*)	*Xanthomonas oryzae pv. oryzae*	2.72
	Brûlure bactérienne des panicules (*Bacterial panicle blight*)	*Burkholderia glumae*	0.87
Maladies Parasitaires	Foreurs de la tige de riz (*Rice stem borers*)	*Scirpophaga incertulas*	5.57
	Rice leaf folder	*Cnaphalocrocis medinalis*	1.92
	Cicadelle brune (*Brown plant hopper*)	*Nilaparvata lugens*	1.31
Maladies Virales	Rayure du riz (*Rice stripe*)	*Rice stripe tenuivirus*	0.27
	Panachure jaune du riz (Rice yellow mottle)	*Rice yellow mottle virus*	0.08

[a] pourcentage de pertes estimées par rapport à la production globale. Savary *et al.* 2019

A.



B.

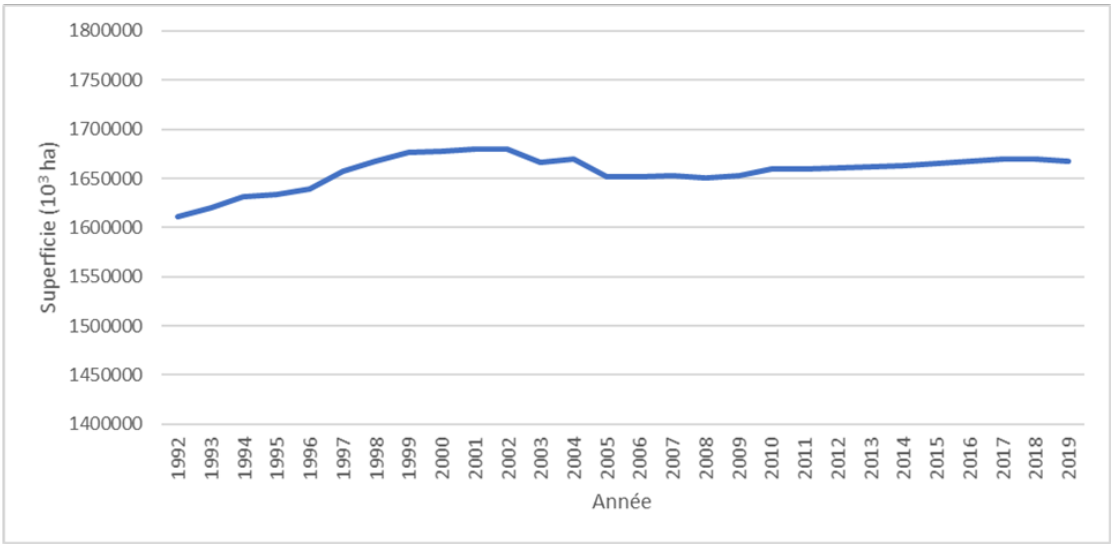


Figure 3 : Evolution de l'usage de pesticides et de la superficie des terres cultivées entre 1992 et 2019.
(A) Quantité de pesticides utilisés à des fins agricoles dans le monde et en Asie. (B) Evolution de la superficie des terres cultivées en Asie. D'après les données de la FAO 2020 (FAO, 2020).

B. Origine, domestication et diversité du genre *Oryza* et de l'espèce modèle *Oryza sativa*

Le riz est une céréale monocotylédone de la famille des Poacées. Les premiers représentants du genre *Oryza* auraient émergé il y a 13 à 15 millions d'années, correspondant au milieu de l'ère Miocène (GRiSP, 2013). Le genre *Oryza* comprend 27 espèces caractérisées pouvant être séparées en 11 groupes distincts d'après leurs caractéristiques génomiques (Figure 4) (Jacquemin et al., 2013; Stein et al., 2018). On distingue six groupes comprenant des espèces à génomes diploïdes (n=12) (AA, BB, CC, EE, FF, GG) et cinq groupes comprenant des espèces à génomes allotétraploïdes (n=24) (BBCC, CCDD, HHJJ, HHKK, KKLL). Le groupe diploïde de génome AA comprend les deux seules espèces cultivées, le riz asiatique *O. sativa*

et le riz africain *O. glaberrima*. Il comprend également 6 espèces sauvages, dont les progéniteurs des espèces cultivées : *O. rufipogon* pour le riz asiatique et *O. barthii* pour le riz africain (Figure 5).

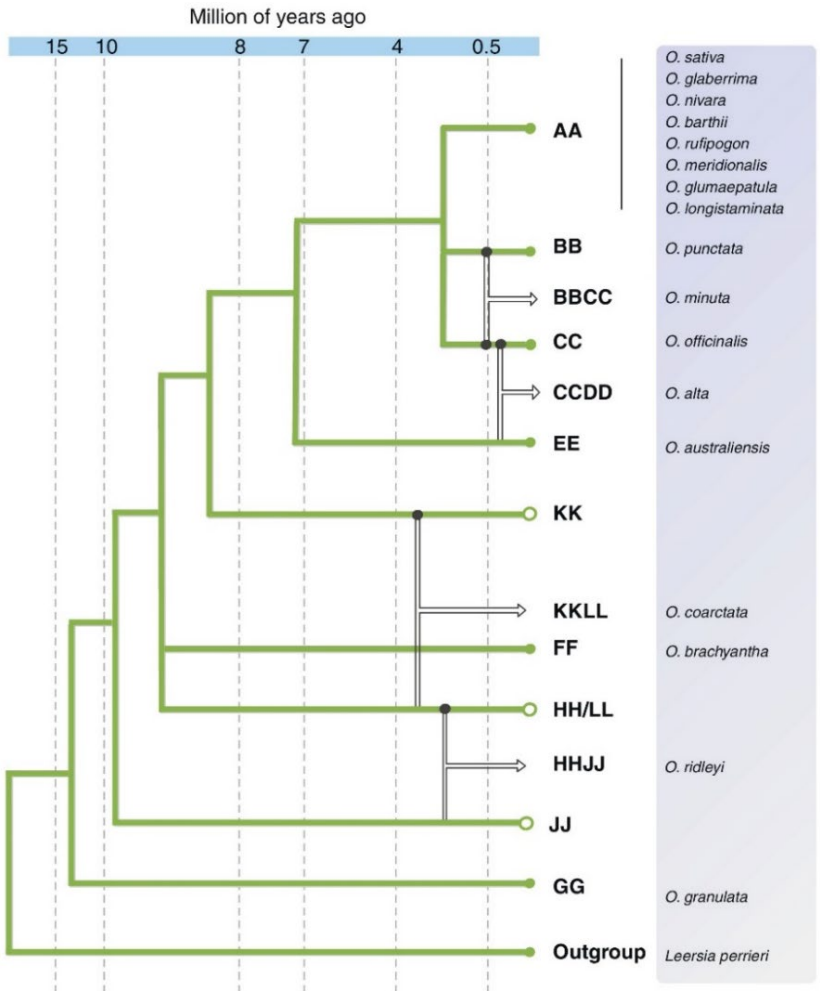


Figure 4 : Phylogénie du genre Oryza.
D'après Jacquemin et al. (2013)

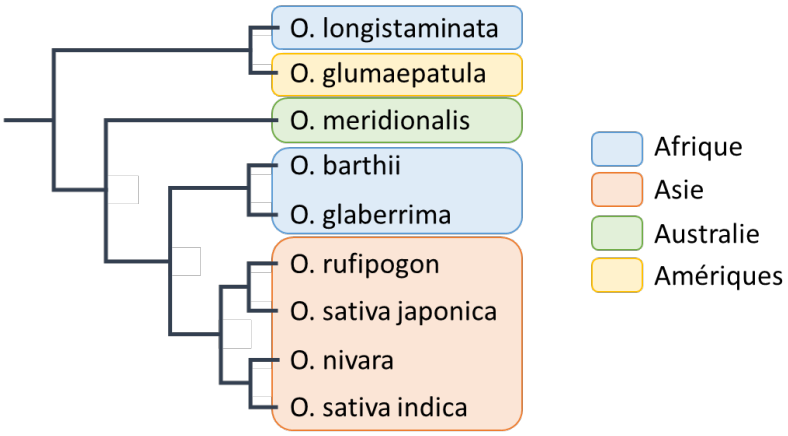


Figure 5 : Phylogénie des espèces de riz AA basée sur les séquences chloroplastiques.
Adaptée de Wambugu et al. (2015)

B.1. Domestication de l'espèce asiatique *Oryza sativa*

La domestication de l'espèce *O. sativa* aurait débuté il y a 8 à 9 000 ans en Asie à partir de différentes sous-populations de *O. rufipogon* (Huang et al., 2012). Les cultivars de l'espèce asiatique *O. sativa* sont séparés en deux grandes sous-espèces : *O. sativa* ssp. *japonica* et *O. sativa* ssp. *indica*. De nombreuses études évoquent un « paradoxe du riz », où des éléments contradictoires quant à la domestication des deux sous-espèces asiatiques, *japonica* et *indica*, ont pu être observés. Ces éléments aboutissent à la formation de deux hypothèses majeures (Choi et al., 2017; Chen et al., 2019). La première suggère une double origine (ou 'Indépendance') des domestications (Oka, 1988). Le riz *japonica* serait issu de la domestication d'une population ancestrale de *O. rufipogon* dans une région du sud de la Chine, puis la sous-espèce *indica* aurait été domestiquée plus tardivement au nord de l'Inde à partir d'une forme annuelle de *O. rufipogon*, aussi appelée *O. nivara*. Cette hypothèse s'appuie sur le fait que les deux populations présentent une différenciation génétique très marquée. Différentes études s'intéressant aux régions inter-géniques ou aux éléments transposables montrent que les cultivars étudiés de *japonica* et *indica* sont respectivement plus proches de différentes accessions de *O. rufipogon* et *O. nivara* qu'ils ne le sont entre eux (Cheng et al., 2003; Huang et al., 2012; Zhao et al., 2018). De plus, les croisements pour former des hybrides *japonica x indica* sont particulièrement difficiles à réaliser et les descendants sont généralement stériles. En 2015, Civan *et al.* (Civan et al., 2015) suggèrent même une triple domestication en incluant les cultivars *aus* (actuellement classés dans la sous-espèce *indica*) qu'ils estiment issus d'une domestication indépendante.

La seconde hypothèse suppose une origine unique des deux sous-espèces domestiquées. Après domestication des formes *japonica* selon le schéma précédent, les génotypes domestiqués précoces « proto-*japonica* » se seraient croisés localement au nord de l'Inde avec des populations annuelles sauvages *O. nivara* pour donner les formes *indica* (Huang and Han, 2015; Chen et al., 2019). Ainsi, il n'y aurait eu qu'un seul évènement de domestication suivie d'une différenciation de la population originelle « proto-*japonica* » en sous-populations *japonica* et *indica* (Vaughan et al., 2008). Cette hypothèse s'appuie sur des études réalisées sur les allèles associés à la domestication. Les différents auteurs retrouvent ces allèles partagés entre *japonica* et *indica* et estiment qu'ils proviendraient d'un seul évènement de domestication (Sang and Ge, 2007; Huang et al., 2012).

B.2. Diversité de l'espèce domestiquée *Oryza sativa*

La diversité des cultivars asiatiques est extrêmement large avec plus de 150 000 variétés recensées qui s'organisent en une phylogénie particulièrement complexe. En 1987, Glaszmann (Glaszmann, 1987) suggère que l'espèce *O. sativa* peut être subdivisée en six groupes sur la base de marqueurs enzymatiques. Plus tard, Garris *et al.* (Garris et al., 2005) subdivisent les deux sous-espèces *japonica* et *indica* en cinq groupes variétaux : les indica, les aus/boro et les riz aromatiques (basmati) pour la sous-espèce *indica*, les *japonica* tempérés et les *japonica* tropicaux pour la sous-espèce *japonica*. Cette subdivision des variétés cultivées est confirmée par l'étude de 2014 portant sur l'analyse de SNP de 3 000 génomes de riz (Rice Genome Project, 2014). Les *japonica* (tempérés et tropicaux) sont génétiquement très proches, présentant une diversité génétique plus faible que les groupes d'*indica* (Garris et al., 2005). Cette diversité importante observée entre les différents groupes variétaux d'*indica* est à l'origine de l'hypothèse d'une triple domestication du riz asiatique développée par Civan *et al*. (Civan et al., 2015).

Quel que soit le scenario de domestication, les compartiments sauvages constituent un réservoir de diversité génétique pour l'amélioration des cultivars sur différents aspects : résistance aux stress hydriques (inondation, sècheresse), adaptation aux changements de température, aux polluants (contamination des sols aux métaux lourds, aux nitrates, etc.) et résistance aux stress biotiques (pathogènes). Il existe plusieurs exemples d'exploitation des espèces sauvages pour améliorer le niveau de résistances aux stress biotiques. Notamment, la variété cultivée *O. sativa japonica cv.* Kitaake a été modifiée pour intégrer deux copies du gène *XA21* (Song et al., 1995) sur le chromosome 6 provenant de l'espèce sauvage *O. longistaminata*. Ce gène confère une résistance accrue au pathogène *Xanthomonas oryzae* pv *oryzae* responsable de la brûlure bactérienne du riz (*rice bacterial blight*). La variété ainsi modifiée est appelé KitaakeX (Jain et al., 2019).

B.3. Le riz, espèce modèle

L'espèce domestiquée *Oryza sativa* a été choisie comme espèce modèle pour les monocotylédones, à l'image d'*Arabidopsis thaliana* pour les dicotylédones. Le génome de l'espèce *O. sativa* est le plus petit des principales céréales cultivées avec 430 Mb (méga-bases,

10^6 bases) environ alors que le génome du maïs fait plus de 2 600 Mb et celui du blé tendre, espèce hexaploïde, plus de 16 000 Mb.

O. sativa ssp *japonica* cv. Nipponbare est le premier génome de riz complètement séquencé et assemblé par le consortium IRGSP *(International Rice Genome Sequencing Project*). L'assemblage complet est publié en 2005 (IRGSP and Sasaki, 2005), moins d'un an après la version finale du génome humain (International Human Genome Sequencing Consortium, 2004). Depuis 2013, suite à l'association de deux projets majeurs, l'IRGSP d'une part, et le *Rice Genome Annotation Project* de l'université du Michigan (MSU) d'autre part, une séquence de référence de très haute qualité est disponible pour le cultivar Nipponbare (IRGSP-1.0) (Kawahara et al., 2013). L'annotation de référence pour cette séquence, RAP-DB (*Rice Annotation Project database*) (Rice Annotation Project, 2008), est fournie par l'IRGSP. Deux autres annotations sont également disponibles et largement utilisées : l'une fournie par la MSU (v7.0) et l'autre fournie par le NCBI (release 102).

Depuis, de nombreux génomes de riz ont pu être séquencés au travers de projets de recherche s'intéressant autant aux variétés cultivées qu'aux espèces sauvages (Jacquemin et al., 2013; Rice Genome Project, 2014; Stein et al., 2018; Wang et al., 2018; Fuentes et al., 2019). Cet essor est favorisé par l'évolution des technologies de séquençage, devenant plus fiables, moins chères et plus accessibles. Le nombre croissant de génomes séquencés pour le riz a ouvert la voie à des analyses globales, comparant des dizaines voire des centaines de génomes en parallèle. Ces données sont également à l'origine de l'essor de la pangénomique dont le but est d'analyser l'ensemble des gènes et allèles portés par les différents génomes de riz pour tenter de capturer l'ensemble de la diversité génétique à différentes échelles au sein du genre *Oryza* (Monat et al., 2017; Cubry et al., 2018; Wang et al., 2018; Zhao et al., 2018).

C. Structure, fonction et diversité des récepteurs LRR

Les mécanismes de signalisation cellulaire sont essentiels à la vie et la survie de toutes les cellules. De nombreuses fonctions biologiques, comme le développement ou la spécialisation cellulaire ou encore la défense contre les pathogènes, reposent sur ces mécanismes. L'initiation d'une réponse cellulaire à un signal quelconque nécessite des récepteurs capables de percevoir ce signal et d'y répondre. Parmi eux, trois familles de récepteurs portant un domaine *Leucine-Rich Repeat* (LRR) sont particulièrement abondantes dans les génomes de

plantes : les LRR-RLK (*LRR Receptor-Like Kinase*), les LRR-RLP (*LRR Receptor-Like Protein*) et les NLR (*Nucleotide Binding-Site LRR* ; aussi appelés NBS-LRR) (Baker et al., 1997). Le domaine LRR est un domaine protéique spécialisé dans les interactions protéine-protéine. Il est composé de 2 à plus de 30 motifs LRR, chacun d'environ 24 acides aminés (cf. section D.1.), répétés en tandem, permettant une modularité importante. Le domaine LRR est commun à plusieurs familles de gènes chez les plantes. On le trouve également dans les protéines des familles F-box-LRR (FBL), Extensin-like LRR (LRX), *plant intracellular Ras group-related LRR* (PIRL) (Forsthoefel et al., 2005) ou *polygalacturonase-inhibiting protein* (PGIP). Les spécificités structurales des différentes protéines et les domaines fonctionnels associés aux domaines LRR permettent d'identifier et de classer ces protéines dans les différentes familles de gènes (Figure 6).

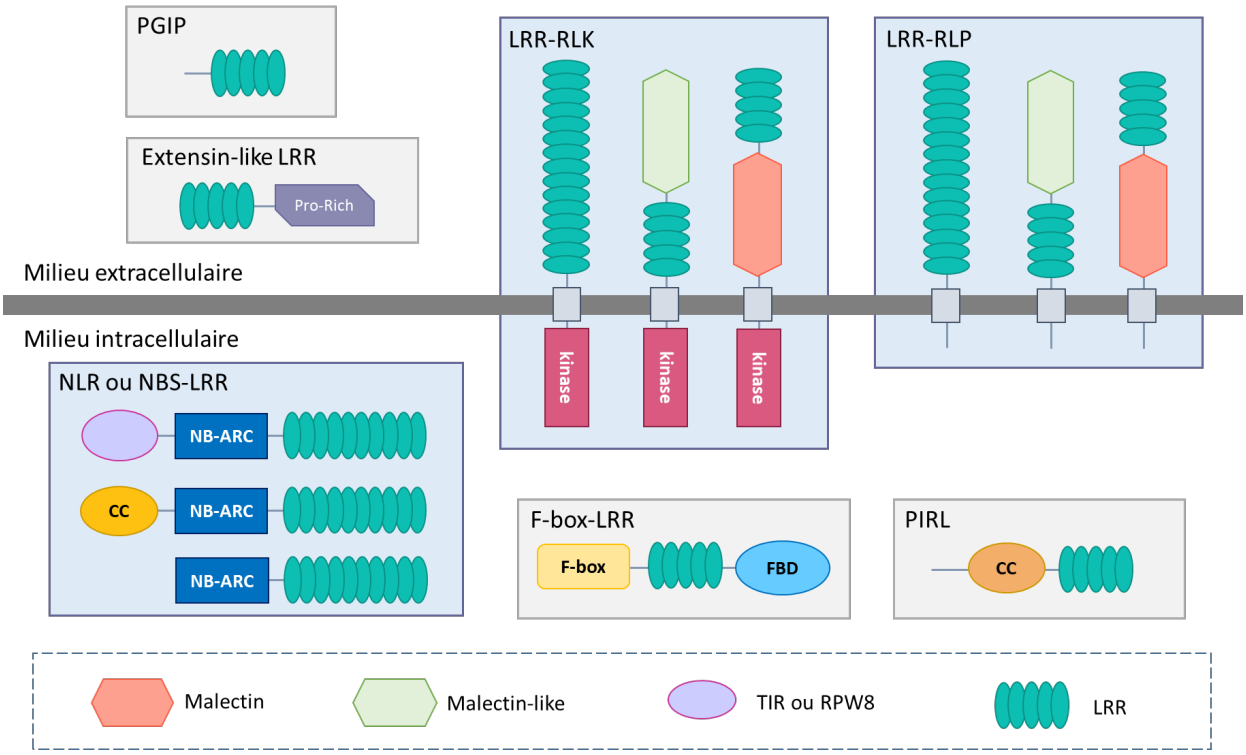


Figure 6 : Représentation schématique de la variabilité structurale des principales familles de récepteurs LRR.

C.1. Récepteurs Membranaires LRR-RLK et LRR-RLP

Les gènes LRR-RLK et LRR-RLP sont des récepteurs transmembranaires. La structure des protéines codées par ces gènes est analogue entre les deux familles et se divise en trois parties : (i) une section extracellulaire appelée « ectodomaine » et portant le domaine LRR ; (ii) un domaine transmembranaire (TM) hydrophobe enchâssé dans la membrane plasmique ; et (iii) une section intracellulaire appelée « endodomaine ». LRR-RLK et LRR-RLP se distinguent par leur partie intracellulaire. Les LRR-RLK portent un domaine kinase et les LRR-RLP présentent une courte section intra-cytoplasmique sans domaine fonctionnel caractérisé (Dodds and Rathjen, 2010; Sekhwal et al., 2015; Couto and Zipfel, 2016; Boutrot and Zipfel, 2017) (Figure 6).

C.1.1. LRR-RLK

Chez les plantes, la famille des LRR-RLK appartient à la grande famille des gènes RLK. Les RLK sont des récepteurs présentant différents types d'ectodomaines associés à la kinase. Ceux dont l'ectodomaine est un domaine LRR sont les plus abondants (Shiu and Bleecker, 2001b). La diversité des domaines extracellulaires associés au domaine kinase des RLK, et l'origine de ces structures au cours de l'évolution des plantes, a fait l'objet d'un travail de revue (Dievart et al., 2020) présenté dans le chapitre 4 de cette thèse. Les RLK constituent une famille hautement multigénique de récepteurs pouvant comporter plusieurs centaines de membres par espèce. Chez *A. thaliana* par exemple, ce sont plus de 610 gènes RLK qui ont été identifiés, dont 216 LRR-RLK (Shiu and Bleecker, 2001a). Les RLK, et LRR-RLK, sont impliqués dans de nombreuses voies de signalisation comme les réponses aux stress abiotiques (salinité, inondation, lésions), les réponses aux stress biotiques (résistance aux pathogènes notamment) et les processus développementaux par exemple. Le domaine LRR représente l'ectodomaine majoritaire des récepteurs RLK impliqués dans les réponses aux stress biotiques. Mais trois autres ectodomaines pouvant être associés au domaine kinase ont également été identifiés dans ce contexte : le domaine *LysM*, le domaine *Lectin*, et le domaine *EGF-Like*. Une partie de ces gènes a été caractérisée d'un point de vue fonctionnel (Baker et al., 1997; Couto and Zipfel, 2016; Boutrot and Zipfel, 2017).

Le premier gène LRR-RLK caractérisé chez *A. thaliana* dans le cadre des résistances aux stress biotiques est *FLS2* (*Flagellin-Sensing 2*) (Gomez-Gomez and Boller, 2000). Ce gène est capable

de reconnaitre le motifs *flg-22* provenant de la flagelline et de déclencher une réponse immunitaire. La flagelline est une protéine constituant les flagelles bactériens. On la retrouve chez un grand nombre d'espèces bactériennes permettant aux gènes *FLS2* de conférer un phénotype de résistance à large spectre.

Les LRR-RLK, comme l'ensemble des RLK chez les plantes, possèdent un domaine kinase Ser/Thr formant un groupe monophylétique. Sur la base de cette phylogénie du domaine kinase, les gènes LRR-RLK peuvent être séparés en 15 clades, LRR-I à LRR-XV, certains étant subdivisés en sous-clades (Dufayard et al., 2017). Il a été remarqué que cette séparation en clade sur la base du domaine kinase respecte la diversité globale des gènes d'un point de vue structural (domaines, intron/exon) (Shiu and Bleecker, 2001a; Liu et al., 2017). Cette analogie de structure entre gènes d'un même clade suggère que ces gènes se sont dupliqués au cours de l'évolution de cette famille. Par ailleurs, il existe un déséquilibre important du nombre de gènes appartenant aux différents clades pour beaucoup d'espèces de plantes. Par exemple, chez *A. thaliana*, le clade LRR-IV comporte 4 gènes tandis que le clade LRR-III en comprend 47 (Dufayard et al., 2017). Ces différences révèlent des évènements d'expansion massive de gènes dans certains clades. Il est noté cependant que ces expansions ne se sont pas toujours produites selon le même schéma au cours de l'évolution des différents phylums de plantes. Tandis que certains clades (LRR-III, LRR-XI) montrent une expansion ancienne antérieur à l'émergence des angiospermes, d'autres ont connu une expansion plus récente spécifique à certaines espèces seulement (LRR-XII) (Dufayard et al., 2017).

C.1.2. LRR-RLP

Les LRR-RLP présentent une structure analogue au LRR-RLK dans la région extracytoplasmique. Ils présentent un domaine LRR extracellulaire et un domaine TM (Figure 6). Le premier gène LRR-RLP identifié et caractérisé fut *Cf-9*. Ce gène identifié chez la tomate confère une résistance au champignon *Cladosporium fulvum* (Jones et al., 1994).

Chez *A. thaliana*, 56 (Fritz-Laylin et al., 2005) ou 57 LRR-RLP (Wang et al., 2008) ont été identifiés. Une partie de ces LRR-RLP présente un domaine LRR très homologue au domaine LRR de gènes LRR-RLK (Shiu and Bleecker, 2003). Par ailleurs, plusieurs LRR-RLP sont localisés au sein de clusters de gènes LRR-RLK. Cela suggère qu'au moins certains de ces gènes LRR-RLP ont une origine commune récente avec des LRR-RLK. Deux scenarios sont alors possibles : les

LRR-RLP pourraient être issus de copies de gènes LRR-RLK, elles-mêmes issues de duplication, ayant perdu le domaine kinase ou inversement les LRR-RLK pourraient être issus de LRR-RLP qui auraient fusionné avec un domaine kinase (par le biais de recombinaisons non-homologues par exemple). Dans le premier scenario, soit la duplication du gène LRR-RLK est partielle, n'emportant que le domaine LRR, soit elle est complète et un autre évènement (de type délétion) est à l'origine de l'élimination du domaine kinase. Ce scénario a été déduit des observations faites pour le gène *XA21D* chez le riz, un LRR-RLP homologue au gène *XA21*. Ce gène est présent dans un cluster. Une duplication d'un gène précurseur aurait fait émerger *XA21* et *XA21D*. Le gène *XA21D* aurait alors perdu le domaine kinase suite à l'insertion d'un élément transposable (Wang et al., 1998).

C.1.3. Récepteurs et corécepteurs

A la suite de la liaison du ligand au niveau extracellulaire, le récepteur membranaire doit transmettre l'information à l'intérieur de la cellule. Ce mécanisme est appelé « transduction du signal ». Les récepteurs LRR nécessitent l'intervention d'un ou plusieurs partenaires membranaires, appelés corécepteurs, pour transmettre le signal via un mécanisme de transphosphorylation des domaines kinases (Figure 7) (Couto and Zipfel, 2016). Les modèles de transduction du signal par les LRR-RLK et les LRR-RLP suggèrent que les LRR-RLK nécessitent l'intervention d'un seul corécepteur, car ils possèdent un domaine kinase intrinsèque. En revanche, les LRR-RLP, dépourvu de domaine kinase, vont nécessiter l'intervention conjointe de deux corécepteurs. Ces corécepteurs présentent souvent un domaine LRR court composé de 3 à 5 motifs LRR.

La formation du complexe protéique nécessite la liaison du ligand au récepteur. Généralement le ligand participe activement à la dimérisation en stabilisant la liaison du récepteur et du corécepteur par un mécanisme appelé « *molecular glue* » (colle moléculaire) (Couto and Zipfel, 2016). Cet effet a notamment été observé entre *FLS2* et son corécepteur *BAK1* (*Brassinosteroid insensitive 1-associated kinase*) chez *A. thaliana* (Sun et al., 2013).

BAK1 est aussi appelé *SERK3* (*Somatic Embryogenesis Receptor-like Kinase 3*). Les gènes *SERK* constituent une famille de corécepteurs LRR-RLK (Hecht et al., 2001). Ils interviennent dans des voies de signalisations diverses en s'associant à différents récepteurs pour assister leur fonction (Aan den Toorn et al., 2015; Ma et al., 2016). Par exemple, *BAK1* interagit avec *FLS2*

mais aussi *BRI1* (*Brassinosteroid Insensitive 1*) (Li et al., 2002). Il intervient donc dans les mécanismes de défense, en assistant *FLS2,* et de développement, en assistant *BRI1*.

BAK1 intervient également pour la signalisation via des récepteurs LRR-RLP. Il a été identifié comme l'un des corécepteurs du gène *RLP23* avec *SOBIR1* (*Suppressor of BAK1-Interacting Receptor kinase 1*). Le complexe reconnait le motif *nlp20* présent dans la plupart de protéines NLPs (*Necrosis and ethylene-inducing peptide 1-like proteins*) (Albert et al., 2015). *SOBIR1* et *RLP23* ont la particularité de former un complexe qui n'est pas dépendant du ligand. Seule l'association tripartite avec *BAK1* est ligand-dépendante. En revanche, *SOBIR1* est nécessaire pour la liaison du ligand avec le récepteur *RLP23* et pour le recrutement de *BAK1*. Les corécepteurs de type *SOBIR1* sont parfois appelés « régulateurs ».

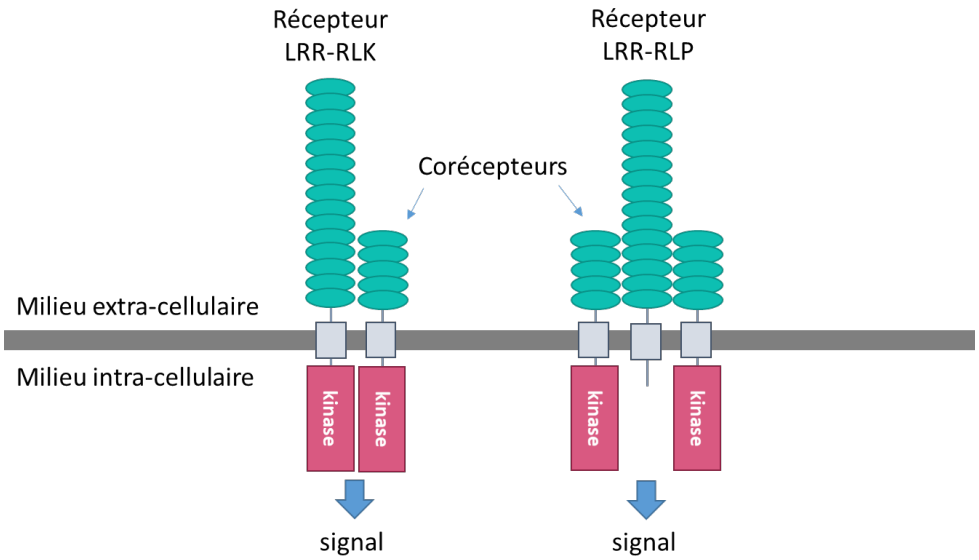


Figure 7 : Représentation schématique des complexes de récepteurs et corécepteurs à LRR pour la transduction du signal.

C.2. Récepteurs intracellulaires NLR

Les récepteurs NLR (*Nucleotide-Binding Site and Leucine-Rich Repeat Receptors*, aussi appelés NBS-LRR) sont des récepteurs intracellulaires présentant un domaine NB-ARC central suivi d'un domaine LRR en C-ter (Figure 6). A ce jour, ces récepteurs sont essentiellement étudiés pour leur rôle dans la reconnaissance d'effecteurs dans des mécanismes de résistances aux pathogènes chez les plantes. Ils sont également appelés « *R genes* », pour *Resistance genes*.

C.2.1. Diversité de structure des gènes NLR

Différentes structures pour ces récepteurs ont été décrites en fonction des domaines trouvés en amont du domaine NB-ARC (Figure 8). Notamment, trois structures principales peuvent être distinguées (Shao et al., 2016b) : les NLR présentant un domaine CC (*Coil coiled*), également appelés CNL, les NLR présentant un domaine TIR (*Toll-Interleukin Receptor*) (Whitham et al., 1994), aussi appelés TNL, et dans une moindre mesure, les NLR présentant un domaine RPW8 (*Resistance to Powdery Mildew 8*) (Zhong and Cheng, 2016), également appelés RNL.

Les récepteurs NLR sont généralement très nombreux dans les génomes de plantes mais présentent une grande variabilité dans leur abondance (Shao et al., 2016a; Shao et al., 2016b). Par exemple, plus de 150 gènes ont été identifiés chez *A. thaliana* alors que près de 500 sont recensés chez le riz (Meyers et al., 2003; Monosi et al., 2004; Shao et al., 2016b). Par ailleurs, chaque structure n'est pas présente chez toutes les espèces. On peut noter en particulier l'absence des récepteurs TNL des génomes des monocotylédones (Meyers et al., 1999; Cannon et al., 2002). En revanche, ces trois structures ont été identifiées dans le génome de *Amborella trichopoda*, une angiosperme basale, reflétant une origine ancienne probablement antérieure à l'émergence des angiospermes (Shao et al., 2016b).

En plus des récepteurs présentant une structure « classique », certains gènes ont été identifiés portant des domaines supplémentaires intégrés à leur structure. On peut donner l'exemple du gène *RRS1* chez *A. thaliana* portant un domaine WRKY en C-ter du domaine LRR (Deslandes et al., 1998). Ces domaines additionnels par rapport à la structure protéique standard sont appelés '*integrated domains*' (ID). Ils peuvent se positionner à différents endroits dans la structure protéique des NLR (Figure 8).

C.2.2. Reconnaissance directe et indirecte

La médiation d'une réponse immunitaire par les NLR nécessite la reconnaissance d'un signal infectieux intracellulaire. Cette reconnaissance peut faire intervenir une interaction directe entre l'effecteur du pathogène et le récepteur NLR, ou une interaction indirecte faisant intervenir des molécules intermédiaires (Figure 9) (Cesari, 2018; Monteiro and Nishimura, 2018).

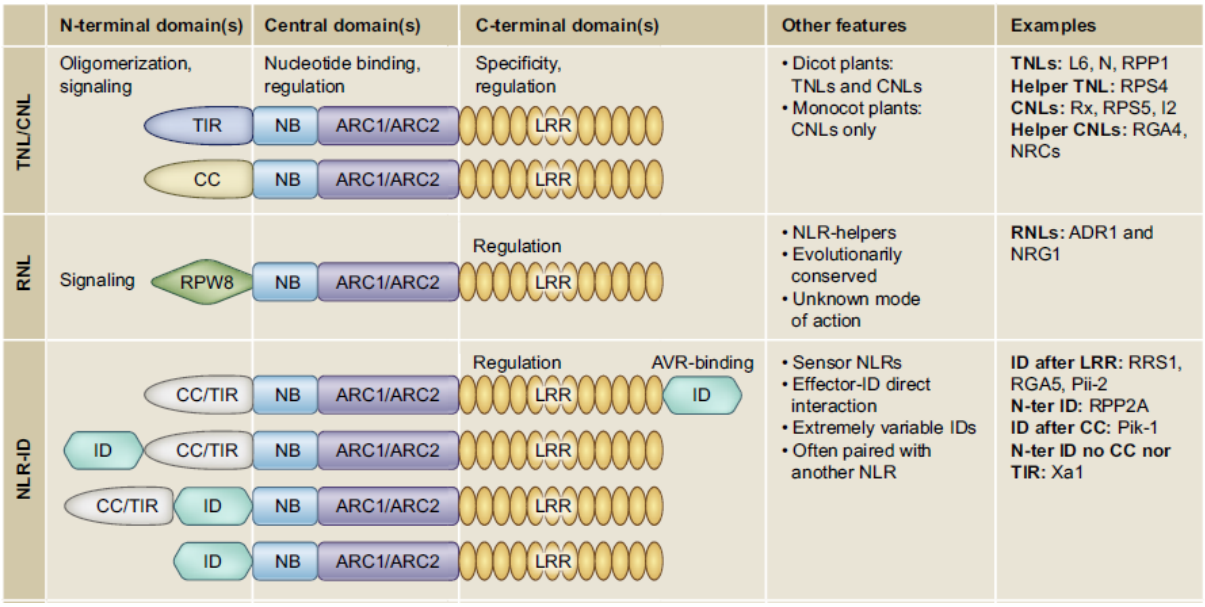


Figure 8 : Diversité structurale des récepteurs NLR chez les plantes.
D'après S. Cesari (2018).

Lorsque les NLR sont capables d'identifier les effecteurs des pathogènes, généralement via leur domaine LRR, on parle de reconnaissance directe (Figure 9a). Cette reconnaissance directe génère une pression de sélection forte chez les pathogènes pour que leurs effecteurs acquièrent des mutations leur permettant d'échapper à cette reconnaissance. Pour maintenir les capacités de défense et contrer les mécanismes d'évitement qui émergent chez les pathogènes, certains NLR fonctionnent par une reconnaissance indirecte des effecteurs. Deux systèmes existent chez les plantes pour cette reconnaissance indirecte. Dans le premier, le récepteur NLR est capable d'identifier une molécule du soi modifiée par un effecteur pathogène. On dit que le NLR « garde » la cible de l'effecteur, et cette cible est appelée '*guardee*' (Figure 9b). Dans le second système, la plante crée une copie de la cible naturelle de l'effecteur, un leurre appelé '*decoy*' en anglais. Ce leurre présente une affinité élevée pour un récepteur NLR lorsqu'il est ciblé par l'effecteur (Figure 9c). Dans ces deux systèmes, le récepteur NLR interagit avec une molécule du soi, la molécule *guardee* ou la molécule *decoy*, modifiée par l'effecteur pathogène, pour initier une réponse immunitaire.

Le dernier mécanisme connu, intervenant dans la reconnaissance des effecteurs par les NLR, implique également une interaction directe entre le récepteur NLR et son effecteur cible. En revanche, cette interaction passe par un domaine annexe, non commun, intégré à la structure du NLR appelé '*integrated domains*' ou '*integrated decoy*'. Ce domaine constitue un leurre

pour l'effecteur pathogène qui va directement interagir avec le NLR (Figure 9d) (Kroj et al., 2016).

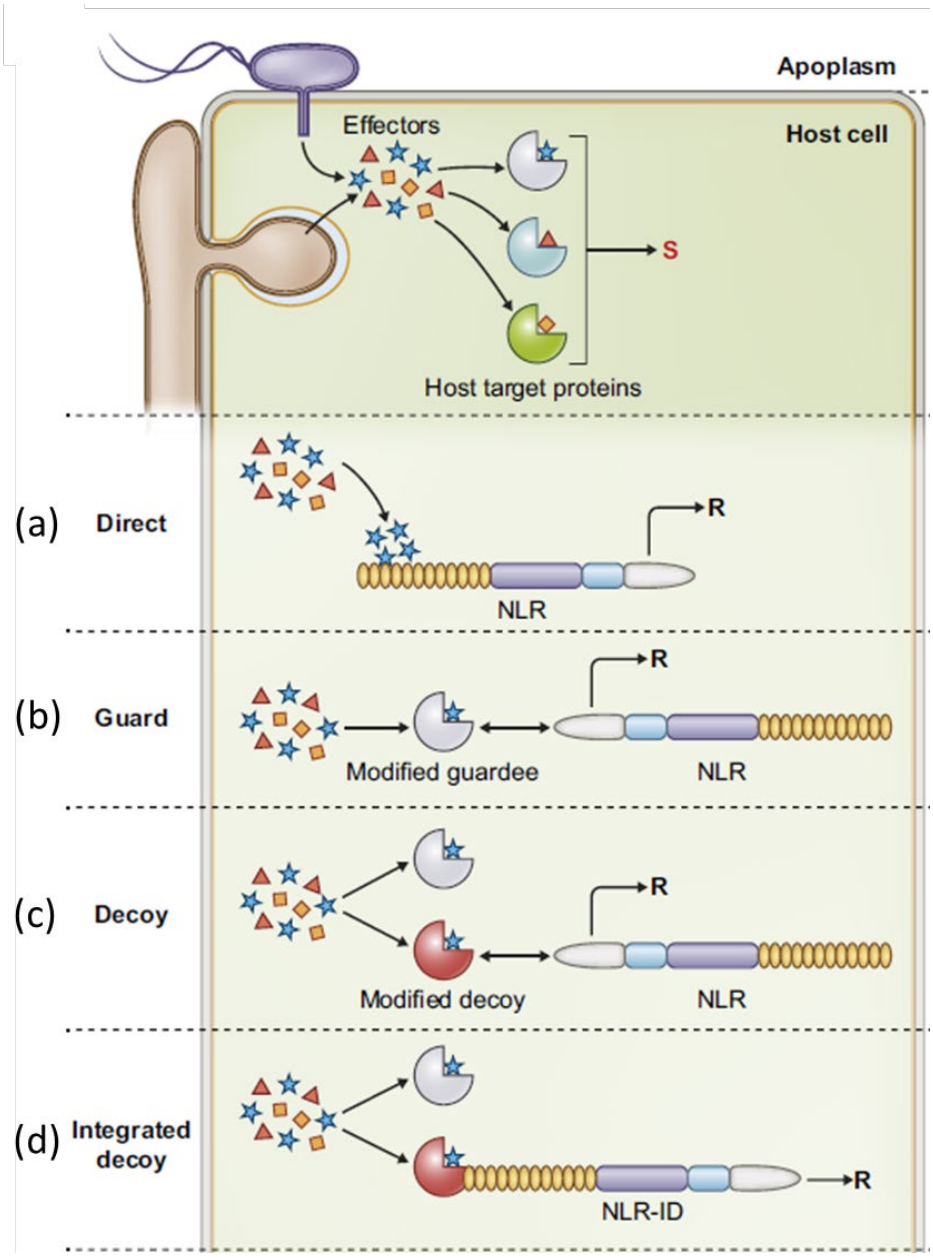


Figure 9 : Différentes stratégies de reconnaissance des effecteurs pathogènes par les récepteurs NLR.
D'après S. Cesari (2018).

C.3. L'immunité chez les plantes

Les trois familles de récepteurs LRR décrites précédemment ont un rôle central dans les mécanismes de résistance chez les plantes. Ils constituent deux lignes de défense distinctes permettant une surveillance complète des plantes : la première portée principalement par des LRR-RLK et des LRR-RLP permet une identification précoce des invasions au niveau des membranes cytoplasmiques. Ces récepteurs sont appelés PRR (*Pattern Recognition Receptors*). Les PRR permettent de déclencher une réponse immunitaire de type PTI (PAMP-*triggered immunity*) après l'identification de motifs pathogènes, appelés PAMP (*pathogen-associated molecular pattern*) ou MAMP (*microbe-associated molecular pattern*), ou de motifs du soi spécifiques d'un signal infectieux, appelés DAMP (*damage-associated molecular pattern*) (Dodds et al., 2006; Couto and Zipfel, 2016; DeFalco and Zipfel, 2021). Les récepteurs membranaires, constituant principalement une défense à spectre large (*broad spectrum*), sont facilement contournés par les pathogènes. Ces derniers utilisent des effecteurs pouvant inhiber les récepteurs ou interférer avec les partenaires intracellulaires des voies de signalisation (Figure 10) (Dodds and Rathjen, 2010).

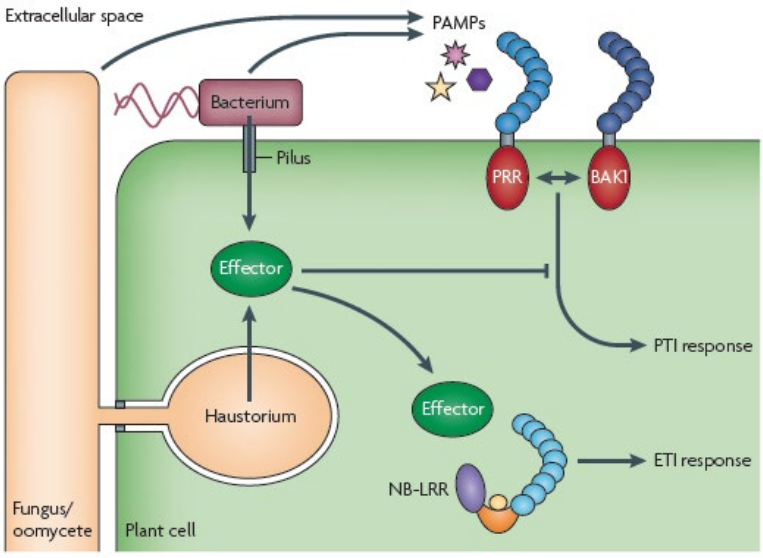


Figure 10 : Représentation schématique des deux lignes de défenses chez les plantes.
D'après Dodds et Rathjen (2010).

En réponse à l'infection, les NLR vont déclencher une réponse immunitaire forte appelée ETI (*Effector triggered immunity*) aboutissant souvent à la mort de la cellule infectée (Cui et al., 2015). Les NLR sont capables d'identifier les effecteurs pathogènes ou les éléments du soi modifiés par l'activité des pathogènes au niveau intracellulaire. Ils constituent ainsi une deuxième ligne de défense, souvent qualifiée de spécifique car chaque récepteur identifie une cible spécifique d'un pathogène (Cesari, 2018). Cette évolution conjointe des deux lignes de défenses et des pathogènes est appelée « course à l'armement » (*arms race*) (Dawkins and Krebs, 1979; Holub, 2001) et implique de nombreux mécanismes évolutifs et adaptatifs au niveau génomique.

C.4. Evolution des gènes LRR

Chez les plantes, on ne connait pas de système de défense médié par des molécules circulantes comme les anticorps chez les animaux. Chaque cellule doit être capable de produire ses propres défenses et doit présenter à chaque instant un arsenal complet de protéines lui permettant de répondre à diverses attaques de pathogènes.

La capacité d'une plante à résister aux attaques de pathogènes est donc définie par le répertoire de gène de résistance codé par son génome. Ce répertoire, majoritairement composé de récepteurs à LRR, évolue à deux échelles différentes et implique des mécanismes différents. Tout d'abord, à l'échelle du génome, on observe une évolution du nombre de gène de résistance par des mécanismes favorisant l'expansion du répertoire de gènes au cours du temps. Ensuite, à l'échelle des gènes eux-mêmes, on observe une évolution de leurs fonctions (notamment de reconnaissance) par des mécanismes favorisant la diversification des copies de gènes et des allèles.

C.4.1. Expansion du répertoire de gènes de résistance

Le principal mécanisme permettant l'expansion du nombre de copies des gènes de résistance dans les génomes est la duplication. Outre les duplications complètes de génome (*whole genome duplication*), le nombre de gènes augmente principalement suite à des duplications en tandem et des duplications segmentales. Lors de la duplication en tandem, la copie dupliquée apparait physiquement proche de la copie originale dans le génome. Lors de la duplication segmentale, la copie dupliquée apparait à une position éloignée de la copie originale, parfois sur un autre chromosome (Leister, 2004). Les récepteurs à LRR sont

particulièrement sujets aux duplications en tandem ce qui a pour conséquence de former de nombreux clusters contenant jusqu'à plusieurs dizaines de copies (Baumgarten et al., 2003; Meyers et al., 2003).

Suite à ces duplications, dans la plupart des cas, les forces de sélection purifiantes, œuvrant au maintien des fonctions des gènes au cours du temps, se relâchent pour les copies dupliquées. Ainsi, les copies de gènes dupliqués peuvent accumuler des mutations avec une fréquence plus élevée que ce qui est observé sur les autres gènes non dupliqués. Des mutations non-sens vont alors pouvoir se fixer et, à terme, rendre les copies non-fonctionnelles, parfois appelées pseudogènes. Ce mécanisme est qualifié de « pseudogénisation ». L'ensemble de ces mécanismes évolutifs, duplications de nombreuses copies de gènes suivies de la pseudogénisation de plusieurs d'entre elles est nommé '*birth and death*' (Michelmore and Meyers, 1998; Nei and Rooney, 2005).

Les copies de gènes issues de duplications, dont certaines sont en cours de pseudogénéisation, présentent une homologie dont le niveau dépend de l'âge de la duplication et du nombre de mutations qui ont pu s'accumuler entre les copies. Lorsque ces duplications sont très récentes et que la dynamique de '*birth and death*' a un rythme élevé, l'homologie entre les copies est forte. Ce cadre particulier peut alors conduire à des erreurs d'identification et d'annotation des gènes d'intérêt et donc impacter l'interprétation de ces données.

C.4.2. Diversification des gènes de résistance

La diversification des récepteurs de résistance permet l'émergence de nouvelles capacités d'identification de ligands. Ce processus est également impliqué dans les dynamiques de coévolution entre récepteurs et ligands pour maintenir la spécificité de reconnaissance au cours du temps (McDowell and Simon, 2006). Les deux mécanismes majoritaires participant à la diversification des gènes sont les mutations ponctuelles et les recombinaisons (notamment illégitimes) (Leister, 2004).

Les mutations ponctuelles

Parmi les mutations ponctuelles, on peut identifier les indels et les substitutions non-synonymes. Ces évènements vont provoquer des changements dans les séquences protéiques correspondantes. Les indels de nucléotides, s'ils sont en multiple de trois, vont éliminer ou ajouter des acides aminés, tandis que les substitutions non-synonymes vont induire un

changement d'acide aminé dans la séquence. Si la plupart des mutations qui affectent la séquence protéique sont éliminées au cours du temps car elles ont un effet négatif sur la *fitness* des individus qui la portent (i.e. sélection purificatrice ou purifiante), certaines d'entre elles peuvent être à l'origine de l'émergence de nouvelles fonctions. Ces mutations favorables vont pouvoir se fixer au cours du temps si elles apportent un avantage sélectif, c'est la sélection positive ou diversifiante. L'émergence de nouvelles fonctions, appelé « néo-fonctionnalisation », est rare mais représente un moteur d'innovation et d'adaptation pour les organismes.

Dans le cas des récepteurs à LRR, plusieurs études ont mis en évidence l'effet de la sélection positive, chez les NLR et les LRR-RLK (Mondragon-Palomino et al., 2002; Wang et al., 2011; Fischer et al., 2016). Ce type de sélection peut être détecté notamment lorsque le taux de substitutions non-synonymes est plus élevé qu'attendu. Ces signatures moléculaires de sélection ont pu être détectées en particulier dans le domaine LRR.

Les recombinaisons

Les mécanismes de recombinaisons permettent l'échange de fragment de séquence entre différents allèles (recombinaison homologue), ou entre différents gènes (recombinaison non homologue ou illégitime, conversion génique). Ces évènements ont lieu entre régions présentant des homologies de séquence élevées, permettant l'appariement des molécules d'ADN. Les récepteurs LRR étant fortement dupliqués et organisés en larges clusters, les évènements de recombinaison apparaissent donc avec une fréquence plus élevée que pour d'autre familles de gènes. En particulier, il a été démontré chez *A. thaliana* qu'au moins 45% des NLR, 34% des LRR-RLK et 69% des LRR-RLP, avaient subi des évènements de conversion génique (Mondragon-Palomino and Gaut, 2005).

Un autre moyen de diversifier les gènes et leur fonction consiste à créer de nouvelles structures en fusionnant des domaines préexistants (Bjorklund et al., 2005). De nouvelles architectures de protéines peuvent apparaitre via des mécanismes de recombinaison non-homologue. Ils sont à l'origine de la variabilité observée pour les récepteurs à LRR comme par exemple les TNL, RNL et CNL pour les NLR. Ils présentent chacun un domaine en N-ter variable issus d'évènements de fusion indépendants mais anciens (Tamborski and Krasileva, 2020). Ces mécanismes sont également à l'origine des NLR présentant des domaines intégrés (*integrated domains*) qui constituent une source d'innovation importante pour les résistances (cf. section

C.2.2.) (Bailey et al., 2018). Une étude récente s'est intéressée à l'impact des variations structurales sur la diversification des LRR-RLK (Man et al., 2020). Les auteurs montrent que la diversification des structures peut impacter : (i) la détection des gènes pour la famille du fait de l'absence de certains domaines caractéristiques, et (ii) les analyses par phylogénie du fait des échanges de séquences entre différents gènes. Ils observent 2,8 % des structures LRR-RLK avec un domaine additionnel non-commun. En revanche, ces fusions ne sont jamais retrouvées dans les orthologues de ces gènes chez les espèces incluses dans l'étude. Ces évènements seraient alors récents et pourraient indiquer que les fusions de domaines sont fréquentes mais rarement conservées. Pour les LRR-RLK, seuls deux évènements de fusions différents semblent avoir été conservés au cours de l'évolution des plantes, et concernent les structures avec un domaine malectin et les structures avec un domaine malectin-like (Figure 6) (Dievart et al., 2020).

Contraste des différents mécanismes pour les récepteurs LRR

Au sein de chaque famille de récepteurs LRR, les gènes n'évoluent pas tous au même rythme et ne sont pas soumis de la même façon aux mutations ponctuelles et aux recombinaisons. Par exemple, les gènes NLR peuvent être distingués selon le mécanisme qui les affecte majoritairement (Kuang et al., 2004; Friedman and Baker, 2007). Les NLR dit de type I sont décrits comme des gènes à évolution rapide. Ces gènes subissent fréquemment des échanges de séquence entre différentes copies de même clade (i.e. des paralogues) par recombinaisons homologues et non-homologues (Baumgarten et al., 2003). Les échanges entrainent la formation de copies 'chimériques' dont chaque partie provient d'une copie d'origine différente. Ces échanges de séquences entre paralogues ont pour effet d'homogénéiser les séquences dans les clusters (appelée 'évolution concertée' (Hickey et al., 1991)) et vont donc impacter la détermination des relations d'orthologie entre espèces (Leister, 2004). A l'inverse, les NLR dit de type II sont décrits comme des gènes à évolution lente. Ces gènes évoluent majoritairement via des mutations ponctuelles et présentent un faible taux de recombinaison, donc peu d'échange de séquence entre différentes copies. Pour ces gènes, les relations d'orthologie sont plus facile à déterminer car les copies orthologues entre génomes ont une homologie plus élevée que les copies paralogues au sein d'un même génome (Friedman and Baker, 2007).

Des tels contrastes ont aussi été observés au sein des LRR-RLK. La plupart de ces récepteurs suivent un modèle *birth and death* avec un rythme parfois très intense et forment de large cluster de gènes dupliqués. En revanche, certains de ces gènes sont retrouvés isolés dans les génomes (24% des LRR-RLK chez *A. thaliana*). Les orthologues de ces gènes sont plus facilement identifiables car ils sont beaucoup plus conservés entre génomes éloignés (Mondragon-Palomino and Gaut, 2005; Fischer et al., 2016).

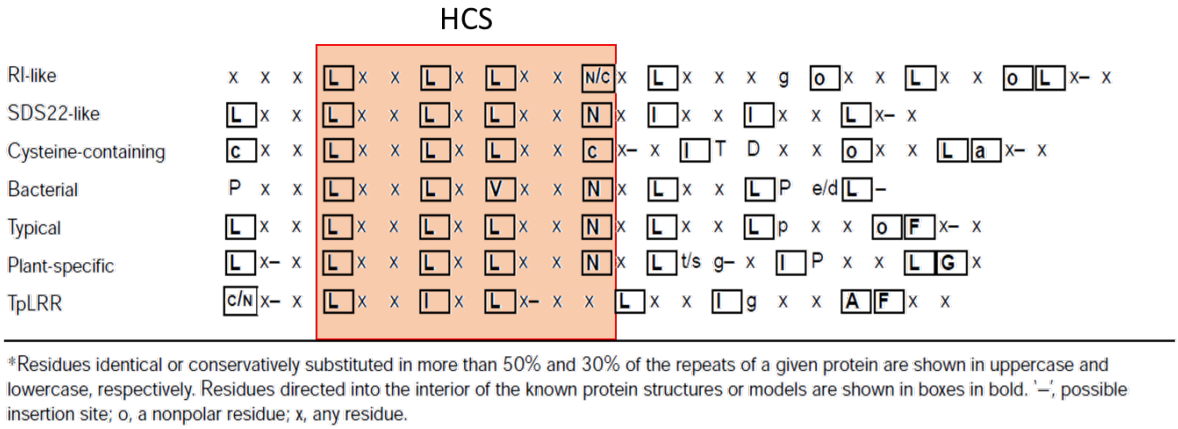
D. Structure et fonction du domaine LRR

Les séquences répétées sont très nombreuses dans les séquences nucléotidiques et protéiques, de tous les organismes. Elles peuvent être réparties en différentes catégories en fonction de leur structure et de leur organisation : (i) les séquences répétées en tandem, dont les répétions sont physiquement proches (micro- et mini-satellites au niveau de zones intergéniques, domaines et motifs répétés dans les gènes) ; (ii) les séquences répétées dispersées (éléments transposables, gènes dupliqués) (Biscotti et al., 2015). Les séquences répétées ont des fonctions biologiques diverses. Elles participent, par exemple, à la stabilité des ARNm via la queue polyA ou à l'organisation structurale de l'ADN via leur présence au niveau des centromères et des télomères. Au niveau protéique, les motifs répétés sont souvent impliqués dans les mécanismes de reconnaissance ou de régulation via des interactions protéine-protéine ou protéine-ligand (Schaper and Anisimova, 2015). C'est le cas des répétitions riches en leucine, ou '*Leucine-Rich Repeats*' (LRR) qui comptent parmi les motifs les plus abondants des protéomes de plantes avec les répétitions PPR (pentatricopeptide repeats) (Bjorklund et al., 2006; Schaper and Anisimova, 2015).

D.1. Différents types de motifs LRR

Le domaine LRR est un domaine protéique retrouvé dans l'ensemble des règnes vivants (virus, bactérie, animaux, plantes). C'est un domaine composé de motifs répétés et riches en résidus aliphatiques, et plus particulièrement en Leucine. La première description d'un tel motif date de 1985 et concerne une glycoprotéine humaine : la *leucine-rich α2-glycoprotein* (Takahashi et al., 1985). Avant la fin des années 90, sept types de motifs LRR différents sont décrits (Figure 11) (Kajava, 1998; Kobe and Kajava, 2001): « *Typical* » (Kajava et al., 1995), « *RI-like* » (Hofsteenge et al., 1988), « *Cys-containing* » (Flick and Johnston, 1991), « *Plant-specific* » (Jones and Jones, 1997), « *SD22-like* » (Ohkura and Yanagida, 1991), « *Bacterial* » (Buchanan

and Gay, 1996) et « *TpLRR* » (Shevchenko et al., 1997). Un 8ème type de motif a été décrit récemment : « *IRREKO* » (Matsushima et al., 2010). Quel que soit le type de motif, une répétition LRR est composée de deux segments : un segment hautement conservé (HCS, *Highly Conserve Segment*) et un segment variable (VS, *Variable Segment*). Le segment HCS est le fragment de séquence portant les résidus Leucine, cette partie est globalement commune à tous les types de motifs LRR. Le segment VS permet lui de distinguer les différents types de motifs (Figure 11).



*Residues identical or conservatively substituted in more than 50% and 30% of the repeats of a given protein are shown in uppercase and lowercase, respectively. Residues directed into the interior of the known protein structures or models are shown in boxes in bold. '–', possible insertion site; o, a nonpolar residue; x, any residue.

Figure 11 : Consensus des différents types de motifs LRR décrits dans la littérature.
Adaptée de Kobe et Kajava (2001)

D.2. Motif LRR chez les plantes

Le motif LRR « *Plant-specific* » est le motif le plus retrouvé dans les protéines de plantes. Il compose notamment les domaines LRR des récepteurs LRR-RLK et LRR-RLP. Il a un consensus de 24 acides aminés représenté par la séquence : LxxLxxLxLxxNxLxGxIPxxLxx. D'autres motifs sont également décrits chez les plantes, en particulier pour des gènes intracellulaires NLR. Ils ressemblent en partie au motif de type « *Cys-containing* » avec un résidu cystéine (C) majoritaire en position 12 à la place de l'asparagine (N). Mais ils présentest également le di-peptide « IP » dans le segment variable des motifs comme pour le motif de type « *Plant-specific* » (Matsushima and Kretsinger, 2016; Martin et al., 2020).

D.3. Structure secondaire et tertiaire des domaines LRR

La première structure cristallisée d'un domaine LRR a été faite pour la protéine *Ribonuclease inibitor* (RI) porcine (Kobe and Deisenhofer, 1993). Chez les plantes, il faut attendre 2011 pour

avoir la première structure cristallisée d'un domaine LRR, qui concerne le récepteur hormonal *BRI1* d'*Arabidopsis thaliana* (Hothorn et al., 2011; She et al., 2011).

Du point de vue du motif LRR, le segment HCS forme un feuillet β et le segment VS va former des structures de type coudes et/ou hélices. Ces structures vont être déterminées principalement par la longueur du motif et la composition de leur segment VS (Figure 12a et b) (Kobe and Deisenhofer, 1994; Kajava, 1998; Bella et al., 2008).

Au sein du domaine, les motifs LRR forment un solénoïde : une structure hélicoïdale dont chaque répétition LRR forme un tour de l'hélice (Figure 12c). Généralement, le solénoïde est courbé, plus ou moins en fer à cheval en fonction du nombre de répétitions LRR. Sur la partie concave, on retrouve les feuillets β parallèles qui se stabilisent entre eux par des liaisons hydrogènes portées par les résidus leucine (ou autres aliphatiques). Sur la partie convexe, on retrouve les segments VS des motifs LRR. L'encombrement généré en fonction de la structure secondaire prise par la partie VS (coude, hélice alpha, hélice 3_{10}) va déterminer le degré de courbure du solénoïde ; plus la structure est encombrante (hélice alpha > hélice 3_{10} > coude) plus le solénoïde sera courbé (Bella et al., 2008).

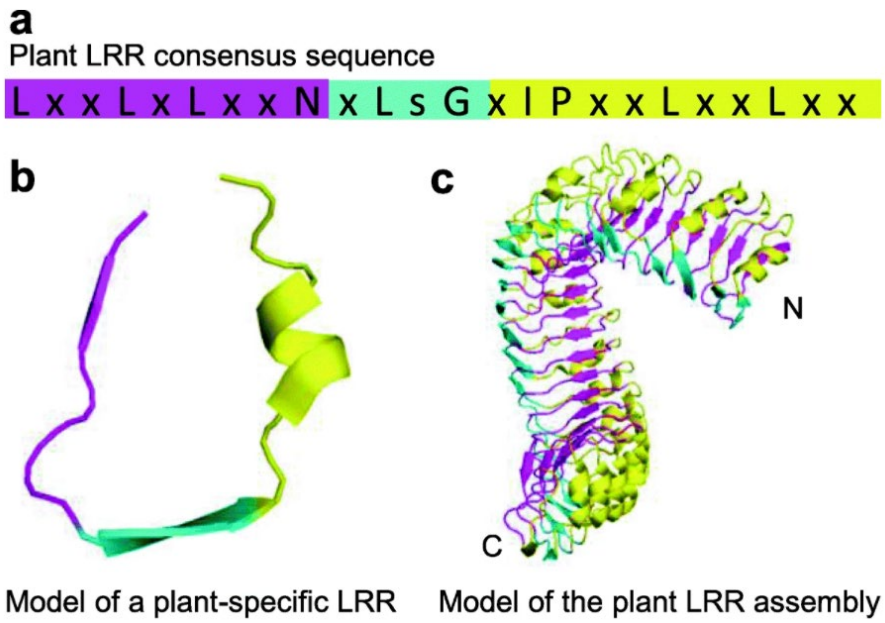


Figure 12 : Modélisation de la structure des motifs LRR.
D'après (Chen, 2021).

D.4. Origine des motifs LRR et évolution du domaine LRR

L'origine des différents types de motifs est débattue depuis le début des années 90. Deux hypothèses s'opposent. La première suppose que les motifs LRR ont une origine commune, et les différents types de motifs ont émergé par duplication et diversification à partir d'un motif ancestral. La seconde évoque plusieurs origines indépendantes pour au moins une partie des différents types de motifs. En faveur de cette hypothèse, Kobe et Deisenhofer (Kobe and Deisenhofer, 1994) évoquent la grande diversité des motifs (en consensus et longueur) ainsi que la spécificité d'occurrence de chacun en fonction des familles de protéines. Cela tend à suggérer une origine complexe des motifs, qui, d'après les auteurs, pourraient avoir émergé *de novo* à différents moments au cours de l'évolution, et dans différents phylums. Cette hypothèse est également défendue par Kajava (Kajava, 1998) qui s'appuie sur la spécificité d'occurrence de chaque motif dans les protéines. Le fait de ne jamais trouver des motifs de différents types au sein d'un même domaine serait la conséquence de contraintes structurales fortes appliquées au segment VS des motifs. Cette observation est interprétée par l'auteur comme le signe que les différents motifs ne seraient pas issus d'un motif ancestral unique. Dans ce cas, les ressemblances observées seraient la conséquence d'une convergence évolutive (Patthy, 2003).

En revanche, il est également évoqué que la distinction entre les différents types de motifs n'est pas absolue et beaucoup de motifs peuvent correspondre à plusieurs de ces types (Andrade et al., 2000). C'est également l'hypothèse privilégiée par Matsushima et Kretsinger (Matsushima and Kretsinger, 2016) qui observent des occurrences de différents types de motifs dans certaines protéines formant des superstructures (Matsushima and Kamiya, 1999; Matsushima et al., 2010). Par ailleurs, certaines caractéristiques partagées par différents gènes portant différents types de motifs tendent également à suggérer une origine commune. Jones et Jones (Jones and Jones, 1997) évoquent le cas des protéines dont chaque motif LRR est porté par un exon indépendant et dont la séparation intron-exon est systématiquement à la même position du motif. Cette caractéristique, spécifique de certains récepteurs membranaires, est partagée par les plantes et les animaux. Pour ces auteurs, ces résultats suggèrent une origine commune de ces motifs-là au moins, avec un motif LRR entouré d'introns préexistants avant la séparation de ces règnes vivants.

Bien que l'origine réelle des différents motifs LRR reste non élucidée pour le moment, il a été montré que le répertoire de ces motifs (comme tout motif répété) tend à s'étendre au cours de l'évolution (Bjorklund et al., 2006; Anisimova et al., 2015). Si la duplication fréquente des gènes complets, liée au modèle évolutif '*birth and death*', participe fortement à cette expansion, il existe d'autres mécanismes qui agissent à l'échelle du domaine ou des motifs pour étendre et diversifier le répertoire des LRR d'un organisme, d'une espèce ou d'un groupe phylogénétique. Les domaines LRR sont notamment affectés par des évènements de conversion génique ou des mutations ponctuelles comme évoqué précédemment (cf. section C.4.2.), mais ils sont aussi affectés par des évènements qui vont modifier le nombre de motifs LRR dans le domaine. Par exemple, on peut observer des indels de motifs complets et des duplications de motifs intra-domaine (Noel et al., 1999). Ces évènements participent ensemble à la diversité locale (motifs intra-gènes/allèles) et globale (motifs intra-génome/espèces) du répertoire LRR.

L'importance et la fréquence de ces différents mécanismes ne sont pas encore bien connues car peu d'études se sont intéressées de manière globale à ces évènements. En particulier, on ne connait pas encore l'impact relatif de chacun des mécanismes (mutations vs duplication et conversion génique) pour les différentes familles, ni le rythme auxquels ils ont lieu à l'échelle du génome. Certains de ces mécanismes sont, en plus, difficiles à distinguer. Par exemple, les conversions géniques et les duplications récentes vont avoir pour conséquence de réduire la diversité observable des motifs. Par ailleurs, le fait de définir un motif par sa séquence, et donc par des bornes précises, peut limiter la puissance des méthodes de détection d'évènements de duplications. En effet, une duplication peut entrainer la fin d'un motif et le début du suivant. Cet évènement forme alors un motif dont la séquence est une chimère entre deux motifs préexistants (Figure 13). Le motif chimérique, défini par des bornes fixes, n'aura alors pas un niveau d'identité plus élevé que l'attendu moyen quand il est comparé aux motifs voisins (Szalkowski and Anisimova, 2013; Anisimova et al., 2015). Björklund *et al.* (2006) utilisent un jeu de données de motifs répétés définis selon des bornes de motif fixes. Ils observent que les motifs LRR se dupliquent majoritairement par groupes de 2 motifs adjacents à la fois. Mais il est possible qu'ils détectent moins de duplications de seulement un motif, à cause de ce choix.

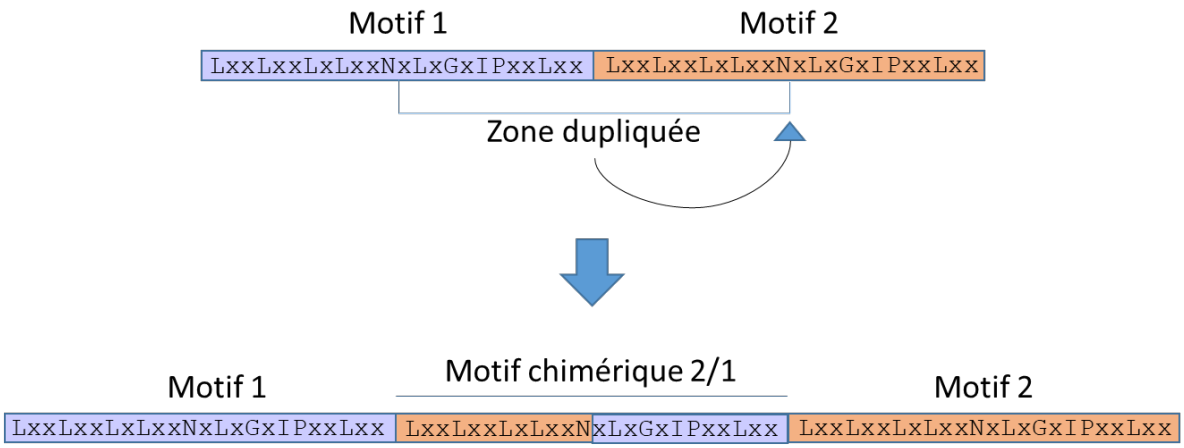


Figure 13 : Conséquence d'une duplication de motif LRR non centrée sur le consensus défini.
La duplication entraîne la formation d'un motif chimérique présentant une similarité de séquence avec le motif 2 sur sa partie N-terminale, et avec le motif 1 sur sa partie C-terminale.

E. Méthodes d'annotation des gènes et des génomes

L'annotation dans le domaine de la bio-informatique consiste à donner aux séquences biologiques brutes une information compréhensible et exploitable. Il existe deux grands types d'annotation : l'annotation structurale et l'annotation fonctionnelle.

E.1. Annotation Structurale : donner du sens à la séquence génomique

L'annotation structurale des génomes consiste à identifier les régions portant des éléments fonctionnels d'intérêts afin de donner du sens à la séquence génomique et ainsi pouvoir l'étudier. On localise les gènes, les promoteurs, les séquences régulatrices, les TE, les ARN non codants, etc. Pour les gènes, l'annotation consiste à identifier et labéliser l'ensemble de ces caractéristiques structurales : les introns, les exons, les UTRs et à repérer la partie codante pour une protéine (CDS).

On distingue généralement deux types de méthodes pour l'annotation structurale des séquences génomiques : les méthodes comparatives et les méthodes dites *ab initio* (Yandell and Ence, 2012).

E.1.1. Méthodes comparatives

Les méthodes comparatives se basent sur des comparaisons de séquences, i.e. des alignements, avec des données de différentes natures pour générer une annotation (König et al., 2018). Elles peuvent être séparées en deux approches : les méthodes basées sur

l'homologie globale, qui consistent à aligner le génome d'intérêt sur des séquences génomiques déjà annotées pour guider la prédiction, et les méthodes basées sur la similarité, qui utilisent des données de type séquences protéiques, cDNA, EST, RNAseq, etc. pour identifier les régions correspondant aux structures à annoter (Figure 14a).

E.1.2. Méthodes ab initio et méthodes hybrides

Les méthodes *ab initio* s'appuient sur des modèles de vraisemblance appris sur un ensemble de données de nature variée. Les modèles sont ensuite appliqués sur la séquence à annoter pour identifier les éléments d'intérêt (Figure 14b). Les modèles combinent par exemple des ensembles de séquences (protéiques, génique, EST, cDNA), des signaux caractéristiques (site d'épissage, codons start et codon stop, promoteurs), ou des informations statistiques diverses (biais d'utilisation du code génétique, pourcentage de GC entre gènes et régions intergéniques). Les modèles qui en découlent sont variés : HMM (*hidden markov model*), GHMM (*generalized HMM*), WAM (*weight array model*), etc. Les outils d'annotations les plus connus sont Augustus (Stanke and Waack, 2003), Gnomon (Souvorov et al., 2010) et FGENESH (Salamov and Solovyev, 2000).

Les méthodes hybrides sont dérivées des méthodes *ab initio* (Yandell and Ence, 2012). Elles reposent essentiellement sur le même principe mais corrigent les prédictions de gènes grâce à des alignements avec des données connues (protéines ou transcrits). Parmi les outils proposant ces stratégies, on trouve : GenomeScan (Yeh et al., 2001), FGENESH+ (Solovyev et al., 2006) et des variantes de EuGene (Sallet et al., 2019), Augustus (Stanke and Waack, 2003) ou Gnomon (Souvorov et al., 2010).

E.1.3. Biais des méthodes d'annotation structurale

La plupart des outils modernes combinent les approches précédentes (comparatives et *ab initio*), et appliquent des règles de décision à travers des « *choosers* » (ou « *combiners* ») pour résoudre les conflits de prédiction et fournir une annotation unifiée pour le génome d'intérêt (Yandell and Ence, 2012).

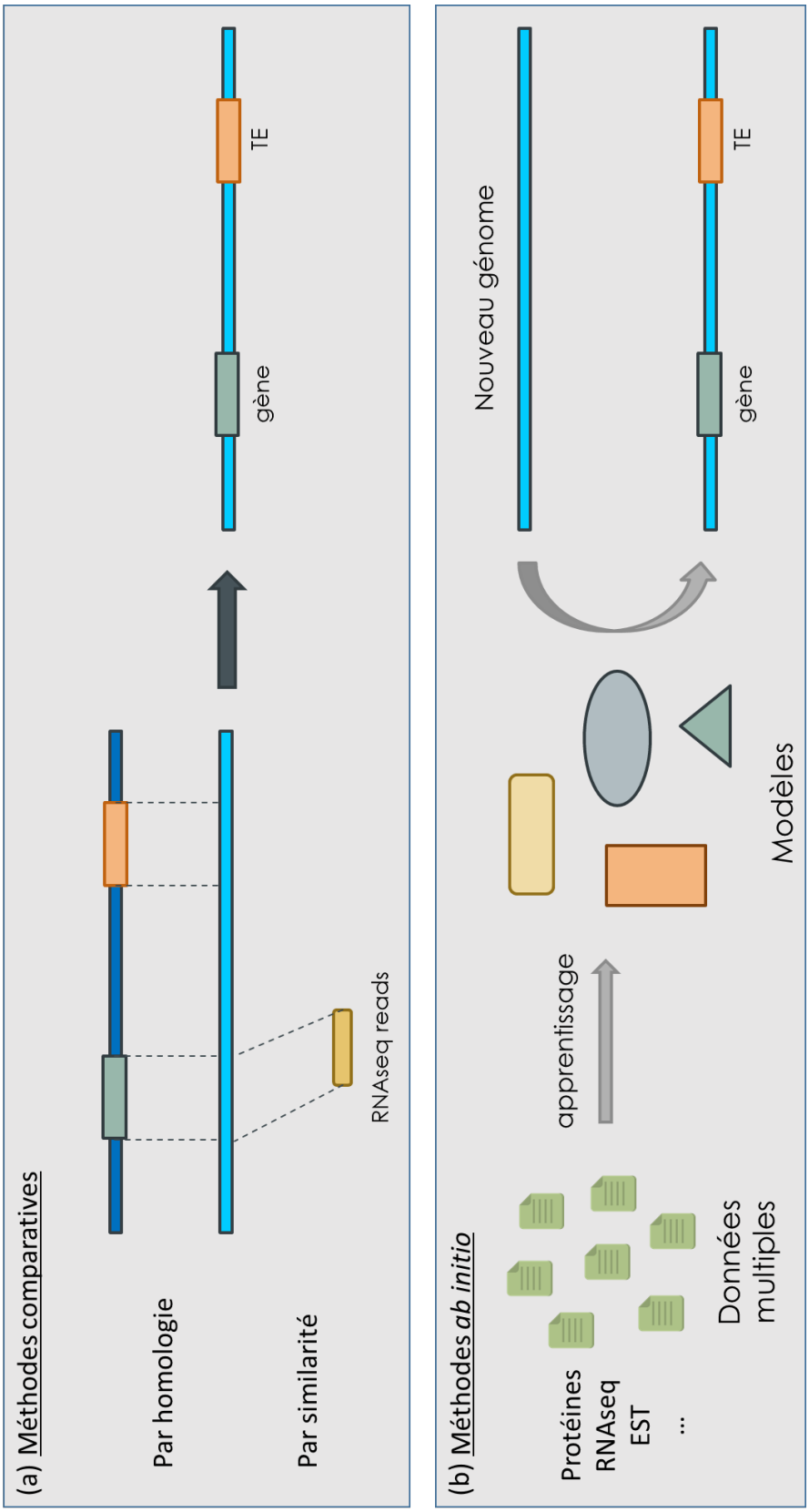


Figure 14 : Deux stratégies d'annotation structurale des génomes.
(a) annotation des génomes par une approche comparative. (b) annotation des génomes via des méthodes *ab initio.*

Cependant, des biais existent dans ces méthodes. (i) Les premiers grands efforts d'annotation ont été réalisés chez les bactéries, les levures puis chez les mammifères, notamment la souris et l'humain. Pendant longtemps, ces méthodes ont simplement été transposées à d'autres espèces et d'autres branches du vivant en exportant des règles qui ne sont pas toutes pertinentes suivant la cible que l'on souhaite annoter. (ii) Les méthodes comparatives utilisent quelques références très connues (humain, souris, drosophile, *Arabidopsis*), et peuvent répandre les erreurs de la référence aux autres génomes. Si ces références, qui sont au centre de nombreux efforts de recherche, sont régulièrement corrigées, les annotations des autres génomes ne le sont pas forcément. (iii) Les méthodes sont validées et les annotations contrôlées via des gènes très connus (souvent des gènes de ménage, appelés « *housekeeping genes* »), mais ces gènes sont généralement les plus simples à annoter car ils sont très conservés entre les espèces et donc très présents dans les bases de données, avec des structures de très bonne qualité. Le contrôle de l'annotation avec des outils comme BUSCO (Simao et al., 2015) ne donne aucune indication sur la qualité de l'annotation des autres gènes que ceux directement intégrés à BUSCO. (iv) Ces méthodes essayent de prédire tous les gènes sur la même base d'information. Or chaque séquence est unique et un modèle capable de générer des prédictions qualitatives pour une famille de gènes, peut être médiocre pour une autre.

Un autre point très important est que ces méthodes ne sont pas conçues pour prédire des gènes impactés par des mutations non-sens, principalement provoquées par des substitutions créant des codons stop prématurés ou des indels provoquant un décalage de cadre de lecture (*frameshif*). La transcription des gènes impactés par ces mutations peut éventuellement aboutir à la synthèse d'un ARN complet mais dont la protéine synthétisée à partir de cet ARN sera tronquée. Cette particularité est difficile à prendre en compte par les outils d'annotation automatique et la qualité des annotations fournies s'en trouve réduite. En effet, certaines annotations peuvent tronquer prématurément le modèle de gène pour correspondre à la protéine partielle qui serait produite (la fin du gène est donc manquée par l'annotation) ou la protéine est annotée complète avec de faux introns introduits pour esquiver les mutations non-sens (cf. chapitre 2).

E.1.4. Impact des redondances pour l'annotation de familles de gènes complexes

La présence de fortes homologies de séquences entre différentes régions du génome qu'on trouve dans le contexte des gènes LRR (motifs répétés, gènes dupliqués), impacte les méthodologies appliquées aux données à différentes échelles. Tout d'abord, elles impactent l'assemblage des séquences de références (génomes et transcrits) à partir des *reads* de séquençage (WGS ou RNAseq) en créant des ambiguïtés de position pouvant conduire à des erreurs de construction. Ensuite, elles impactent l'annotation structurale des séquences génomiques, c'est-à-dire la prédiction des gènes et la détermination de leurs structures introns-exons. Pour pallier ce problème, il est fréquent de passer par une étape de masquage des séquences répétées (« *repeat masking* ») avant de procéder à l'annotation. Mais cette étape constitue un biais supplémentaire car il est fréquent que les gènes de résistances (souvent présents en clusters) sont en partie masqués par cette étape (Bayer et al., 2018).

Les familles de gènes complexes, en particulier multigéniques et dupliquées, sont particulièrement impactées par ces biais (Fawal et al., 2014). Ce sont des familles souvent mal caractérisées dont les séquences sont de mauvaise qualité dans les bases de données ce qui participe à la propagation de mauvaises annotations pour les gènes de ces familles.

Plusieurs études ont noté la présence de modèles de gènes erronés dans les données d'annotation publiques pour les récepteurs LRR, et plus particulièrement pour les NLR (Meyers et al., 2003; Jupe et al., 2012; Jupe et al., 2013; Andolfo et al., 2014; Man et al., 2020). Certains ont fait le choix de ré-annoter manuellement les gènes concernés avant d'analyser les données. Parmi les erreurs récurrentes relevées, on observe l'annotation de gènes tronqués ou de pseudogènes (séquences avec des mutations non-sens) annotés comme des gènes complets fonctionnels (Meyers et al., 2003). En effet, les outils d'annotation automatiques forcent les modèles pour être canoniques, c'est-à-dire qui respectent les attendus biologiques d'un gène fonctionnel (codon *start*, sites de *splicing*, taille d'intron, codon *stop* terminal etc.). Si un codon stop ou un *frameshift* survient dans une copie de gène, les outils peuvent tronquer le modèle ou intégrer de faux introns pour obtenir un modèle canonique. Ces modèles incomplets ou erronés participent à une mauvaise caractérisation des familles complexes et au maintien d'annotations chimériques dans les génomes et les bases de données.

E.2. Annotation des séquences protéiques

L'annotation des séquences protéiques consiste à identifier les domaines et motifs protéiques présents dans la séquence. L'identification de ces motifs ou domaines appartient au champ de l'annotation fonctionnelle. Différentes stratégies existent pour annoter des séquences protéiques (Wang et al., 2021). Ces méthodes font la balance entre la précision, l'exhaustivité et l'utilisation des ressources computationnelles (mémoire et temps de calcul).

E.2.1. Méthodes comparatives (alignements)

La stratégie la plus simple consiste à rechercher des similarités de séquences entre l'élément à annoter (la requête) et un ensemble de données déjà annotées (les cibles). Fréquemment utilisées, ces méthodes « comparatives » reposant sur des alignements entre séquences sont très simples à mettre en œuvre et permettent d'annoter de grandes quantités de données (Wang et al., 2021). En revanche, l'efficacité de ces méthodes peut être impactée par différents aspects : (i) Le temps d'exécution est dépendant de la taille de la base de données cible et peut rapidement devenir très long pour les grandes bases de données. (ii) Il existe des biais de représentation de certaines familles de gènes ou certains domaines protéiques dans les bases de données. Les domaines les mieux annotés sont aussi les plus représentés, et les domaines les plus représentés sont les plus faciles à annoter. (iii) L'homologie entre la séquence requête et le jeu de données cible peut être variable et la conservation de séquence au sein d'une même famille peut être faible et limiter les possibilités d'identification. Ces stratégies basées sur les alignements peuvent alors générer des erreurs d'annotation correspondant à : (i) des faux négatifs (i.e. ne pas identifier un domaine alors qu'il est présent) lorsque la séquence à annoter est trop divergente ou lorsque la base de données possède peu ou pas d'homologues, et (ii) des faux positifs (i.e. identifier un domaine qui est absent) lorsque des ressemblances aléatoires entre des séquences n'étant pas liées par une même fonction sont identifiées.

E.2.2. Profils HMM pour l'annotation des protéines

Une fonction est généralement portée par un ou plusieurs domaines fonctionnels. L'avantage sélectif conféré par la fonction crée une pression de sélection purificatrice sur les domaines fonctionnels. Leur séquence primaire est donc conservée entre les différents gènes qui les portent au cours de l'évolution. Ainsi, la détection de domaines fonctionnels conservés

constitue une approche plus sensible pour l'annotation fonctionnelle que la simple comparaison de séquence (Bolger et al., 2018). Dans cette optique, l'utilisation des profils HMM (*Hidden Markov Model*) est très répandue. C'est une méthode simple, peu gourmande en ressources et permettant de discriminer efficacement la présence d'un motif ou d'un domaine particulier dans les séquences biologiques, qu'elles soient protéiques ou nucléotidiques. Un profil HMM est un modèle statistique construit à partir d'un alignement multiple permettant de retranscrire la diversité observée à chaque résidu du domaine considéré. Le profil HMM représente des probabilités d'émission associées à des états et des probabilités de transitions entre ces états. Dans le cadre des séquences biologiques, et plus particulièrement des motifs protéiques, un état d'un HMM correspond à une position dans ce motif. Les probabilités d'émission associées à cet état détermineront les fréquences d'observation de chaque résidu acide aminé à cette position. Les probabilités associées aux transitions détermineront le prochain état. A partir d'un état donné, les transitions possibles sont le passage à l'état suivant ('*match*'), l'insertion (rester à l'état actuel) ou la délétion (aller à un état postérieur à l'état suivant). Lorsqu'un profil HMM est appliqué sur une séquence ou un fragment de séquence, il permet d'extraire une probabilité, sous forme d'un score, que cette séquence en question puisse être produite par ce profil. Si ce score surpasse certains seuils prédéfinis, alors on considère que le motif testé est présent dans la séquence ('*hit*'). Les outils permettant de comparer des profils à des séquences biologiques (HMMER, MEME) fournissent une *e*-value associée à chaque comparaison, permettant de savoir si le *hit* aurait pu être obtenu par simple hasard (Eddy, 1996, 1998).

L'outils d'identification fourni par la suite HMMER offre une *e*-value supplémentaire, dépendante du nombre de *matchs* dans la séquence. Ainsi, un *hit* de *e*-value supérieur au seuil peut être quand même considéré comme significatif si le même profil a eu d'autres *hits* dans la même séquence, ce qui a pour effet de diminuer la probabilité que ce *hit* divergent soit dû au hasard. Cette fonction est particulièrement utile pour l'annotation des motifs répétés (Anisimova et al., 2015).

E.2.3. Annotation des motifs répétés dans les séquences protéiques

La détection et l'annotation de motifs répétés au sein de protéines est difficile et se base largement sur la conservation entre les répétitions. Or, les domaines de motifs répétés forment des structures 3D conservées (Kajava, 2012). Dans le cas des LRR, cette structure est

une hélice dont chaque tour est composé d'un motif (cf section D.3.). Les répétitions au sein d'un même domaine sont généralement imparfaites car les séquences accumulent des mutations, substitutions et *indels*, au cours de l'évolution qui ne sont pas nécessairement contre sélectionnées. La variabilité de ces motifs semble participer à la stabilité des structures 3D (Jorda et al., 2010) mais impacte négativement les capacités de détection des motifs par les outils d'annotation.

L'annotation des domaines protéiques contenant des répétitions fait face à trois challenges. Premièrement, il est important d'identifier l'ensemble des répétitions d'une séquence, permettant de fournir une annotation exhaustive du domaine. Ensuite, il est nécessaire de ne pas identifier de faux motifs (faux positifs) qui conduiraient à une annotation erronée du domaine. Enfin, il est important d'identifier de manière constante les bornes de ces motifs afin de pouvoir les comparer entre eux.

Il existe deux stratégies principales pour l'identification et l'annotation des motifs répétés. La première, sans a priori, consiste à faire une recherche de cyclicité en alignant la séquence étudiée avec elle-même à l'aide d'outils dédiés tels que ceux implémentés dans MEME (Bailey et al., 2006). La seconde est basée sur une recherche de motifs prédéfinis, généralement modélisés sous forme de profils HMM ou de PSSM ('*Position Scoring Matrix*'). L'utilisation de profils HMM pour annoter les motifs est une méthode simple et fréquemment utilisée dans le cas des motifs LRR (Mondragon-Palomino et al., 2002; Wang et al., 2008; Li et al., 2010; Dievart et al., 2011; Guo et al., 2011; Sun and Wang, 2011; Wang et al., 2011; Schulze et al., 2015; Fischer et al., 2016; Magalhaes et al., 2016; Dufayard et al., 2017; Liu et al., 2017; Sun et al., 2017; Die et al., 2018; Sun et al., 2018; Van de Weyer et al., 2019; Man et al., 2020).

Différentes stratégies ont été développées pour l'identification des motifs LRR avec les HMM tenant compte de leurs caractéristiques intrinsèques :

(i) Afin d'identifier indépendamment les différents types de motifs connus, on peut utiliser un HMM ne représentant que la partie HCS du motif car cette partie est commune aux différentes familles de LRR (cf. tableau 1). En revanche, cette partie étant courte, et majoritairement composée de résidus aliphatiques (fréquents dans la composition des protéines) cette stratégie tend à générer de nombreux faux positifs ;

(ii) A l'inverse, utiliser des profils représentant un motif complet permet généralement de s'affranchir des faux positifs. En revanche, la partie variable du motif (VS) étant assez différente entre les répétitions des différentes familles de motifs LRR, chaque profil tend à n'identifier que les motifs de sa famille. Avec cette stratégie, l'utilisation d'un profil générique à l'échelle des plantes tend alors à générer de nombreux faux négatifs ;

(iii) Enfin, il est possible d'utiliser des profils multi-motifs, c'est-à-dire représentant plusieurs répétitions consécutives (2 ou 3 en général). Souvent, les parties conservées (HCS) des motifs sont bien caractérisées et les parties variables (VS) sont faiblement caractérisées. Ainsi, le profil offre généralement une bonne sensibilité pour l'ensemble des familles LRR, mais limite les faux positifs en s'appuyant sur l'occurrence multiple du segment conservé avec une distance définie entre chacune des occurrences. En revanche, les résultats des recherches effectuées sur la base de ce type de profil ne permettent pas de déduire directement une annotation motifs par motifs avec des bornes standardisées.

Une façon de contourner les limitations évoquées ci-dessus serait d'utiliser un profil représentant un motif complet, pour éviter les faux positifs, mais qui s'adapte au jeu de donnée et aux familles de gènes, pour éviter les faux négatifs.

F. Problématiques et objectifs de la thèse

L'étude de l'histoire évolutive des récepteurs à LRR est une étape indispensable à une meilleure compréhension de l'immunité chez les plantes, en particulier des mécanismes évolutifs à l'origine de l'émergence de nouveaux allèles d'intérêt dans les espèces sauvages et cultivées.

Dans la littérature, les analyses évolutives de ces gènes chez les plantes sont conduites majoritairement sur les LRR-RLK ou NLR car elles sont basées sur l'analyse des domaines fonctionnels associés, i.e. le domaine kinase pour les LRR-RLK et le domaine NB-ARC pour les NLR. Se focaliser sur les domaines fonctionnels associés facilite les analyses car ces domaines sont plus conservés et donc plus facile à identifier et à aligner, mais ne permet pas, par exemple, d'analyser la famille des LRR-RLP. C'est pour cette raison que les analyses évolutives sur les LRR-RLP sont nettement moins nombreuses que celles sur les LRR-RLK et les NLR. Par ailleurs, les analyses réalisées sur ces familles sont généralement indépendantes les unes des autres. Elles ne permettent donc pas de comparaisons entre ces trois familles de gènes,

notamment au niveau des domaines LRR. Pourtant, la spécificité de ces 3 familles de récepteurs pour une cible étant principalement portée par le domaine LRR, étudier l'évolution de ces récepteurs et les mécanismes à l'origine de la diversité de leurs domaines LRR, est d'un intérêt majeur pour comprendre leurs fonctionnements et leurs mécanismes d'adaptation chez de nombreuses espèces.

Dans ce champ de recherche, les questions qui se posent actuellement sont de deux types :

(i) est-ce que ces trois familles de récepteurs évoluent de la même façon, et dans quelle mesure les spécificités structurales de leurs domaines LRR jouent-elles un rôle dans cette évolution ?

(ii) comment caractériser et mesurer la vitesse d'évolution des domaines LRR en tenant compte de plusieurs facteurs spécifiques aux séquences protéiques répétées (évolution du nombre de motif, de leur ordre, de la séquence même des motifs etc.).

Pour répondre à ces questions, un prérequis est que les trois familles de récepteurs à LRR, et en particulier leurs domaines LRR et par extension leurs motifs LRR, soient identifiés et annotés de manière homogène, exhaustive et reproductible dans les génomes étudiés.

Ma thèse s'inscrit dans cette thématique scientifique et contient trois objectifs :

- Le premier objectif concerne le développement d'un outil d'annotation efficace et reproductible des motifs LRR dans les séquences protéiques,
- Le second objectif concerne l'annotation des récepteurs LRR (modèle intron-exon) dans les génomes avec une attention particulière pour les structures affectées par des mutations non-sens (appelées « non-canoniques »),
- Le troisième objectif concerne les stratégies et bonnes pratiques pour la diffusion, la reproductibilité, la portabilité et la transparence des outils et données générées pendant la thèse.

Ma thèse se compose de quatre chapitres. Le **premier chapitre** décrit les travaux que nous avons menés sur l'annotation des motifs LRR au sein des séquences protéiques. Dans ce chapitre, nous proposons une stratégie d'annotation des motifs LRR basée sur l'utilisation de profils HMM. Nous utilisons des profils représentant un motif unique et complet. Cela nous permet, d'une part, d'avoir des bornes fixes définies pour tous les motifs identifiés, et d'autre

part, d'éviter les faux positifs inhérents à l'utilisation de profils représentant uniquement la partie conservée des motifs (cf. section E.2.3.). En revanche, nous faisons en sorte que les profils s'adaptent aux motifs LRR de chaque famille de récepteurs dans le génome étudié afin d'éviter les faux négatifs pouvant apparaitre avec l'utilisation de profils génériques. Cette stratégie repose sur un processus itératif permettant d'adapter progressivement un profil LRR aux données étudiées. Nous avons implémenté notre méthode dans un outil dédié à l'annotation des séquences LRR que nous avons appelé « LRRprofiler ».

Le **deuxième chapitre** s'intéresse aux annotations automatiques des récepteurs LRR-RLK, LRR-RLP et NLR chez le riz et présente une ré-annotation manuelle de leurs modèles de gènes. Ce travail a fait l'objet d'un article publié dans la revue *The Plant Journal* (Gottin et al., 2021). Ces trois familles de gènes d'intérêts présentent deux niveaux de complexités pouvant impacter la qualité des données d'annotation. Premièrement, le domaine LRR est un domaine à unités répétées en tandem jusqu'à plus de 30 fois. Deuxièmement, les gènes codant ces récepteurs LRR sont dupliqués, complets ou partiels, dans les génomes. En particulier, les génomes de riz (et d'autres monocotylédones) ont connu une très forte expansion des gènes LRR. Ce contexte est alors propice à l'apparition de biais et d'erreurs dans les données d'annotation pouvant fortement impacter les analyses faites sur ces gènes. Le génome de référence du riz (*O. sativa* ssp *japonica* cv. *Nipponbare*) a été annoté par trois consortiums suivant trois stratégies d'annotation différentes. Ce contexte particulier offre la possibilité de comparer ces stratégies d'annotation pour nos gènes d'intérêt et de mettre en lumière d'éventuels biais ou erreurs.

Le **troisième chapitre** présente l'ensemble des données et outils développés et les moyens mis en œuvre pour les rendre facilement accessibles et réutilisables. En effet, l'étude et la compréhension des mécanismes à l'origine de l'évolution et de l'adaptation des domaines LRR des familles d'intérêt nécessite des outils adaptés et des efforts collectifs. Il est alors essentiel de partager les données et outils dédiés dans des formats facilement exploitables et mobilisables. Ce chapitre aborde en premier lieu les efforts réalisés pour mettre à disposition de la communauté l'ensemble des données de ré-annotation à travers un site web dédié développé en collaboration avec Marilyne Summo (ingénieure CIRAD). Le site permet une visualisation graphique des gènes et de leurs annotations, pour l'ensemble des espèces intégrées à l'étude, offrant un appui graphique pour les analyses comparatives. Le site met également à disposition les données dans différents formats standards les rendant ainsi

exploitables par d'autres outils. Ce chapitre présente ensuite les méthodes et stratégies utilisées pour favoriser la portabilité, la reproductibilité et la transparence de nos outils.

Enfin, le **quatrième chapitre** présente un travail de revue, auquel j'ai eu l'opportunité de participer au début de ma thèse, sur l'origine des récepteurs RLK chez les plantes. Cette revue a été publiée dans le journal *Annual review of plant biology* (Dievart et al., 2020). Pour ce travail, nous nous sommes intéressés aux différentes structures ou associations de domaines dans les RLK chez les plantes. Nous avons retracé l'histoire de leur découverte, et présentons des résultats originaux sur l'émergence des différentes structures au cours de l'évolution des plantes.

Chapitre 1

Chapitre 1 :
Annotations des motifs LRR dans les protéines et classification des séquences en sous-familles

Table des matières

A. Introduction

Les motifs LRR sont très abondants dans les protéomes de plantes. Ils interviennent notamment dans la composition des récepteurs LRR-RLK, LRR-RLP et NLR impliqués dans les fonctions essentielles de développement et de défenses. Le domaine LRR de ces trois familles se compose de 2 à plus de 30 motifs répétés en tandem. A ce jour, 8 consensus différents de motifs LRR ont été décrits (Tableau 2). Les motifs LRR des récepteurs membranaires LRR-RLK et LRR-RLP présentent un consensus conservé de 24 acides aminés appelé '*plant specific*'. Les protéines NLR portent des motifs dont le consensus diverge légèrement du consensus *plant-specific*, avec notamment l'asparagine (N) en position 12 fréquemment remplacée par un résidu cystéine (C) ou tyrosine (T) (Jones and Jones, 1997; Kajava, 1998) (cf. introduction, section D.2.).

Pour pouvoir étudier les domaines LRR des différentes familles de récepteurs, il est important de pouvoir annoter chaque répétition de façon fiable, exhaustive et reproductible. L'utilisation de profils HMM semble pertinent pour l'annotation de ces motifs. Nous avons identifié trois stratégies différentes dans la construction des profils HMM pour la recherche de motifs répétés (cf. introduction, section E.2.3.). Chacune de ces stratégies présente des avantages et inconvénients. La première consiste à utiliser un profil construit à partir du segment HCS des motifs. Cela permet de traiter l'ensemble des motifs avec le même profil, mais le segment HCS étant court, cette stratégie tend à générer des faux positifs. La seconde stratégie consiste à utiliser un profil représentant un motif complet. Mais le segment VS étant très variable entre les différents types de motifs, l'utilisation d'un profil générique tend à générer des faux négatifs. Enfin, il est possible d'utiliser un profil représentant plusieurs motifs consécutifs en autorisant une grande variabilité sur les segments VS pour limiter autant les faux positifs que les faux négatifs. En revanche, les résultats des recherches effectués sur la base de ce type de profil ne permettent pas de déduire directement une annotation motifs par motifs avec des bornes standardisées. Ainsi, pour contourner ces limitations, nous utilisons des profils représentant chacun un motif complet (permettant d'éviter les faux positifs), mais adaptés aux jeux de donnée et à chaque famille de gènes d'intérêt (pour éviter également les faux négatifs).

Tableau 2 : Séquences consensus des différents types de motifs LRR.
Majuscules : résidus conservés dans plus de 50% des répétitions. Minuscules : résidus conservés dans plus de 30% des répétitions. o : résidus non-polaires. x : n'importe quel résidus. Les résidus L peuvent être remplacés par des I, A, V et F. Orange : segment conservé HCS.
D'après Kajava (1998), Kobe et Kajava (2001) et Matsushima *et al.* (2010).

	1	2	3	**4**	5	6	**7**	8	**9**	10	11	**12**	13	**14**	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Typical	**L**	x	x	**L**	x	x	**L**	x	**L**	x	x	**N**	x	**L**	x	x	L	p	x	x	o	F	x	x				
RI-like	x	x	x	**L**	x	x	**L**	x	**L**	x	x	**NC**	x	**L**	x	x	x	g	o	x	x	L	x	x	o	L	x	x
Cys-containing (CC)	**c**	x	x	**L**	x	x	**L**	x	**L**	x	x	**c**	x	x	I	T	D	x	x	o	x	x	L	a	x	x		
Plant-specific (PS)	**L**	x	x	**L**	x	x	**L**	x	**L**	x	x	**N**	x	**L**	ts	g	x	I	P	x	x	L	G	x				
SD22-like	**L**	x	x	**L**	x	x	**L**	x	**L**	x	x	**N**	x	**I**	x	x	I	x	x	L	x	x						
Bacterial	**P**	x	x	**L**	x	x	**L**	x	**V**	x	x	**N**	X	**L**	x	x	L	P	ed	L								
TpLRR	**CN**	x	x	**L**	x	x	**I**	x	**L**	x	x	x	L	**x**	x	I	g	x	x	A	F	x	x					
IRREKO	**NLQx**	x	x	**L**	x	x	**L**	x	**LC**	x	x	**N**	x	**L**	x	x	L	D	L	x	x							

Il existe de nombreuses banques de données mettant à disposition des profils HMM pour une large variété de domaines et de motifs. Parmi les plus utilisées, on trouve les banques Pfam (Mistry et al., 2021) et SMART (Letunic et al., 2021). Ces banques proposent en particulier plusieurs profils HMM différents pour l'identification des motifs LRR. Nous avons voulu, dans un premier temps, déterminer quels étaient les profils HMM publics les plus performants pour l'identification des motifs LRR, i.e. permettant d'obtenir l'annotation la plus complète possible des domaines LRR des gènes d'intérêt LRR-RLK, LRR-RLP et NLR. Dans un second temps, nous avons développé une procédure permettant de créer des profils HMM pour les motifs LRR qui soient à la fois spécifiques du taxon étudié et de la famille de gènes étudiée, i.e. LRR-RLK et LRR-RLP puis NLR. Cette procédure, qui adapte automatiquement le profil HMM aux séquences étudiées, permet d'améliorer la détection des motifs et l'annotation des domaines LRR, en particulier pour les gènes NLR. Elle constitue la base de notre pipeline global, nommé *LRRprofiler*, qui identifie, annote et classe en sous-familles les séquences contenant des motifs LRR d'un protéome donné *de novo*. Ce pipeline est libre d'accès et est publié en matériel supplémentaire du papier présenté en chapitre 3.

B. Matériels et méthodes

B.1. Données protéiques de référence

Les séquences protéiques « expertisées » de la base de données Swiss-Prot (Boutet et al., 2007) disponibles pour l'espèce *Arabidopsis thaliana* ont été téléchargées au format fasta via le serveur ftp (version datant du 01-2021). Cette base fournit des données de qualité, annotées manuellement. Parmi les informations disponibles, la famille de gènes à laquelle chaque séquence appartient (LRR-RLP, LRR-RLK, NLR ou Autre) ainsi que l'annotation des domaines protéiques et des motifs ont été extraites pour les séquences portant des répétitions LRR. Cette extraction a été faite via le système de recherche de l'interface web. Le jeu de données se compose de 16 031 séquences protéiques, dont 548 portent des répétitions LRR (172 LRR-RLK, 61 LRR-RLP, 91 NLR et 224 Autres). Les détails concernant ce jeu de données sont fournis dans le tableau 3.

Tableau 3 : Description des protéines LRR présentes dans le jeu de données test d'A. thaliana de la base de données Swiss-Prot.

	LRR-RLK	LRR-RLP	NLR	Autres	Total
Nombre de protéines	172	61	91	224	548
Nombre de motifs LRR	1 954	1 333	695	1 223	5 205
Nombre moyen de LRR par protéine (min-max)	11,4 (2-31)	21,9 (2-32)	7,6 (0-23)	5,5 (0-25)	12,3 (0-32)

B.2. Profils HMM LRR des bases de données Pfam et SMART

Les profils HMM de motifs LRR des bases de données SMART (Letunic et al., 2021) et Pfam (Mistry et al., 2021) ont été téléchargés. Cela inclut les profils LRR_1 (PF00560), LRR_2 (PF07723), LRR_3 (PF07725), LRR_4 (PF12799), LRR_6 (PF13516), LRR_8 (PF13855) de Pfam et les profils LRR (SM00370), LRR_TYP (SM00369), LRR_CC (SM00367), LRR_RI (SM00368), LRR_BAC (SM00364) et LRR_SD22 (SM00365) de SMART. Une présélection a été effectuée pour ne tester que les profils représentant un seul motif LRR et connus pour générer des hits significatifs chez les plantes supérieures. Ceux-ci ont été identifiés en se basant sur les informations fournies dans les descriptions des profils de chacune des bases de données. Les profils Pfam LRR_4 et LRR_8 n'ont pas été testés car ils représentent plusieurs motifs LRR consécutifs. Leur intérêt dans notre contexte est limité car ils ne permettent pas une annotation précise et reproductible des motifs LRR. Les profils représentant plusieurs motifs consécutifs vont identifier des répétitions dont les bornes de début et de fin de motif ne sont pas régulières. Cela nous pose des problèmes de standardisation pour la détermination des frontières entre chaque motif et rend donc leur comparaison compliquée. Ces deux profils ne seront donc pas utilisés pour l'annotation des séquences LRR. Ainsi sept profils ont été retenus : quatre profils de Pfam (LRR_1, LRR_2, LRR_3 et LRR_6) et trois profils de SMART (LRR, LRR_TYP et LRR_CC).

Ces sept profils HMM ont été utilisés, avec le programme *hmmsearch* de la suite HMMER, pour identifier les répétitions LRR au sein des séquences protéiques d'*A. thaliana*. La recherche a été effectuée en utilisant différents seuils de *e*-value qui peuvent être déterminés via les options `-E` (*e-value* sur le motif) et `--domE` (*e-value* sur le domaine complet). Lancer le programme *hmmsearch* avec les options `-E 1000 --domE 1000` permet de supprimer le filtrage des résultats en fonction des seuils de *e-value*.

B.3. Outils utilisés pour la reconstruction itérative de profils HMM

Deux sous-jeux de données indépendants sont constitués, l'un contenant les protéines identifiées avec un domaine kinase et l'autre contenant celles identifiées avec un domaine NB-ARC. Les domaines kinases sont détectés via l'outil iTAK (v1.7a) (Zheng et al., 2016) en utilisant les paramètres par défaut. Les domaines NB-ARC sont identifiés avec le profil Pfam NB-ARC (PF00931) et le programme *hmmsearch* de la suite HMMER (v3.1b2) (Eddy, 2011) en utilisant les options `-E 1000` et `--domE 1000`. Le profil LRR initial du processus itératif est le profil LRR (SM00370) de SMART. La recherche des motifs au sein des séquences protéiques est faite à l'aide du programme *hmmsearch* avec les options `-E 1000` et `--domE 1000` également et les nouveaux profils HMM sont construits via le programme *hmmbuild* de la suite HMMER en utilisant les paramètres par défaut. Les alignements protéiques sont réalisés avec MAFFT (v7.271) (Katoh and Standley, 2013) en utilisant les paramètres par défaut.

B.4. Développement et validation du pipeline LRRprofiler

B.4.1. Annotation des motifs LRR

L'annotation des motifs LRR intègre les profils construits itérativement pour les LRR-RLK et pour les NLR, ainsi que les profils publics LRR (SM00370), LRR_TYP (SM00369), LRR_CC (SM00367), LRR_RI (SM00368), LRR_BAC (SM00364) et LRR_SD22 (SM00365) de SMART. La recherche des profils dans les séquences protéiques est réalisée avec le programme *hmmsearch* et les options `-E 1000` et `--domE 1000`.

B.4.2. Annotation des domaines protéiques

L'annotation des domaines protéiques associés aux LRR intègre les profils publics NB-ARC (PF00931), TIR (PF01582), TIR_2 (PF13676), RPW8 (PF05659), F-box (PF00646), FBD (PF08387), malectin (PF11721) et malectin-like (PF12819) de Pfam. La recherche des profils dans les séquences protéiques est réalisée avec le programme *hmmsearch* et les options `-E 1000` et `--domE 1000`. Les domaines kinases sont détectés via l'outil iTAK en utilisant les paramètres par défaut. Les domaines transmembranaires (TM) sont détectés via TMHMM (v2.0c) (Sonnhammer et al., 1998) avec l'option `-noplot` (sans construction des images, permettant de rendre l'exécution plus rapide et réduire la taille des résultats en mémoire).

B.4.3. Distribution du pipeline

Les scripts du pipeline sont sauvegardés et disponibles librement sur github (https://github.com/cgottin/LRRprofiler). Afin de faciliter la distribution et l'utilisation du pipeline, l'ensemble des dépendances (HMMER, TMHMM, iTAK, MAFFT et les profils HMM publics) et des scripts sont intégrés à un container Singularity disponible publiquement sur le *cloud* Sylabs (https://cloud.sylabs.io/library/_container/600ea381517f0358917abf0a). L'exécution du pipeline se fait en ligne de commande avec une seule instruction :

```
singularity run lrrprofiler.sif --in_proteome <fasta_file> --name <string>
```

B.4.4. Données supplémentaires pour la validation du pipeline

Les séquences protéiques du protéome complet pour *A. thaliana* ont été téléchargées depuis le site *tair* (https://www.arabidopsis.org/). Les données correspondent aux séquences représentatives (une protéine par gène) de la version TAIR10 (27 416 séquences, 06-2020).

C. Résultats

C.1. Les profils publics Pfam et SMART des motifs LRR sont perfectibles

Nous avons testé sept profils HMM LRR. Le tableau 4 présente les consensus de ces profils centrés sur les six premiers résidus du segment conservé « LxxLxL ».

Tableau 4 : Consensus des profils LRR publics issus des bases de données Pfam et SMART.
Les séquences consensus sont centrées sur le segment HCS (surligné en jaune).

Identifiant	Source	Nom	Taille	Séquence Consensus
PF00560	Pfam	LRR_1	23	x **LxxLxL** xxNxxxxxxxxxFxxL
PF07723	Pfam	LRR_2	26	x **LKxLxL** xxVxFxxxxxxxxLLxxCP
PF07725	Pfam	LRR_3	20	x **LxxLxL** xxSxxxxLWxGxx
PF13516	Pfam	LRR_6	24	xxxx **LxxLxL** xxNxIxxxGxxxLx
SM00370	SMART	LRR	24	Lxx **LxxLxL** xxNxLxxLPxxxFxx
SM00369	SMART	LRR_TYP	28	xxx **LxxLxL** xxNxLxxxGxxxLxxxLxx
SM00367	SMART	LRR_CC	24	Lxx **LxxLxL** xxNxLxGxIPxxLxx

Les performances de ces profils HMM ont été comparées en les testant sur le jeu de données de référence composé des 16 031 séquences protéiques d'*Arabidopsis thaliana* issues de la base de données Swiss-Prot. Les tests ont été réalisés en utilisant le programme *hmmsearch* avec différents seuils de *e*-value pour retenir un *hit*. La spécificité des profils pour la structure

des motifs LRR est très élevée et aucun faux positif n'a été détecté même lors du retrait total du seuil de *e-value*. Quand un profil identifie plus de motifs que décrit par l'annotation de Swiss-Prot, une vérification manuelle a été systématiquement réalisée. Toutes ces vérifications ont confirmé la validité des motifs prédits. Les résultats présentés ci-dessous sont ceux obtenus après retrait des seuils de *e*-value.

Le tableau 5 présente le nombre de protéines et de motifs identifiés avec chacun des profils testés. Les profils Pfam LRR_1 et LRR_6 et les profils SMART LRR et LRR_CC présentent de bons résultats pour les protéines membranaires LRR-RLP et LRR-RLK, avec au minimum 85% des protéines attendues identifiées. En revanche, les résultats pour les protéines intracellulaires NLR ne sont pas satisfaisants avec au maximum seulement 54% des protéines NLR attendues identifiées (49 protéines NLR sur 91 identifiées avec le profil LRR_1 de Pfam) (Tableau 5, Figure 15).

Tableau 5 : Résultats d'identification des protéines LRR et de détection des motifs LRR via les profils publics dans le jeu de données test.
Les données de références de la base de données Swiss-Prot sont rappelées.

	LRR-RLK		**LRR-RLP**		**NLR**	
	Nb. protéines	*Nb. motifs*	*Nb. protéines*	*Nb. motifs*	*Nb. protéines*	*Nb. motifs*
Swiss-Prot (ref)	**172**	**1 954**	**61**	**1 333**	**91**	**695**
HMM LRR_1	171	1 765	61	1 163	49	350
HMM LRR_2	27	139	10	64	1	6
HMM LRR_3	0	0	0	0	32	84
HMM LRR_6	145	1 266	58	1 045	11	51
HMM LRR	171	1 661	61	1 174	33	209
HMM LRR_TYP	98	552	38	547	7	37
HMM LRR_CC	172	2 013	61	1 313	42	225
Nouveau profil LRR-RLK	171	2 013	61	1 325	35	192
Nouveau profil NLR	60	303	26	390	85	905

Pour comprendre ces différences de performance nous avons comparé les consensus de ces profils avec ceux obtenus après avoir aligné les motifs LRR annotés dans la base de données Swiss-Prot avec le programme MAFFT (Figure 16). Les alignements des motifs LRR de Swiss-Prot issus de gènes LRR-RLK et LRR-RLP présentent un consensus similaire correspondant au consensus « *plant specific* » décrit dans la littérature (Tableau 2). En revanche, les motifs présents dans les protéines NLR ont un consensus global qui diffère du consensus observé chez les LRR-RLP et LRR-RLK. Les motifs LRR des gènes NLR présentent des similitudes avec le

consensus « *cys-containing* » (Tableau 2). On y observe un résidu cystéine en position 12 à la place de l'asparagine, mais ils ne présentent pas la cystéine en position 1 (Figure 16c). Il présente également le dipeptide 'LP' conservé en position 18 et 19 comme pour le consensus « *plant specific* ».

Pour conclure sur les comparaisons des profils publics, un profil très satisfaisant existe dans la base de données SMART pour l'identification des motifs LRR des gènes LRR-RLK et LRR-RLP (LRR_CC, SM00367). En revanche, aucun des profils testés issus de la base de données SMART ne présente des niveaux satisfaisants de sensibilité vis-à-vis des motifs LRR pour les protéines NLR.

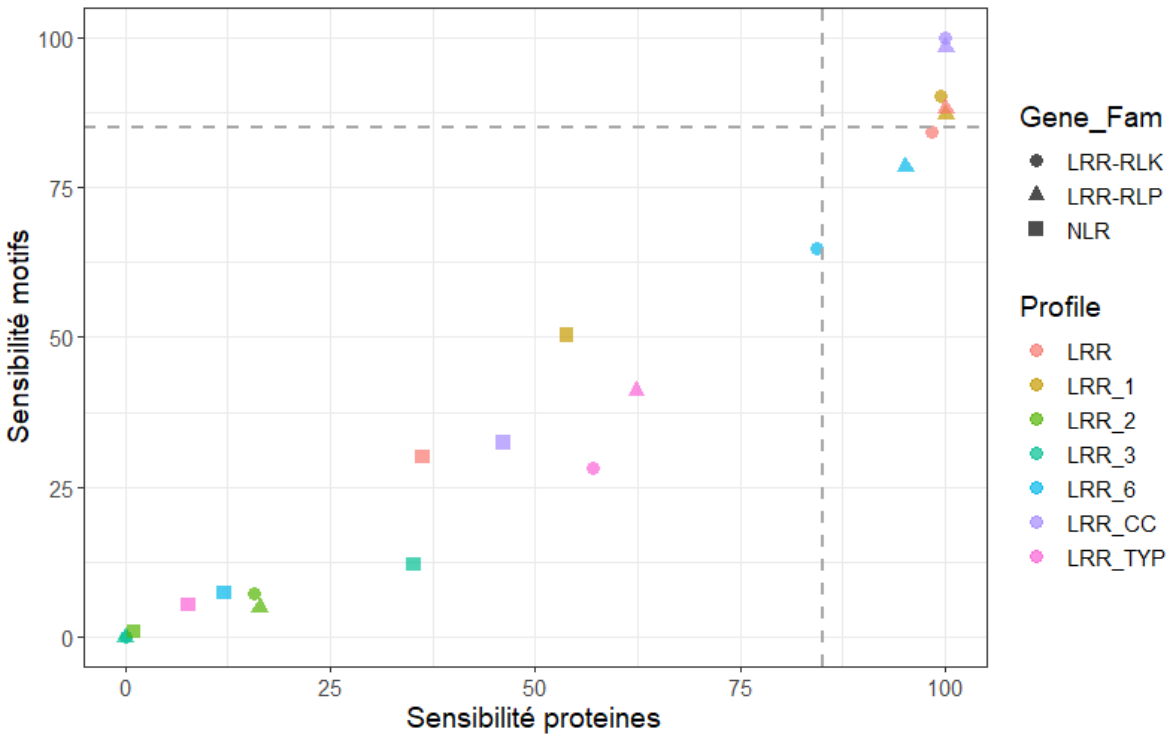


Figure 15 : Sensibilité des profils publics pour l'identification des protéines d'intérêt et l'annotation des motifs LRR de ces protéines.
Les lignes pointillées grises représentent la sensibilité de 85% pour les protéines (verticale) et les motifs (horizontale).

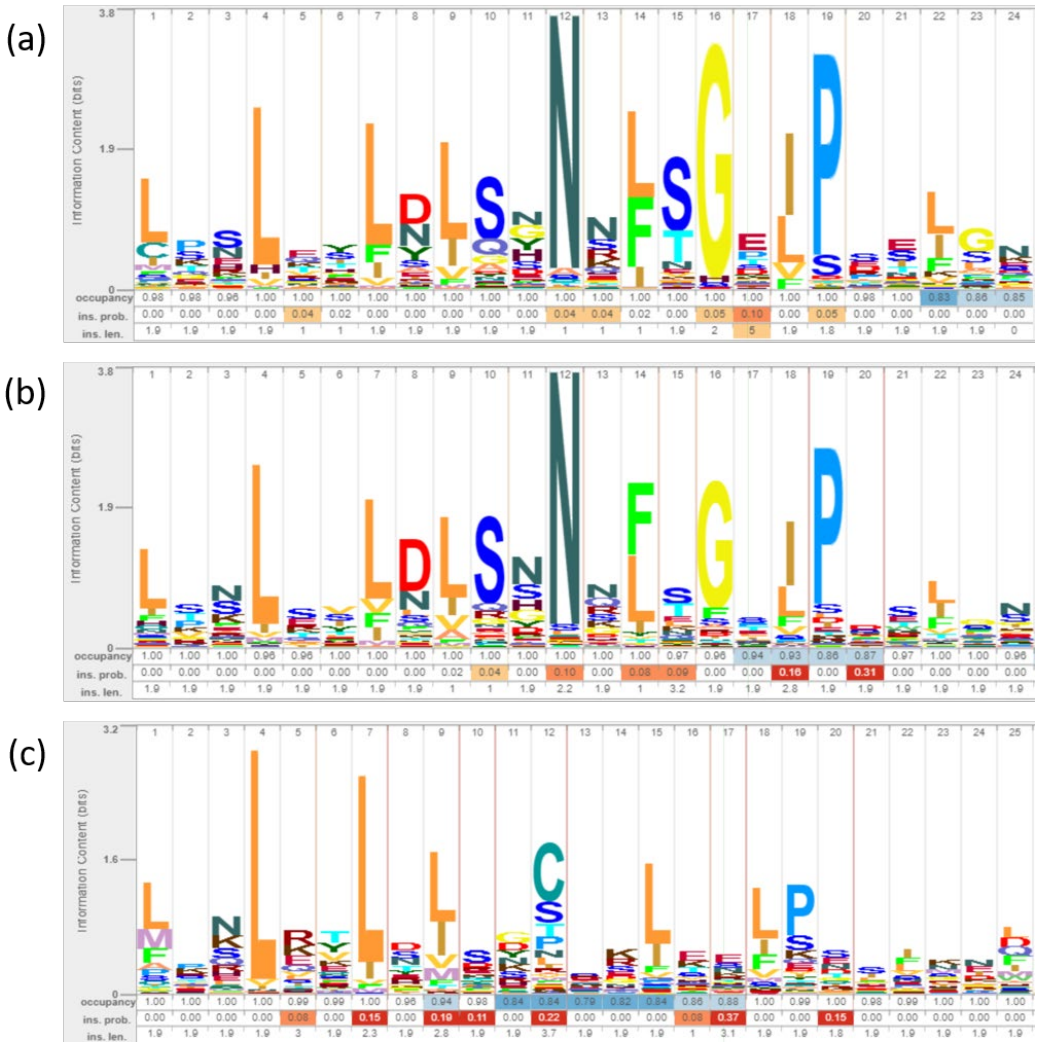


Figure 16 : Weblogo des alignements de motifs LRR issus des séquences protéiques d'A. thaliana issues de Swiss-Prot.
(a) *LRR-RLK, (b) LRR-RLP et (c) NLR. Images générées sur skylign.org.*

C.2. Une reconstruction itérative et famille-spécifique des profils HMM permet d'améliorer la détection et l'annotation des motifs LRR des récepteurs NLR

Les profils issus de la base de données SMART semblent plus pertinents car ils débutent avec un résidu ayant un haut niveau de conservation ce qui facilite l'annotation des bornes des motifs (Tableau 4). Afin d'améliorer nos capacités d'annotation des motifs pour les gènes NLR, nous avons construit un nouveau profil spécifique à cette famille et ayant les mêmes frontières de consensus que le profil LRR_CC. De cette façon, les annotations des motifs LRR générées avec les profils SMART et avec le nouveau profil NLR pourront être comparables.

C.2.1. Une stratégie itérative pour adapter les profils HMM

Différentes stratégies existent pour augmenter la spécificité d'un profil vis-à-vis d'un jeu de données particulier. L'une d'elle consiste à modifier l'alignement servant à construire le profil au vu des séquences identifiées par ce profil dans le jeu de donnée cible (Ng et al., 2011; Terrapon et al., 2012; Anisimova et al., 2015). Plus précisément, à partir d'un profil initial, on recherche des motifs dans le jeu de donnée cible. Les motifs identifiés sont ensuite alignés, avec ou sans les motifs du profil initial. Un nouveau profil est construit à partir de cet alignement. Le nouveau profil est alors plus apte à identifier les motifs de ce jeu de données vu qu'il en intègre directement certaines spécificités. Cette stratégie a déjà été utilisée avec succès dans le contexte des LRR par Ng *et al.* (Ng et al., 2011) pour une étude chez l'humain.

Nous avons adapté cette stratégie au cas des LRR en développant une approche itérative qui est représentée de manière schématique dans la Figure 17a. Chaque itération est constituée de trois étapes. La première étape consiste à identifier des motifs LRR dans les séquences protéiques du jeu de séquences candidates. Le programme *hmmsearch* de la suite HMMER est utilisé pour identifier les motifs LRR au sein des séquences protéiques grâce au profil HMM « LRR » de SMART (SM000370). La seconde étape consiste à filtrer ces motifs, en ne gardant que ceux passant différents critères de qualité, puis à aligner ces motifs avec le programme MAFFT. L'alignement ainsi obtenu sert alors à l'étape trois pour construire le profil HMM, avec le programme *hmmbuild* de la suite HMMER. Ce nouveau profil HMM sera celui utilisé pour la recherche de motifs de l'itération suivante. Dans cette approche, l'alignement initial au début d'une itération est totalement remplacé (et non complété) par les motifs identifiés à la fin de cette itération, cela permet de limiter les biais liés à une surreprésentation de certains motifs dans l'alignement. A chaque itération, le nombre de protéines identifiées, le nombre de motifs annotés (nbMotifs) et la taille moyenne de ces motifs (lgMoyMotifs) sont calculés et conservés. En l'absence de faux positifs (cf. introduction, section E.2.3.), le nombre de résidus annotés comme faisant partie d'un motif LRR (i.e. nbMotifs * lgMoyMotifs) est utilisé comme indicateur de la qualité d'une annotation. Cet indicateur à tendance à s'améliorer au fil des itérations puis à se stabiliser une fois le profil HMM optimal atteint. Il peut cependant arriver que cet indicateur baisse légèrement avant de repartir à la hausse. Le critère utilisé pour définir la condition d'arrêt de notre procédure itérative prend donc cette possibilité d'optimums locaux en compte. Ainsi la reconstruction des profils prend fin non pas dès qu'une

itération n'améliore pas le critère par rapport à la meilleure solution trouvée jusque-là, mais au bout de la troisième fois où cela se produit.

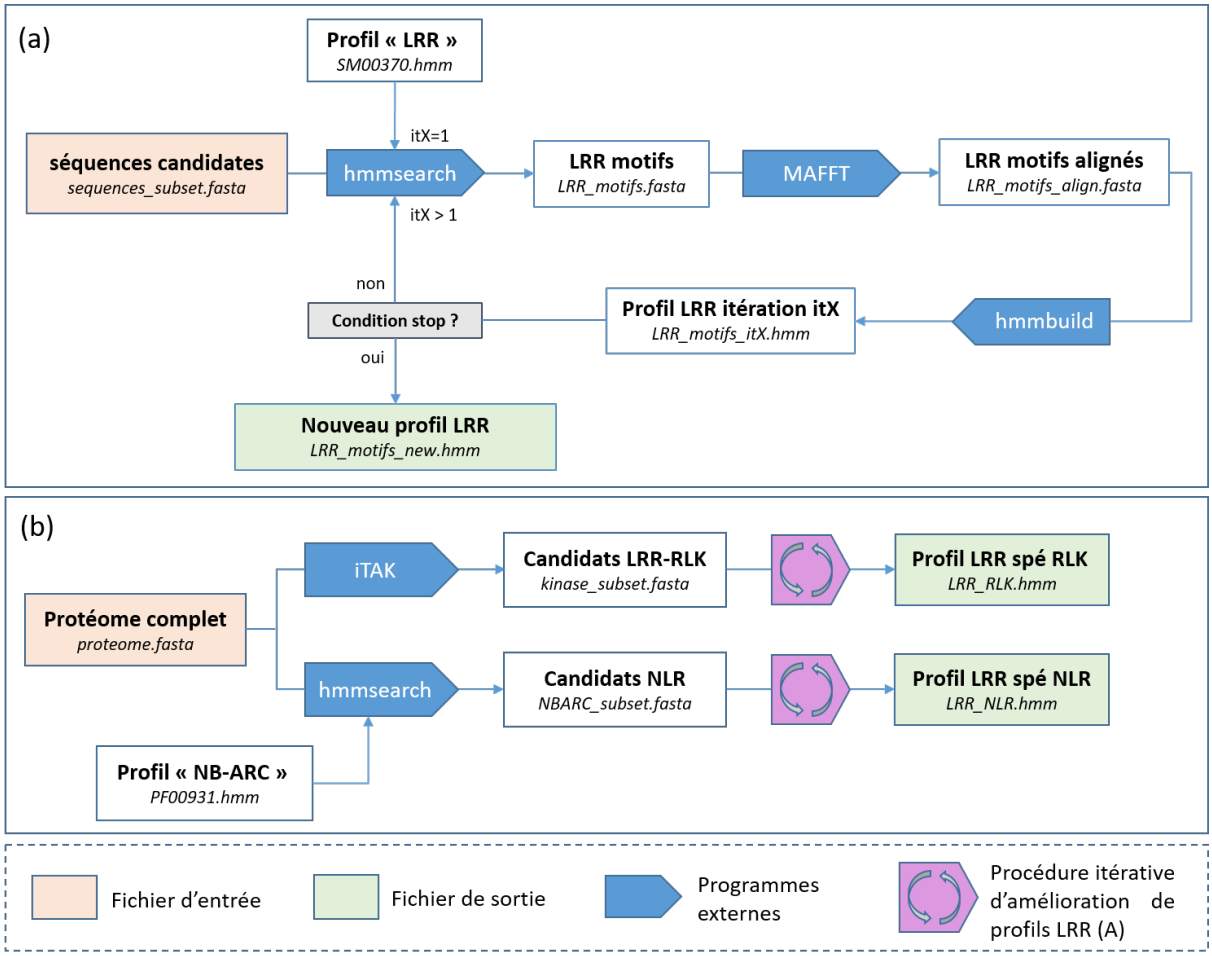


Figure 17 : Procédure d'amélioration des profils HMM pour les motifs LRR.
(a) Schéma de la procédure itérative. (b) Représentation schématique de la création des jeux de données candidats pour les séquences LRR-RLK et NLR.

D'après les observations faites sur les différences de consensus LRR entre les récepteurs membranaires et les NLR, nous choisissons de traiter séparément les différentes familles. Nous pouvons facilement identifier des séquences candidates pour les LRR-RLK et les NLR dans un protéome donné grâce à la présence mutuellement exclusive d'un domaine kinase chez les LRR-RLK et d'un domaine NB-ARC chez les NLR. Ainsi, la procédure d'amélioration présentée en figure 17a est exécutée indépendamment pour les LRR-RLK d'un côté, et les NLR de l'autre. Pour ce faire, le protéome d'intérêt est soumis à une recherche du domaine kinase et du domaine NB-ARC pour constituer deux jeux de données candidats pour les motifs issus des LRR-RLK et ceux issus des NLR respectivement (Figure 17b). L'outil iTAK est utilisé pour

identifier les séquences portant un domaine kinase. Le domaine NB-ARC est identifié par le profil HMM « NB-ARC » (PF00931) de la base de données Pfam à l'aide du programme *hmmsearch*.

C.2.2. Validation de la stratégie proposée

Afin de valider notre stratégie, nous utilisons le protéome d'*A. thaliana* de Swiss-Prot sur lequel l'annotation attendue est connue. Le principe de la validation est d'appliquer notre stratégie d'amélioration de profil HMM en partant d'un profil ayant des performances moyennes et de voir si cela permet d'obtenir une annotation aussi bonne que celles obtenues à l'aide des meilleurs profils disponibles. Plus spécifiquement, c'est l'amélioration du profil SMART « LRR » (SM000370) qui présente des résultats moyens pour les LRR-RLK, qui nous permettra de vérifier la faisabilité de la stratégie. Le nouveau profil créé pourra être comparé au profil « LRR_CC » de SMART qui, lui, présente de très bonnes performances pour l'annotation des protéines membranaires LRR-RLK.

Le protéome d'*A. thaliana* est soumis à une recherche du domaine kinase avec iTAK pour constituer le jeu de données candidat pour les motifs issus des LRR-RLK. A chaque itération, on extrait le nombre de protéines LRR-RLK identifiées (présentant au moins un hit de profil LRR) ainsi que le nombre de motifs détectés. Les résultats présentés en figure 18 montrent que pour les motifs LRR-RLK, un profil plus performant est obtenu dès la première itération. En effet, si ce nouveau profil n'identifie qu'une seule protéine LRR-RLK supplémentaire (172 vs 171) que le profil initial SM00370, il identifie par contre 465 motifs LRR supplémentaires (1988 vs 1523) et les motifs identifiés sont en moyenne beaucoup plus long (20,0 vs 14,6 résidus) ce qui correspond davantage à la taille attendue d'un motif LRR qui est de 24 acides aminés. Le profil exporté par la procédure comme étant le « meilleur » profil pour les LRR-RLK est celui de l'itération n°5 (Figure 19). Le nouveau profil permet d'identifier 172 protéines et 2027 motifs LRR pour les LRR-RLK. Cela représente le même nombre de protéines et 0,6% de motifs en plus que le profil LRR_CC. Par ailleurs, il permet d'identifier 61 protéines et 1325 motifs LRR pour les LRR-RLP soit autant de protéines et 0,9% de motifs en plus que le profil LRR_CC. Les performances du nouveau profil créé sont donc meilleures que celles du profil initial et similaires au profil LRR_CC, ce qui nous permet de valider la pertinence de la procédure.

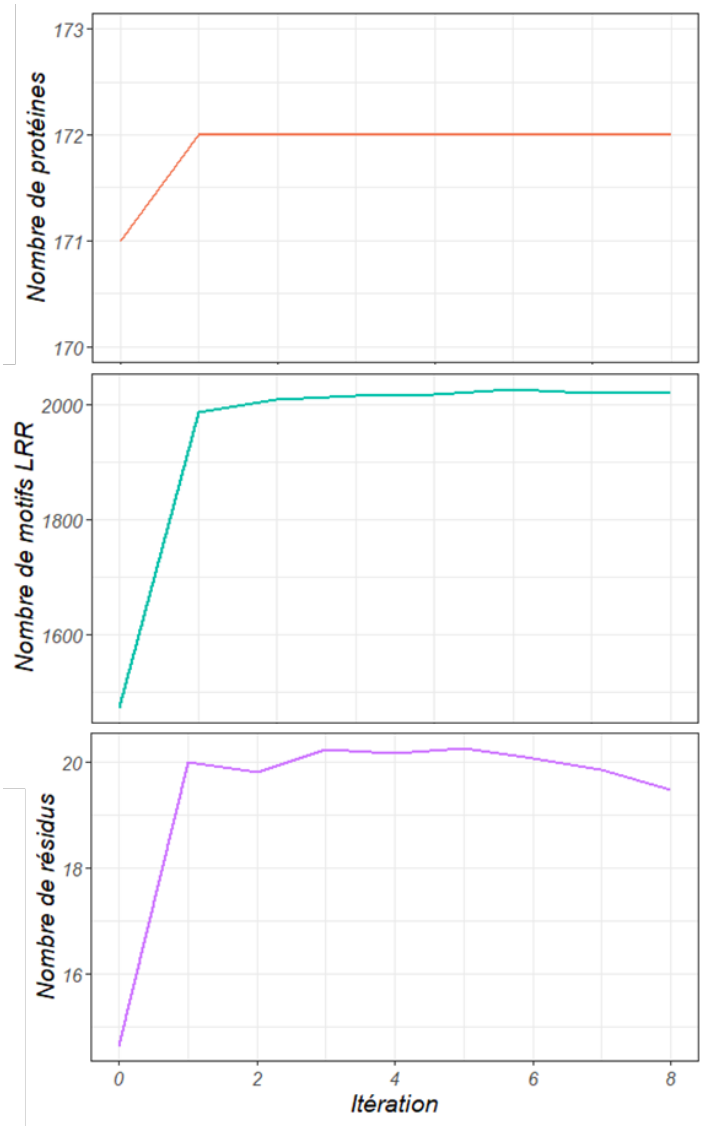


Figure 18 : Performances du profil HMM en cours de construction pour l'annotation des récepteurs LRR-RLK en fonction de l'itération.
La figure présente : (i) en orange, le nombre de protéines identifiées par le profil à chaque itération ; (ii) en vert, le nombre de motifs identifiés par le profil à chaque itération ; (iii) en violet, le nombre moyen de résidus identifiés par motifs LRR. Les valeurs sont également données sous forme de tableau en annexe 1.

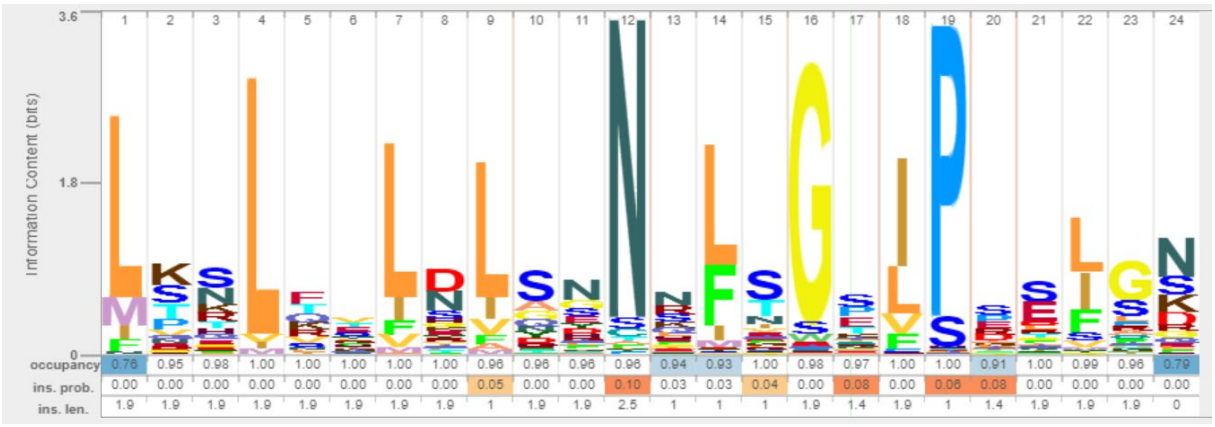


Figure 19 : Weblogo représentant le profil LRR final pour les motifs LRR-RLK.
Image générée via https://skylign.org.

C.2.3. Résultats du nouveau profil pour l'annotation des motifs LRR des récepteurs NLR

La procédure d'amélioration est alors appliquée aux protéines NLR avec le même profil initial SM00370 (« LRR »). L'évolution du nombre de protéines, de motifs LRR et de résidus identifiés en fonction des itérations est présentée en figure 20. Pour les gènes NLR, le gain majeur d'efficacité apparait à la deuxième itération, où 44 protéines et 660 motifs de plus sont identifiés. On remarque également une diminution marquée du nombre de motifs détectés à la septième itération. Le profil extrait par la procédure est celui de la 6ème itération (Figure 21). Ce nouveau profil permet d'identifier 97,8% (89) des protéines NLR annotées par Swiss-Prot et 130% (905) des motifs. A noter que quatre gènes identifiés « NLR » dans Swiss-Prot n'ont aucune annotation de motifs LRR dans cette base de données. De même, pour 81 séquences NLR (89%), la base de données Swiss-Prot annote moins de motifs LRR que ce qui est identifié avec le nouveau profil. Après expertise manuelle de ces motifs supplémentaires, nous avons conclu que ces nouveaux motifs sont bien présents, avec un consensus LRR parfaitement identifiable (Tableau 6).

Tableau 6 : Séquences des motifs LRR identifiés avec le nouveau profil NLR dans les quatre gènes NLR sans annotation de motifs dans Swiss-Prot.

Identifiant	Début	Fin	**L**xx**L**xx**L**x**L**xx**C**xx**L**xx**IP**xx**L**xx
P59584	577	601	**L**TL**L**RV**L**D**L**SW**V**KFE**G**GK**LP**CS**I**GG
P59584	602	624	**L**IH**L**RY**L**R**L**YG**A**V**V**SH**LP**ST**M**RN
P59584	625	648	**L**KL**L**LY**L**N**L**SV**H**NED**L**IH**VP**NV**L**K
P59584	695	717	**M**TK**L**RN**L**T**V**SL**S**ERYNFKT**L**SSS
P59584	749	770	**F**IH**L**KE**L**G**L**VV**R**MSK**IP**DQHQF
P59584	844	867	**M**PC**L**RT**L**T**I**HD**C**EK**L**KE**LP**DG**L**KY
Q9C646	576	600	**V**KL**L**RV**L**D**L**VQ**A**KFK**G**GK**LP**SD**I**GK
Q9C646	601	622	**L**IH**L**RY**L**S**L**KD**A**K**V**SH**LP**SS**L**R
Q9C646	624	647	**L**VL**L**IY**L**D**I**RT**D**FTDIF**VP**NV**F**MG
Q9C646	660	682	**M**HEKTK**L**E**L**SN**L**EK**L**EA**LE**NFST
Q9C646	693	718	**M**VR**L**RT**L**V**I**IL**S**EGTS**L**QT**LS**AS**V**CG
Q9C646	745	766	**F**TY**L**KK**L**T**L**SI**E**MPR**LP**KIQHL
Q9C646	767	788	**P**SH**L**TV**L**D**L**SY**C**C**L**EE**DP**MP**I**L
Q9C646	791	814	**L**LE**L**KD**L**S**L**DY**L**SFSGRKMVCSAG
Q9C646	840	862	**M**SR**L**HT**L**S**I**WS**S**T**L**KE**LP**DG**L**RF
Q9C646	863	887	**I**YS**L**KN**L**I**M**GK**S**WMER**L**SERGEE**F**Y
Q9M667	609	632	**L**RF**L**QT**L**F**V**SD**N**YF**I**EETIDLRKL
Q9STE7	658	682	**G**VN**L**QT**L**RSIS**S**YSWSK**L**NHEL**L**RN
Q9STE7	740	762	**F**PS**L**ES**L**T**L**VG**T**T**L**EENS**MP**A**L**Q

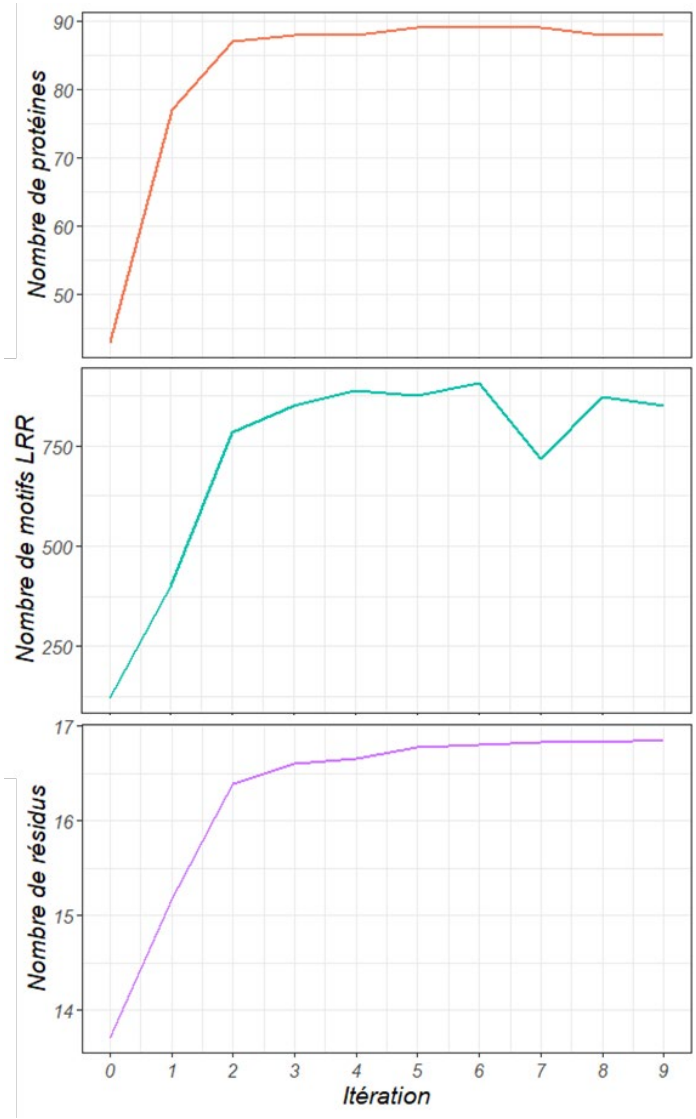


Figure 20 : Performances du profil HMM en cours de construction pour l'annotation des récepteurs NLR en fonction de l'itération.
La figure présente : (i) en orange, le nombre de protéines identifiées par le profil à chaque itération ; (ii) en vert, le nombre de motifs identifiés par le profil à chaque itération ; (iii) en violet, le nombre moyen de résidus identifiés par motifs LRR. Les valeurs sont également données sous forme de tableau en annexe 2.

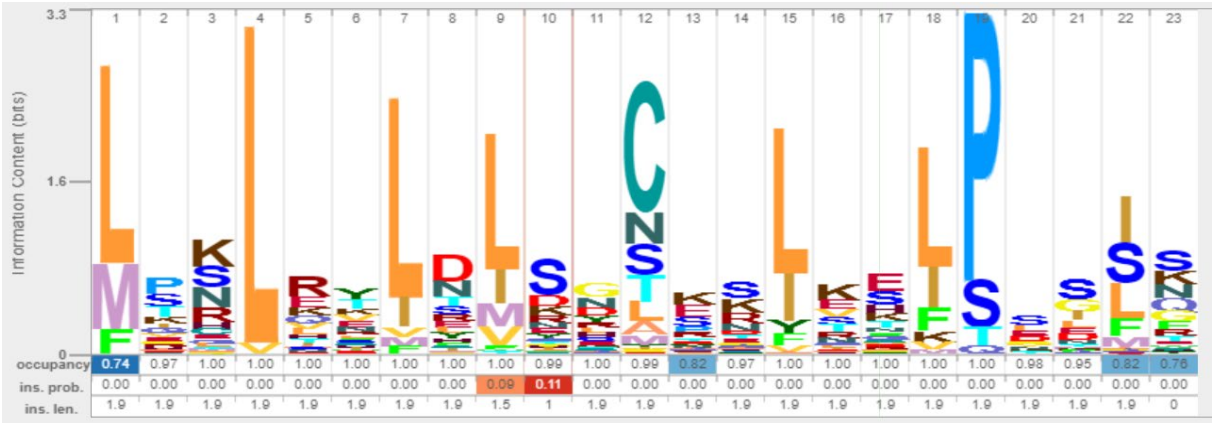


Figure 21 : Weblogo représentant le profil LRR final pour les motifs NLR.
Image générée via https:// skylign.org

C.3. LRRprofiler. Un outil intégratif pour la détection et l'annotation des récepteurs à motifs LRR

Afin d'obtenir une annotation complète standardisée des récepteurs LRR dans un protéome donné, nous avons développé un pipeline appelé LRRprofiler. Ce pipeline permet d'identifier les récepteurs LRR au sein d'un protéome, d'annoter leurs domaines et motifs, et de les classer au sein de sous-familles LRR sans *a priori*.

C.3.1. Développement du pipeline LRRprofiler

Ce pipeline est divisé en 3 étapes (Figure 22). La première implémente la construction spécifique des profils HMM pour les motifs LRR des gènes LRR-RLK et des gènes NLR tel que présenté dans la section précédente (section C.2., Figure 17).

La seconde étape consiste à identifier les motifs LRR au sein des séquences protéiques. L'ensemble du protéome est soumis à une recherche *hmmsearch* avec les deux nouveaux profils HMM et les six profils HMM LRR de la base de données SMART (LRR, LRR_CC, LRR_RI, LRR_TYP, LRR_BAC, LRR_SD22). Les sorties du programme *hmmsearch* pour chaque profil sont 'brutes' et présentent (i) des redondances, c'est-à-dire qu'un motif peut être identifié par plusieurs profils, et (ii) des imprécisions surtout sur les extrémités des motifs qui peuvent être améliorées. Pour améliorer la précision des prédictions et fournir une annotation finale unifiée, une étape de filtration et de correction a été implémentée. L'ensemble des motifs prédits est alors filtré et corrigé selon la procédure présentée en figure 23. Les prédictions redondantes ou chevauchantes sont filtrées, en donnant la priorité à la prédiction de meilleur score pour une position donnée. La taille moyenne des motifs LRR chez les plantes est de 24 acides aminés. Les différents profils utilisés ont une taille de 20 à 28 acides aminés. Lors d'une prédiction de motifs avec un profil donné, l'alignement à l'origine de cette prédiction peut être partiel. Par exemple pour une prédiction avec un profil de taille 24 acides aminés, l'alignement peut s'étendre des position 2 à 21 uniquement, les positions 1, 22, 23 et 24 restant non alignées. Dans ce cas, cette prédiction est étendue pour tenter de compléter le motif. La priorité est donnée à l'extension en 5', jusqu'à atteindre la position 1 du profil si possible ou, en cas de chevauchement avec le motif précédent, jusqu'à rejoindre la fin de celui-ci. On procède ensuite à l'extension en 3' du motif jusqu'à atteindre la fin du profil (ici la position 24) ou, idem en cas de chevauchement, jusqu'à atteindre le début du motif suivant. L'extension prioritaire en 5' est réalisée simultanément pour l'ensemble des motifs prédits

pour une séquence. De même, l'extension en 3' est ensuite réalisée simultanément pour l'ensemble des motifs de la protéine. Les tailles des motifs pouvant varier, si un écart de moins de trois acides aminés existe entre les prédictions de deux motifs consécutifs après l'étape d'extension, les résidus du *gap* sont ajoutés à la fin du motif précédant ce *gap*. Les écarts plus larges sont identifiés comme des fragments inter-LRR. Finalement, une dernière recherche visant à identifier des motifs irréguliers est réalisée. Elle consiste en une recherche *blastp* (Altschul et al., 1990) entre les motifs LRR identifiés par les profils HMM et les fragments de protéines identifiés « inter-LRR » avec un seuil de *e*-value fixé à 1. La recherche BLAST étant moins stringente que la recherche avec des profils HMM, elle permet d'identifier des motifs dont certains résidus respectent moins le consensus attendu.

La troisième étape consiste à identifier les autres domaines protéiques pouvant être associés aux domaine LRR et à classer chaque séquence dans la famille appropriée en fonction de ces résultats. L'annotation des domaines kinases et NB-ARC est récupérée de la première étape (iTAK + hmmsearch). Les profils TIR (PF01582), TIR_2 (PF13676), RPW8 (PF05659), Malectin (PF11721), Malectin-like (PF12819), F-box (PF00646) et FBD (PF08387) de Pfam sont utilisés avec *hmmsearch* pour annoter les domaines correspondants. Les domaines TM sont identifiés via TMHMM. La famille LRR de chaque séquence est déduite à partir de l'annotation complète des domaines et motifs (Figure 24). Les séquences présentant un domaine F-box ou FBD ou les deux sont classées « F-box-LRR ». Les séquences présentant un domaine NB-ARC seul ou associé à un domaine TIR ou RPW8 sont classées « NLR ». Les séquences restantes présentant au moins un domaine kinase sont classées « LRR-RLK ». Ensuite, les séquences ne présentant aucun des domaines précédents et présentant un domaine Malectin ou Malectin-like sont classées « LRR-RLP ». De même, les séquences présentant des LRR associés uniquement à un domaine TM sont classées LRR-RLP. En revanche, les domaines TM sont difficiles à identifier. Des tests réalisés sur les séquences LRR-RLP d'*A. thaliana* issues de Swiss-Prot montrent que TMHMM ne parvient pas à identifier de domaine TM pour 12 des 61 séquences (20%). Ainsi, les structures avec uniquement des motifs LRR (*LRR-only*) sont également classées LRR-RLP si elles présentent plus de 13 motifs LRR et si plus de 60% des motifs identifiés le sont avec un consensus *plant specific*. Toutes les structures restantes sont classées « autre » (*other*).

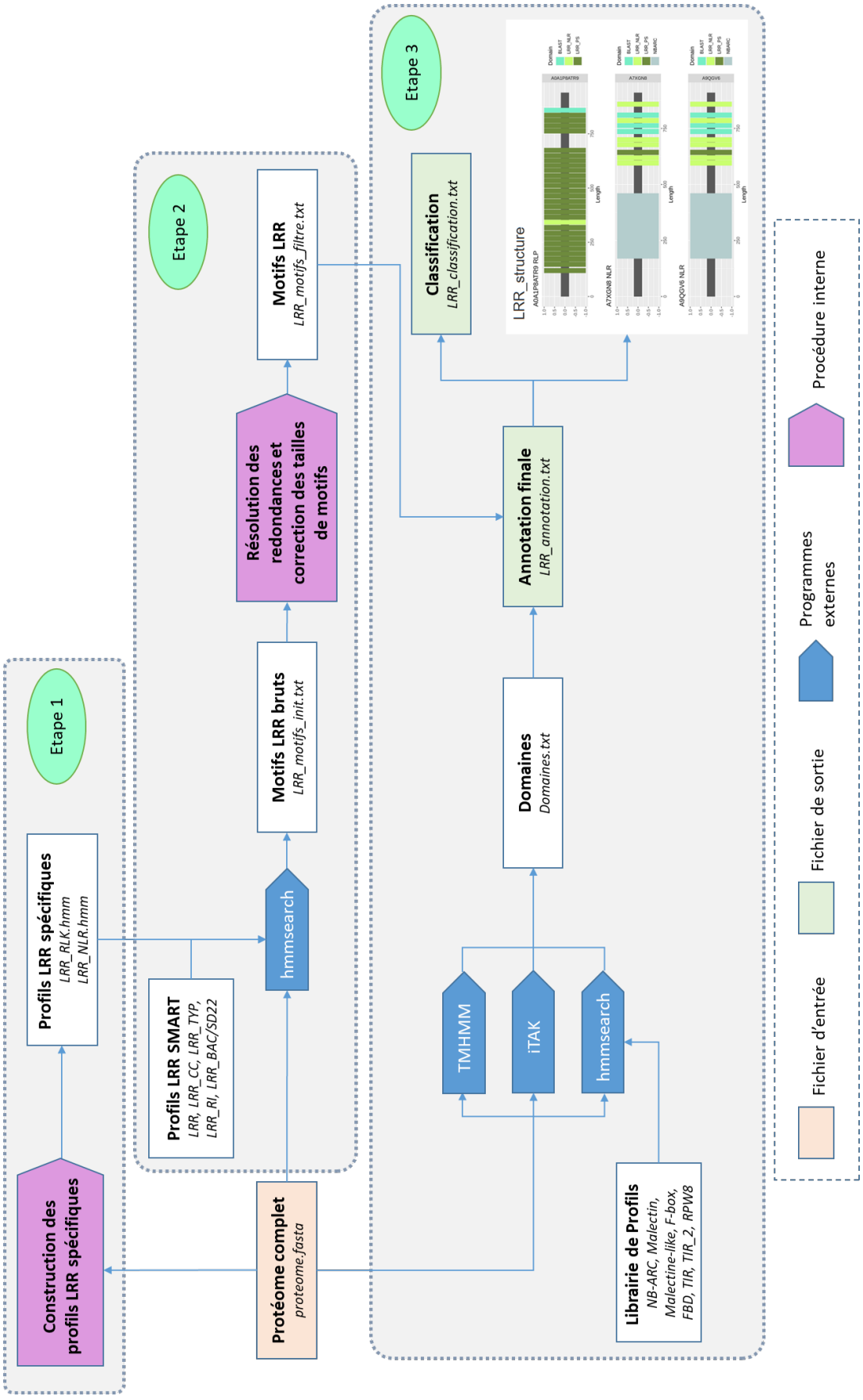


Figure 22 : Représentation schématique du pipeline LRRprofiler.

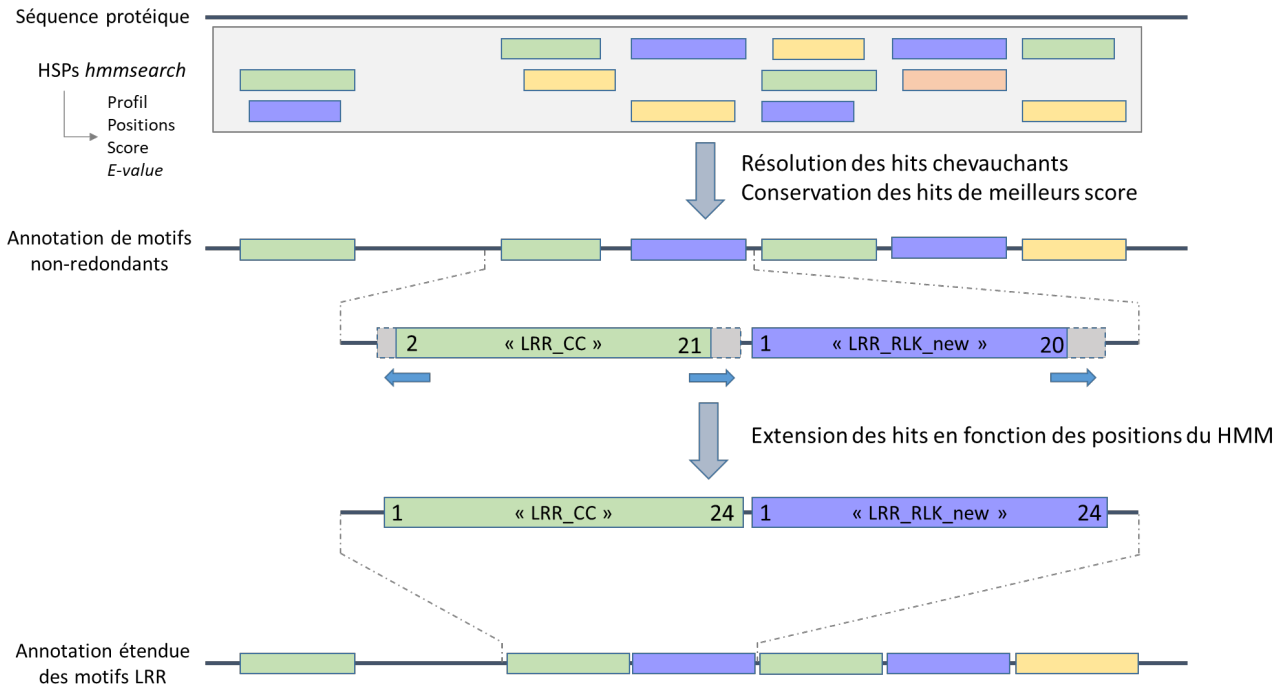


Figure 23 : Procédure de filtrage et de correction des prédictions de motifs LRR par HMM.

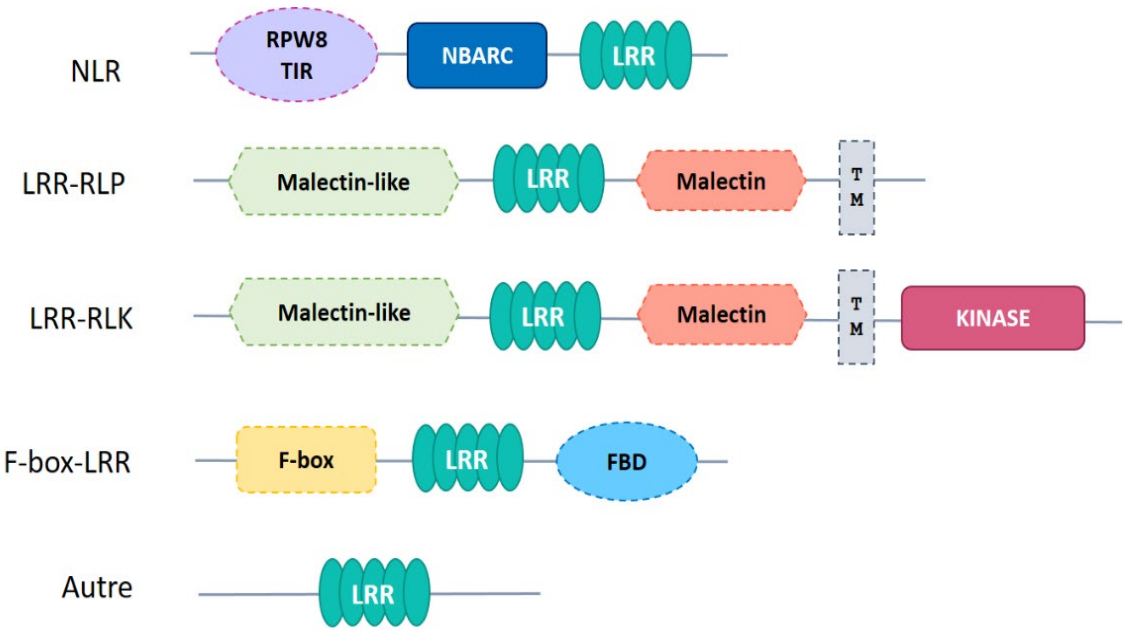


Figure 24 : Représentation des domaines protéiques pour la classification en famille.
Les domaines avec tour pointillés sont des domaines dont la prédiction n'est pas obligatoire pour la classification de la protéine dans la sous-famille correspondante.

C.3.2. Validation du pipeline

Le pipeline LRRprofiler a été testé dans un premier temps sur le jeu de 16 031 protéines d'*A. thaliana* issues de Swiss-Prot. Les résultats obtenus ont été comparés aux annotations expertisées de cette base de données. Dans un second temps, LRRprofiler a été également évalué sur le protéome plus large de TAIR 10 (27 416 protéines) et comparé à un outil alternatif publié très récemment (Martin et al., 2020).

Comparaison des résultats obtenus sur un protéome de Swiss-Prot dont les annotations sont expertisées

Le pipeline LRRprofiler a permis d'identifier 100% des protéines attendues pour les trois familles d'intérêt du protéome de Swiss-Prot (LRR-RLK, LRR-RLP et NLR). Le tableau 7 présente les mesures de rappel (ou sensibilité) et de précision pour la classification des protéines en sous-famille. Le rappel mesure la proportion de protéines correctement classées parmi les protéines attendues :

$$rappel = \frac{vrais\ positifs}{vrais\ positifs + faux\ négatifs}$$

Et la précision mesure la proportion de protéines correctement classées parmi l'ensemble des protéines identifiées :

$$précision = \frac{vrais\ positifs}{vrais\ positifs + faux\ positifs}$$

Les résultats de LRRprofiler sont comparés à l'annotation fournie par Swiss-Prot. Concernant les protéines LRR-RLK et NLR, le pipeline a identifié, et correctement classé, 100% des séquences attendues. Les mesures de sensibilité et précision sont de 100% pour ces deux sous-familles. Concernant les LRR-RLP, 100% des protéines ont été identifiées et 58 des 61 protéines ont été correctement attribuées à la sous-famille LRR-RLP. Les trois autres séquences n'ont pas pu être attribuée à une famille et ont été classées dans le groupe « autre » portant le rappel à 95%, et la précision à 100%. Pour ces 3 séquences classées « autres », TMHMM n'a pas identifié de domaine TM et le nombre de LRR identifiés est inférieur à 13.

Tableau 7 : Performance du pipeline LRRprofiler pour la classification des protéines LRR d'A. thaliana issues du jeu de données de Swiss-Prot.
Les performances sont mesurées pour les trois familles d'intérêt : LRR-RLK, LRR-RLP et NLR. VP = vrais positifs; VN = vrais négatifs ; FP = faux positifs ; FN = faux négatifs.

	LRR-RLK	LRR-RLP	NLR	Total 3 familles
Rappel (VP/VP+FN)	100 %	95 %	100 %	99 %
Précision (VP/VP+FP)	100 %	100 %	100 %	100 %

Comparaison des résultats obtenus sur TAIR10 avec ce qui est connu dans la littérature

Le pipeline LRRprofiler a aussi été testé sur une autre version (assemblage et annotation) du protéome complet d'*A. thaliana* : « TAIR10 ». Ce protéome est plus complet que celui de Swiss-Prot (27 416 protéines vs 16 031) mais il ne fournit pas d'annotation expertisée des gènes LRR. Pour ce protéome, nous avons donc comparé les résultats de notre annotation des gènes LRR à ce qui est connu dans la littérature. Le pipeline a identifié 227 LRR-RLK, 58 LRR-RLP, 148 NLR et 221 autres protéines LRR. Le nombre de protéines à LRR par famille est variable dans la littérature allant de 213 à 251 LRR-RLK (Mondragon-Palomino et al., 2002; Wang et al., 2011; Liu et al., 2017; Man et al., 2020), de 48 à 57 LRR-RLP (Fritz-Laylin et al., 2005; Mondragon-Palomino and Gaut, 2005; Wang et al., 2008) et de 147 à 206 NLR (Meyers2003, Wang2011, VanDeWeyer2019, Mondragon2005, Guo2011). L'identification des séquences LRR de TAIR10 par le pipeline LRRprofiler donne donc des résultats en accord avec les informations disponibles dans la littérature pour l'espèce *A. thaliana*.

L'analyse de Meyers *et al*. (2003) est aujourd'hui encore une référence pour l'analyse des protéines NLR d'*Arabidopsis*. Ces auteurs ont identifié 149 protéines NLR portant des répétitions LRR. Parmi ces NLR, quatre sont absents de nos résultats. Trois sont des séquences absentes du jeu de données TAIR10 à notre disposition (AT3G25515 ; AT4G14610 ; AT5G40920), et la quatrième séquence ne présente aucun motifs LRR (AT3G15700). Par ailleurs, le pipeline LRRprofiler identifie trois séquences NLR absentes de l'étude de Meyer et al. (AT1G57650 ; AT5G45500 ; AT5G45520). Ces séquences pouvaient être absentes des banques de données disponibles au moment de leur étude. Ces résultats montrent que le pipeline permet une identification et une classification automatique des séquences protéiques NLR aussi fiable qu'un travail manuel spécifique minutieux.

Comparaison de LRRprofiler et de LRRpredictor sur un échantillon de protéines de Swiss-Prot

LRRpredictor (Martin et al., 2020) est un outil développé spécifiquement pour la détection de motifs LRR irréguliers au sein des protéines NLR. La détection fonctionne néanmoins pour l'ensemble des trois familles d'intérêt. Les capacités de détection des motifs LRR de LRRprofiler ont été comparées à celles de cet outil. L'outil a été testé, grâce à une machine virtuelle (VM) mise à disposition par Biosphère (https://biosphere.france-bioinformatique.fr/), sur une sélection de 104 séquences du jeu de données Swiss-Prot précédent. Les 104 protéines sélectionnées correspondent aux 53 LRR-RLK, 18 LRR-RLP et 33 NLR appartenant au chromosome 1 d'*A. thaliana*. Nous n'avons testé que le chromosome 1 car nous sommes temporellement limités par l'outil LRRpredictor qui présente une exécution protéine par protéine d'environ 10 à 15 minutes chacune.

En moyenne, LRRpredictor détecte 1,39 motifs LRR de plus par protéines que LRRprofiler. L'écart entre les performances des deux pipelines varie entre les familles considérées. Pour les protéines membranaires LRR-RLK et LRR-RLP, les performances des deux pipelines sont très similaires : LRRpredictor détecte 0,57 motifs LRR de plus en moyenne. Pour les protéines NLR, la différence est plus importante LRRpredictor détecte en moyenne 3,15 motifs LRR de plus par protéine que LRRprofiler.

D. Discussion et perspectives

Notre objectif dans cette partie était d'être capable d'annoter avec une grande précision les répétitions LRR au sein des séquences protéiques. L'utilisation de profils HMM pour annoter des domaines et motifs protéiques a plusieurs avantages : c'est une stratégie facile, rapide, demandant peu de ressources informatiques et facilement automatisable.

Les Profils HMM publics seuls ne permettent pas une annotation complète des domaines LRR des trois familles d'intérêt

Nous avons testé sept profils disponibles publiquement dans les bases de données Pfam et SMART. Les tests ont été réalisés sur un jeu de 16 031 protéines d'*A. thaliana* issues de la base de données Swiss-Prot. Les séquences de cette base sont des données de haute qualité, dont les annotations font l'objet d'une expertise manuelle et de mises à jour régulières, fréquemment utilisées comme référence pour des comparaisons ou tests d'outils. Nous nous sommes intéressés à deux facteurs : (i) la capacité d'un profil à identifier une protéine portant

des LRR (un profil « identifie » une protéine comme LRR si l'on observe au moins un *hit* du profil dans la séquence), et (ii) la capacité d'un profil à identifier chaque motif LRR d'une protéine.

En s'intéressant aux consensus des profils HMM LRR issus de Pfam et SMART, nous avons noté des incohérences dans les noms donnés aux profils SMART (Tableau 2, Tableau 4). En particulier :

- le profil « LRR » présente un consensus de motif *Typical LRR*,
- le profil « LRR_TYP » présente un consensus *RI-Like LRR*,
- le profil « LRR_RI » présente un consensus *Cysteine-containing LRR*,
- le profil « LRR_CC » présente un consensus *Plant Specific LRR*,
- les profils « LRR_SD22 » et « LRR_BAC » présentent un alignement identique et de consensus *Bacterial LRR*.

Ces incohérences ont été signalées aux responsables de la plateforme sans qu'une solution puisse être mise en œuvre dans le délai de la thèse. Pour respecter la notation actuelle des données, et faciliter la reproductibilité, les noms des profils n'ont pas été modifiés. Nous retiendrons en particulier que le profil SMART LRR_CC présente le consensus *plant specific* de la littérature.

Les protéines LRR-RLK et LRR-RLP partagent des motifs LRR de même consensus

Nous avons remarqué que les profils testés présentaient une spécificité pour les motifs LRR soit des gènes membranaires (LRR-RLK et LRR-RLP), soit des gènes intracellulaires (NLR). Le fait que les profils HMM sensibles pour les LRR-RLK soient également sensibles pour les LRR-RLP laisse penser que les histoires évolutives des domaine LRR de ces deux familles sont intimement liées. Les alignements de motifs LRR issus de ces deux familles montrent un même consensus conservé dit *plant specific* (Tableau 2, Figure 16). L'origine des gènes LRR-RLP, possiblement dérivés de troncations de gènes LRR-RLK est possible mais difficile à établir. Certaines séquences LRR-RLP montrent des similarités très élevées avec certaines séquences LRR-RLK, avec parfois des signaux de perte du domaine kinase de certains gènes LRR-RLK, les faisant basculer dans la famille des LRR-RLP. En ce sens, Man *et al*. (Man et al., 2020) ont pu identifier 14 LRR-RLP d'*A. thaliana* (sur 57 au total dans leur analyse) comme étant des variants structuraux de gènes LRR-RLK. La forte conservation observée de manière générale sur les

motifs LRR entre LRR-RLK et LRR-RLP, permettant à un seul profil HMM d'être aussi performant pour l'une ou l'autre des sous-familles, pourrait également être le résultat d'une origine ancestrale commune de leur domaine LRR. Si l'existence de gènes LRR-RLP issus de gènes LRR-RLK ayant perdu leur domaine kinase a pu être démontrée, ce n'est pas le cas de tous les LRR-RLP. Parallèlement, l'hypothèse que les LRR-RLK seraient la conséquence d'une fusion de domaines préexistants est aussi très forte (Dievart et al., 2020). Certains LRR-RLP seraient alors ancestraux aux LRR-RLK. Les fusions de domaines protéiques et leur apparition au cours de l'évolution, particulièrement pour les récepteurs kinase, ont été largement étudiés (cf. chapitre 4).

Il est tentant d'expliquer les similarités entre séquences par l'existence d'une séquence ancestrale commune récente. Il faut cependant rester prudent, surtout en présence de duplications, car différents mécanismes au niveau génomique peuvent aboutir à une similarité élevée entre gènes de différentes familles. C'est le cas par exemple de la conversion génique ou des recombinaisons non-homologues qui pourraient permettre l'échange, partiellement ou totalement, des domaines LRR entre gènes. Ces mécanismes sont connus pour être très actifs parmi les gènes LRR du fait de leur caractère répétitif, au niveau génique (motifs répétés) et génomique (gènes dupliqués) (Mondragon-Palomino and Gaut, 2005). Ces échanges de séquences entre différentes copies de gènes compliquent fortement la détermination des relations d'orthologie entre gènes de différents génomes et la reconstitution de l'histoire évolutive de ces sous-familles de gènes.

La reconstruction de profils HMM LRR spécifiques des familles de protéines permet d'améliorer l'annotation des NLR

Afin de pouvoir annoter l'ensemble des protéines d'intérêt avec la même stratégie, nous avons mis en place une procédure de construction de profils HMM LRR spécifique d'un taxon et d'une sous-famille d'intérêt. Lors de cette procédure, nous avons remarqué que les deux à trois premières itérations permettaient une amélioration rapide des capacités de détection des motifs dans la sous-famille considérée. On observe ensuite une alternance entre des phases d'amélioration et des phases de régression de la qualité des résultats de détection. Les profils HMM permettent un équilibre entre rappel (trouver tous les positifs) et précision (ne trouver que de vrais positifs). Le rappel et la précision reposent sur la diversité de séquence représentée par le profil. A chaque itération, la diversité « capturée » par la procédure varie.

Elle augmente lors des phases d'amélioration puis devient trop importante pour conserver une bonne spécificité. L'alignement ne représente plus efficacement la diversité des motifs, et le rappel du profil suivant diminue. Nous arrêtons la procédure après trois phases de dégradation des résultats. Le profil présentant les meilleures capacités d'annotation est extrait comme nouveau profil pour la sous-famille d'intérêt.

La stratégie a été validée sur les protéines d'*A. thaliana* en comparant les résultats du profil LRR_CC avec ceux d'un nouveau profil construit avec des motifs issus des LRR-RLK. Les deux profils, nouveau et LRR_CC, montrent des résultats équivalents pour l'identification et l'annotation des protéines LRR-RLK et les LRR-RLP.

L'application de cette stratégie aux protéines NLR améliore significativement nos capacités d'annotation des motifs LRR dans ces protéines. En effet, le nouveau profil seul permet d'identifier 97,8% des protéines NLR disponibles dans le jeu de données Swiss-Prot et permet d'annoter 905 motifs LRR dans ces gènes, soit 210 motifs de plus qu'attendus d'après l'annotation Swiss-Prot (695 motifs).

Les protéines NLR présentent des motifs LRR qui divergent de ceux des LRR-RLK et LRR-RLP

La reconstruction de profil LRR pour les NLR via la procédure itérative aboutit à un profil de consensus LxxLxxLxLxxCxxLxxLPxxxxx (Figure 21). Ce consensus diffère du consensus *plant specific* observé pour les protéines LRR-RLK et LRR-RLP : LxxLxxLxLxxNxLxGxIPxxLxx (Tableau 2). Un consensus semblable pour les motifs LRR des NLR est décrit par Martin *et al.* (Martin et al., 2020), avec un résidu cystéine conservé en position 12 à la place d'un résidu asparagine. Le même type de consensus était déjà rapporté par Jones et Jones (Jones and Jones, 1997) pour les protéines LRR intracellulaires impliquées dans les mécanismes de résistance. Ce consensus présente des similitudes avec les motifs de type *Cysteine Containing* initialement observé dans la protéine GRR1 de la levure (*Saccharomyces Cerevisiae*) (Flick and Johnston, 1991) (Tableau 2). En 1998, Kajava place les gènes intracellulaires de plante dans cette catégorie de motif *Cysteine Containing* sur la base des observations faites pour la protéine RSP2 de *A. thaliana* (Kajava, 1998). En revanche, contrairement aux motifs de types *Cysteine Containing*, les motifs trouvés dans les séquences NLR ne présentent pas de cystéine en première position et présentent le di-peptide « IP » en position 18-19 du motif comme pour les motifs de type *Plant-Specific*. Matsushima et Kretsinger (Matsushima and Kretsinger, 2016)

ont donné le nom « PS_sub2 » (*Plant specific subprofil 2*) à ce consensus trouvé chez les NLR, pour conserver la référence au consensus « *Plant-Specific* ».

L'origine des motifs LRR au cours de l'évolution est une question qui est au cœur de nombreuses recherches. L'origine commune des différents types de motifs est souvent évoquée (Andrade et al., 2000; Matsushima and Kretsinger, 2016) même si les divergences observées entre les différentes familles tendent à suggérer une origine différente (Kobe and Deisenhofer, 1994; Kajava, 1998) (cf. introduction, section D.4.). Les motifs des différentes familles peuvent avoir une origine commune, mais suffisamment ancienne pour générer une divergence marquée observable. La question se pose en effet du fait de l'occurrence de domaines LRR au sein de familles de gènes analogues entre plantes et animaux, alors que chaque famille expose des motifs de différentes classes. On trouve par exemple des gènes analogues aux LRR-RLK, les *Toll-Like Receptors* (TLR), et aux NLR de plantes, les *NOD-Like Receptor* chez les métazoaires (animaux et insectes). Chacune de ces familles présente des motifs LRR de consensus différents. Les TLRs sont des gènes membranaires présentant un domaine LRR extracellulaire et un domaine TIR intracellulaire. Tout comme les LRR-RLK, ces gènes sont des récepteurs impliqués dans le développement et l'immunité innée (Janeway and Medzhitov, 2002). Les *NOD-Like Receptor* sont des gènes intracellulaires possédant un domaine NACHT central, de structure et fonction analogue au domaine NB-ARC. Le domaine NACHT est suivi en 3' d'un domaine LRR. Ils peuvent présenter différents types de domaines en 5' comme les NLR de plantes. Les similitudes de structure et de fonctionnement observées ont amené les scientifiques à faire un parallèle entre ces gènes (Ausubel, 2005; Zipfel and Felix, 2005; Jones et al., 2016). Cependant, le lien entre ces gènes, qui indiquerait une origine commune des récepteurs, n'est pas établi. Au contraire, les dernières études sur le sujet tendent à infirmer cette hypothèse. Par exemple, si les domaines centraux des NLRs chez les métazoaires et les plantes (NACHT et NB-ARC) semblent être originaires d'une structure ancestrale commune (famille STAND), Yue *et al.* (Yue et al., 2012), tout comme Urbach et Ausubel (Urbach and Ausubel, 2017), concluent que l'association des LRR avec les domaines NATCH d'un côté et NB-ARC de l'autre résultent de deux évènements de fusion de domaines indépendants. Les similitudes observées existeraient donc par convergence évolutive sans origine commune, et cela bien que les différents domaines (domaine LRR et domaine ancestral

à NACHT et NB-ARC) aient préexisté au sein du dernier ancêtre commun entre plantes et animaux.

L'annotation des motifs LRR est complexe et aucune méthode n'est réellement optimale

La question de l'annotation des motifs LRR est très présente dans la communauté, comme en témoigne la publication de plusieurs articles très récents proposant des solutions d'annotations des LRR pour les NLR (LRRpredictor (Martin et al., 2020)), pour les RLK (PhytoLRR-prediction (Chen, 2021)) ou d'identification de gènes de résistances (RRGpredictor (Santana Silva and Micheli, 2020), NLRtracker (Kourelis et al., biorxiv preprint)).

Nous avons testé l'outil LRRpredictor sur une sélection de séquences protéiques de Swiss-Prot et comparé les résultats avec ceux de LRRprofiler. Cet outil présente des résultats intéressants en particulier sur les NLR dans lesquels il identifie en moyenne plus de 3 motifs LRR de plus que LRRprofiler. L'outil LRRpredictor est une application dédiée à la détection des motifs LRR irréguliers fréquemment retrouvés dans les NLR. Il s'appuie sur la prédiction des structures secondaires des protéines et sur des approches de *machine learning* pour prédire la présence des motifs LRR. La partie d'apprentissage s'appuie sur les données d'Uniprot nécessitant de télécharger 50 Go de données protéiques de cette base. L'installation de l'outil a été compliquée (et a finalement échoué sur le cluster Montpelliérain du CIRAD malgré une vingtaine de mails échangés avec les développeurs). De plus, son fonctionnement protéine par protéine et son temps d'exécution moyen de 10 à 15 minutes par séquence sont des freins à son utilisation à large échelle sans parallélisation. L'outil a finalement pu être testé, en utilisant une VM mise à disposition par Biosphère, et constitue néanmoins un atout très utile ponctuellement pour compléter l'annotation des gènes d'intérêt de la sous-famille NLR.

Ces différences de résultats entre les outils utilisant des stratégies différentes montrent les limites des prédictions ne se basant que sur la séquence primaire des protéines. Les motifs non-identifiés par LRRprofiler ont vraisemblablement perdu, au niveau de la séquence primaire, certaines des caractéristiques utilisées par les profils HMM pour identifier une séquence comme étant un LRR. En revanche, ils auraient conservé les caractéristiques de structure secondaire leur permettant d'être identifiés par LRRpredictor. Les prédictions au niveau de la structure primaire s'appuient sur la conservation de la séquence au cours de l'évolution. Or, la fonction des domaines est principalement déterminée par sa structure 3D.

Dans certains cas, la structure 3D peut être maintenue malgré une divergence importante de la séquence primaire. C'est tout l'intérêt des outils intégrant des prédictions sur la base des structures primaire et secondaire comme LRRpredictor.

Par ailleurs, l'utilisation de HMM force les répétitions à être définies selon des bornes précises. La prédiction du nombre de motifs dans une séquence peut être impactée par ces bornes si elles ne correspondent pas aux limites biologiques du motif. La structure 3D est la seule façon d'obtenir de façon fiable et sans erreur l'annotation d'un domaine répété, c'est-à-dire d'obtenir le bon nombre de motifs et leurs bornes (Bella et al., 2008). Néanmoins, d'après Bella *et al.*, les HMM constituent une alternative assez fiable, particulièrement efficace pour identifier des séquences d'intérêt, mais qui pourraient générer des faux négatifs en fonction de la définition du consensus choisi. Ils relèvent également qu'il y a moins d'erreurs lorsque la partie la plus conservée est mise en début de profil comme c'est le cas pour les HMM que nous utilisons.

La prédiction avec les HMM reste néanmoins pertinente. Nous faisons ce choix comme un compromis entre l'exhaustivité de l'annotation d'une part et la rapidité et la facilité d'utilisation de la méthode d'autre part. Les annotations fournies sont exploitables car homogènes même si certains motifs ne sont pas détectés.

Les données « gold standard » restent incomplètes pour ces familles complexes

Quelle que soit la famille, les deux programmes (LRRprofiler et LRRpredictor) trouvent souvent plus de motifs que ceux annotés par Swiss-Prot. Notamment, LRRprofiler identifie plus de motifs que l'annotation Swiss-Prot pour 237 séquences avec en moyenne 1,3 motifs de plus chez les LRR-RLK (110 séquences), 1,5 motifs de plus chez les LRR-RLP (42 séquences) et 5,2 motifs LRR de plus pour les NLR (85 séquences). Ces observations révèlent les difficultés à avoir des annotations complètes pour les séquences LRR, en particulier pour les NLR, même pour des données expertisées et *gold standard*. Ces données de référence, surtout pour une espèce modèle comme *A. thaliana*, sont des appuis pour la caractérisation et la compréhension de l'évolution de ces familles de gènes. Les annotations incomplètes pour ces gènes peuvent alors participer à de mauvaises interprétat et à la propagation d'erreurs dans les données.

E. Conclusion

L'annotation des motifs LRR au sein des séquences protéiques et une étape essentielle pour étudier les mécanismes évolutifs de ce domaine. Nous avons développé une procédure itérative permettant de créer un profil HMM LRR spécifique d'une sous-famille pour un protéome donné. Le gain de performance pour l'annotation des motifs NLR nous a permis de développer un pipeline complet et intégratif permettant d'identifier, d'annoter et de classer en sous-famille *de novo* les protéines LRR d'un protéome donné. LRRprofiler est un outil simple et rapide, mis à disposition sous forme de container *Singularity* permettant à l'utilisateur de s'affranchir des dépendances d'outils. Il reste malheureusement dépendant d'une annotation fiable des modèles de gènes au niveau génomique (structure intron/exon). En effet, en cas de mauvaise annotation d'un gène, les protéines d'intérêt correspondantes peuvent être manquées, ou mal classées, du fait de l'absence des motifs LRR ou des domaines fonctionnels associés. Cette problématique est abordée dans le chapitre 2.

Chapitre 2

Chapitre 2 :
Annotation des récepteurs à LRR chez le riz cultivé Nipponbare

Table des matières

A. Introduction

L'article présenté dans ce chapitre aborde les biais liés aux méthodes d'annotations des gènes appartenant à des familles multigéniques complexes telles que celles des récepteurs à LRR chez le riz, les conséquences de ces biais sur la qualité des données disponibles et les solutions que nous avons conçues.

Cet article est construit selon 3 axes. **Le premier axe** est consacré à l'analyse des annotations disponibles des gènes LRR d'intérêt pour l'espèce *O. sativa* ssp. *japonica* cv. Nipponbare. Il existe trois annotations publiques pour le génome de référence Nipponbare (IRGSP-1.0) (Kawahara et al., 2013). Ces annotations ont été produites par : (i) l'*International Rice Genome Sequencing Project* (annotation identifiée « IRGSP » dans l'article), (ii) la *Michigan State University* dans le cadre du *Rice Genome Annotation Project* (identifiée « MSU » dans l'article) et (iii) le *National Center for Biotechnology Information* (identifiée « NCBI » dans l'article). Ces trois annotations différentes reposent sur des pipelines dont les stratégies d'annotations sont également différentes. Ce contexte particulier du riz nous permet de comparer des résultats d'annotation des gènes LRR issus de stratégies d'annotation différentes, d'analyser les causes des différences et de rechercher d'éventuels biais et erreurs. Les comparaisons des trois annotations pour les gènes LRR montrent de nombreuses incohérences. Le nombre de loci annotés est très variable avec, par exemple, 282 loci NLR annotés par IRGSP contre 418 annotés par MSU. Les modèles de gènes fournis pour un même locus sont également variables. Par exemple, les modèles prédits par IRGSP sont fréquemment plus courts que ceux fournis par les deux autres annotations. Il semble qu'en cas de mutation non-sens, IRGSP tend à fournir des modèles tronqués, s'arrêtant au premier stop précoce rencontré. Les deux autres annotations présentent des modèles de gènes plus long, couvrant la totalité du gène initial, mais en prédisant fréquemment des introns probablement erronés. En réalisant les mêmes analyses sur des facteurs de transcription comme contrôles de nos expériences, nous avons montré que ces variations sont également observées pour les facteurs de transcriptions mais dans une moindre mesure. La complexité de la famille des récepteurs à LRR (fréquence de duplication élevée, nombreuses copies dont beaucoup contiennent des mutations non-sens) semble donc accentuer des biais d'annotations qui sont inhérents aux méthodes employées.

Le deuxième axe de l'article présente une ré-annotation manuelle des récepteurs à LRR appartenant aux trois familles d'intérêt (LRR-RLK, LRR-RLP et NLR) pour le génome de Nipponbare. Cette ré-annotation nous a semblé nécessaire au vu de la qualité des données disponibles pour ces gènes chez Nipponbare. Pour ré-annoter les gènes à LRR, la stratégie que nous avons suivie est d'identifier la totalité des séquences codantes à chaque locus d'intérêt. Pour les gènes affectés par des mutations non-sens, nous avons choisi d'annoter la totalité du locus « ancestral », c'est-à-dire en prenant en compte l'ensemble de la séquence codante qui présente une forte similarité avec des copies de la même famille de part et d'autre de la ou des mutations. Pour identifier les séquences codantes ou anciennement codantes, nous nous sommes appuyés sur la conservation d'homologie avec différents types de données nucléiques et protéiques (RNAseq, EST, protéines…) disponibles dans les bases de données du NCBI pour le riz mais aussi pour d'autres espèces de monocotylédones. Cette stratégie d'annotation fournit des modèles de gènes dont certains portent des codons *stop* précoces, des *frameshifts* ou des mutations ayant fait perdre le codon *start* ou *stop* terminal. Ces modèles particuliers sont identifiés et labellisés « non-canoniques ». La ré-annotation a abouti à l'identification de 1 058 loci LRR (350 LRR-RLK, 147 LRR-RLP, 289 NLR et 58 non-classés). Parmi eux, 328 sont issus d'un modèle de gène public modifié, et 306 sont non-canoniques. Les modèles de gènes modifiés sont majoritairement non-canoniques (83,5 %) confirmant que les gènes présentant des mutations non-sens sont plus facilement touchés par les erreurs d'annotation, et ont plus souvent nécessité une modification manuelle des modèles publics.

Le troisième axe présente le transfert de cette annotation expertisée vers un autre cultivar de riz *japonica* proche de Nipponbare (Kitaake) et les analyses comparatives du répertoire de récepteurs LRR entre ces deux individus. Le choix du cultivar Kitaake s'est fait pour faciliter le développement de la méthode de transfert. Dans leur publication consacrée au séquençage et à l'assemblage *de novo* du génome de KitaakeX (variété de Kitaake avec ajout de deux transgènes Xa21), Jain *et al.* montrent que l'identité globale entre ces deux cultivars au niveau génomique est de plus de 98% (Jain et al., 2019). En revanche, la majorité des variations qu'ils observent concernent des loci associés aux domaines NB-ARC, LRR et Kinase. L'identité génomique globale facilite les comparaisons post-transfert, mais les variations du contenu en gène d'intérêt permettent de développer efficacement des méthodes dédiées qui prennent

en compte les situations complexes (loci spécifiques, paralogues, différents niveaux de divergence...).

L'article a été publié dans la revue *The Plant Journal* en août 2021 (https://doi.org/10.1111/tpj.15456).

B. Article associé : *'A New Comprehensive annotation of Leucine-Rich Repeat-Containing Receptors in rice'*

the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2021) doi: 10.1111/tpj.15456

A new comprehensive annotation of leucine-rich repeat-containing receptors in rice

Céline Gottin[1,2], Anne Dievart[1,2], Marilyne Summo[1,2], Gaëtan Droc[1,2], Christophe Périn[1,2], Vincent Ranwez[1] and Nathalie Chantret[1,*]

[1]*UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France, and*
[2]*CIRAD, UMR AGAP Institut, F-34398 Montpellier, France*

Received 11 February 2021; revised 23 July 2021; accepted 30 July 2021.
*For correspondence (e-mail nathalie.chantret@inrae.fr).

SUMMARY

Oryza sativa (rice) plays an essential food security role for more than half of the world's population. Obtaining crops with high levels of disease resistance is a major challenge for breeders, especially today, given the urgent need for agriculture to be more sustainable. Plant resistance genes are mainly encoded by three large leucine-rich repeat (LRR)-containing receptor (LRR-CR) families: the LRR-receptor-like kinase (LRR-RLK), LRR-receptor-like protein (LRR-RLP) and nucleotide-binding LRR receptor (NLR). Using LRRPROFILER, a pipeline that we developed to annotate and classify these proteins, we compared three publicly available annotations of the rice Nipponbare reference genome. The extended discrepancies that we observed for LRR-CR gene models led us to perform an in-depth manual curation of their annotations while paying special attention to nonsense mutations. We then transferred this manually curated annotation to Kitaake, a cultivar that is closely related to Nipponbare, using an optimized strategy. Here, we discuss the breakthrough achieved by manual curation when comparing genomes and, in addition to 'functional' and 'structural' annotations, we propose that the community adopts this approach, which we call 'comprehensive' annotation. The resulting data are crucial for further studies on the natural variability and evolution of LRR-CR genes in order to promote their use in breeding future resilient varieties.

Keywords: LRR-receptor-like kinase, LRR-receptor-like protein, nucleotide-binding LRR receptor, annotation curation, pseudogenes, *Oryza sativa*, disease resistance gene.

INTRODUCTION

Modern agriculture is at a critical juncture, as the world's population continues to grow but there is a call to shift away from chemical treatments to deal with current environmental issues. Crop pest and pathogen susceptibility is one of the main causes of annual crop yield loss (FAO, 2018; Savary et al., 2019). Despite an awareness of the harmful environmental impact, massive pesticide use remains a common means to prevent plant diseases today. Studying and understanding plant disease resistance and the underlying evolutionary mechanisms are of utmost importance to make effective widespread use of known sources of resistance through specific breeding programs, while also promoting new resistance engineering for crop sustainability (Bailey-Serres et al., 2019; Tamborski and Krasileva, 2020). The elucidation of resistance mechanisms in plants has highlighted a trove of resistance genes to combat the great and evolving genetic diversity of plant pathogens. The leucine-rich repeat (LRR)-containing receptors (LRR-CRs) are at the forefront of these genes. LRR-CRs share the common structural and functional LRR domain. This domain contains between two and 30+ repetitions of an approximately 24-amino-acid motif, characterized by a conserved skeleton composed mostly of leucine residues (Bella et al., 2008; Kajava, 1998; Kajava, 2012; Matsushima and Miyashita, 2012). These LRR-CRs are classified in three main gene families: LRR receptor-like kinase (LRR-RLK, also named LRR-RK but referred to herein as LRR-RLK), LRR receptor-like protein (LRR-RLP) and nucleotide-binding-site LRR (NBS-LRR or NLR) (Han, 2019; Sekhwal et al., 2015) (Figure 1). The LRR-RLKs and LRR-RLPs are transmembrane receptors composed of an extracellular LRR domain and an intracellular domain. The intracellular domain is a kinase domain for LRR-RLK (Shiu and Bleecker, 2001a; Shiu and Bleecker, 2001b) and a short cytoplasmic tail for LRR-RLP (Fritz-Laylin et al., 2005; Jones

and Jones, 1997). Some LRR-RLKs and LRR-RLPs play roles in intercellular communication involved in disease resistance (such as pattern-recognition receptors, PRRs), stress responses or developmental processes (Boutrot and Zipfel, 2017; van der Burgh and Joosten, 2019). Other LRR-RLKs and LRR-RLPs also act as co-receptors or regulators in these signaling pathways (Couto and Zipfel, 2016). NLRs are intracellular receptors composed of a central nucleotide-binding domain (NB-ARC domain) followed by the LRR domain (Burdett et al., 2019; Sekhwal et al., 2015; Tamborski and Krasileva, 2020; Xiong et al., 2020). These proteins can contain other functional domains, such as the toll/interleukin receptor (TIR) domain, the coiled-coil (CC) domain or the resistance to powdery mildew 8 (RPW8) domain, located upstream of the NB-ARC domain.

Over the past few decades, advances in sequencing have provided the research community with an ever-increasing number of complete genomes. These resources have made it possible to revisit gene evolution at the level of entire families and on different evolutionary timescales. LRR-CR genes have been inventoried in many angiosperm genomes, and their numbers have also been compared in a phylogenetic framework to shed light on their evolutionary dynamics (for just some of the more recent articles, see Andersen et al., 2020; Furumizu and Sawa, 2021; Hosseini et al., 2020; Lee et al., 2021; Man et al., 2020; Prigozhin and Krasileva, 2021). A large proportion of LRR-CR genes are thought to evolve through a so called birth-and-death model (McDowell and Simon, 2006; Michelmore and Meyers, 1998; Nei and Rooney, 2005; Richter and Ronald, 2000). In this model, the gene copy number expands by recurrent duplication events and duplicated copies can then follow different evolutionary pathways, such as keeping the original function, acquiring a new function (neofunctionalization) or, more frequently, undergoing a non-functionalization process by accumulating nonsense mutations (Innan and Kondrashov, 2010; Leister, 2004). This model explains why LRR-CR genes are found in multiple copies, often organized in large gene clusters, with some genes no longer being functional (Meyers et al., 2003; Mizuno et al., 2020).

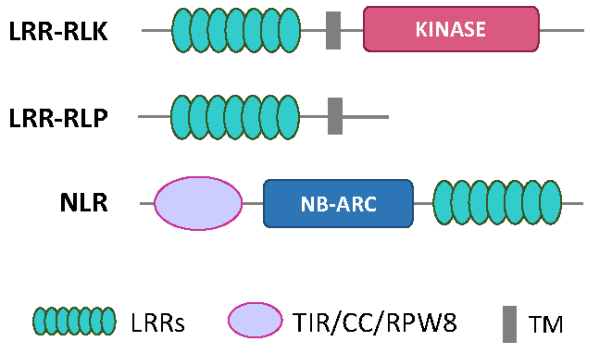


Figure 1. Schematic protein structure of the three LRR-CR subfamilies: LRR-RLK, LRR-RLP and NLR. TM, transmembrane domain; CC, coiled coil domain; TIR, toll-interleukin receptor; RPW8, resistance to powdery mildew 8 domain.

Comparative genomic studies have led to considerable progress in understanding the evolutionary dynamics of LRR-CR gene families, but these studies are highly dependent on the accuracy of annotation procedures. Given the increasing avalanche of sequence data, the most reasonable approach is to rely on automatic annotation. Gene and protein sequence annotation are thus crucial and the target of considerable effort. Structural gene annotation is geared towards identifying coding sequences within genomic data and documenting the associated gene features (e.g. introns, exons and untranslated regions, UTRs) (Wilming and Harrow, 2009). The most widely used structural annotation pipelines, such as the Ensembl pipeline for gene annotation (Aken et al., 2016), Augustus (Stanke and Waack, 2003) and Gnomon (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/), rely on: (i) *ab initio* gene structure determination according to rules learned on pre-existing annotations; and/or (ii) comparative approaches, i.e. using sequence homology with available RNAseq data and/or with a closely related annotated genome. Those methods allow large-scale studies with standardized approaches, yet they are not completely reliable, especially for complex multigene families. Indeed, repetitions are known to impair gene annotations (Bayer et al., 2018; Fawal et al., 2014) and there are also genome assembly issues (Torresen et al., 2019). The difficulty is twofold in the case of LRR-CRs: several similar genes are present in the genome as a result of gene duplication events, whereas each gene contains several similar motifs because of the repetitive structure of the LRR domain. The automatic annotation and classification of LRR-CRs is thus especially challenging. For example, although multiple studies have reported that there are more than 800 LRR-CR loci in the rice variety Nipponbare, the number of genes per family is variable, e.g. 374 498 NLR proteins (Li et al., 2010; Li et al., 2016; Shao et al., 2016; Stein et al., 2018; Zhou et al., 2004), 292 435 LRR-RLKs (Dufayard et al., 2017; Hwang et al., 2011; Man et al., 2020; Sun and Wang, 2011) and 90 LRR-RLPs (Fritz-Laylin et al., 2005). These variations are to a large extent linked to the annotation version chosen for the analysis and to the decision rules for gene detection and classification. Scientists sometimes perform the manual curation of gene annotations to limit these uncertainties and achieve high-quality comprehensive analyses, as in the case of Arabidopsis and *Solanum lycopersicum* (tomato) NLR genes (Jupe et al., 2013; Meyers et al., 2003; Van de Weyer et al., 2019) or *Oryza sativa* (rice) Nipponbare LRR-RLK genes (Sun and Wang, 2011).

Rice was the first monocotyledon plant to have its genome entirely sequenced and three different annotations of

its reference genome, *O. sativa* ssp. *japonica* cv. Nipponbare (Kawahara et al., 2013), are currently available: one from the Michigan State University Rice Genome Annotation Project (MSU, http://rice.uga.edu) (Yuan et al., 2003), one from the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) and the current reference genome from the Rice Annotation Project of the International Rice Genome Sequencing Project (IRGSP, https://rapdb.dna.affrc.go.jp) (Sakai et al., 2013). We first implemented the LRRPROFILER pipeline to compare them with regards to the LRR-CR protein repertoire. This program builds subfamily- and genome-specific LRR hidden Markov model (HMM) profiles, detects LRR-CR proteins that contain LRR motifs and accurately locates LRR motifs within these proteins. We ran the LRRPROFILER pipeline in parallel on the three rice predicted proteomes and found that they greatly differed in terms of the number of LRR-CR genes and their structural annotations. We therefore performed a manual curation of the whole Nipponbare LRR-CR repertoire annotation. To do so, for gene models that diverged between the three annotations, we looked for the reasons of divergence and decided, when appropriate, to supplement the gene models with sequence fragments undoubtedly derived from LRR-CR-encoding genes. In turn, we provided objective information, i.e. whether the gene models were canonical or non-canonical. To be qualified as canonical a gene model had to fulfil all of these conditions: presence of a start codon; presence of a terminal stop codon; absence of an in-frame stop codon; absence of frameshifts; and absence of unexpected intron splicing sites. Conversely, any gene violating at least one of these constraints was qualified as non-canonical. Finally, we also propose a strategy to transfer these manually curated LRR-CR gene annotations to Kitaake, the closest related japonica genome that has been sequenced (Jain et al., 2019). We then analyzed the observed variations in gene numbers and LRR motifs between Nipponbare and Kitaake genotypes while using the available automatic annotations and our manually curated annotations (hereafter referred to as 'comprehensive'). This comparison demonstrated how erroneous conclusions can readily be drawn when relying solely on automatic structural and functional annotations for this complex gene family. The curated comprehensive LRR-CR annotation introduced in this article is available online through a dedicated website (https://rice-genome-hub.southgreen.fr/content/geloc).

RESULTS

Inconsistencies among three publicly available Nipponbare rice LRR-CR annotations

We used LRRPROFILER, a newly developed pipeline (see Experimental procedures and Data S1, Methods S1, Figures S1 and S2 and Table S1 for LRRPROFILER validation results, performed on a manually reviewed *Arabidopsis thaliana* protein data set and on the whole Arabidopsis proteome, including a comparison with the LRRPREDICTOR tool; Martin et al., 2020), to identify, annotate and classify into gene subfamilies the LRR-CR protein sequences of the three publicly available Nipponbare proteomes (MSU, IRGSP and NCBI). The total number of LRR-CRs identified varied markedly according to the annotation: we identified 1226 LRR-containing sequences in the MSU predicted proteome, 1047 in that of IRGSP and 1073 in that of NCBI (Table 1). The distribution patterns of these proteins in the different subfamilies also varied according to the annotations. For instance, the number of predicted genes fluctuated less for the LRR-RLP subfamily than for the NLR subfamily, for which 60% more NLRs were detected in the MSU proteome (418 proteins) compared with the IRGSP proteome (282 proteins). For comparison, we conducted a similar analysis on nine transcription factor (TF) subfamilies, for which we assumed that the annotation process would be easier as they had a more conserved structure and, although having undergone expansion events, were not evolving under a birth-and-death model (Lai et al., 2020). The TF data set contained between 874 and 1041 genes, according to the annotations, and this number was similar to that of LRR-CR. To assess whether the identified genes were at the same genomic location or not, we measured the overlap of the three predicted gene sets. The percentage of loci for which a gene model was present in all three annotations was 52.3% for LRR-CR genes and 69.5% for TF (Figure 2a), indicating that the three annotations were more congruent for TF genes. Moreover, the percentage of loci in which only one annotation detected a gene was 19.5% for LRR-CR genes, compared with only 12% for TF genes.

Even when a gene was predicted by the different annotations, the predicted structure of the gene sometimes varied between predictions. One way to address this issue is to compare the length of the predicted proteins for genes positioned at the same locus. Note that this is a conservative approach. Indeed, although a predicted protein length difference between two gene models indicated that the gene models differed, the reverse was not true, as identical

Table 1 Number of LRR-CR sequences in the predicted proteomes from three publicly available annotations for the Nipponbare rice reference genome. Sequences were identified and classified into subfamilies using the LRRPROFILER pipeline

	Total	LRR-RLKs	LRR-RLPs	NLRs	Others[a]
IRGSP	1047	237 (22.6%)	160 (15.3%)	282 (26.9%)	368 (35.1%)
MSU	1226	329 (26.8%)	141 (11.5%)	418 (34.1%)	338 (27.6%)
NCBI	1073	305 (28.4%)	121 (11.3%)	361 (33.6%)	286 (26.7%)

[a]F-box-LRR and unclassified (UC) sequences.

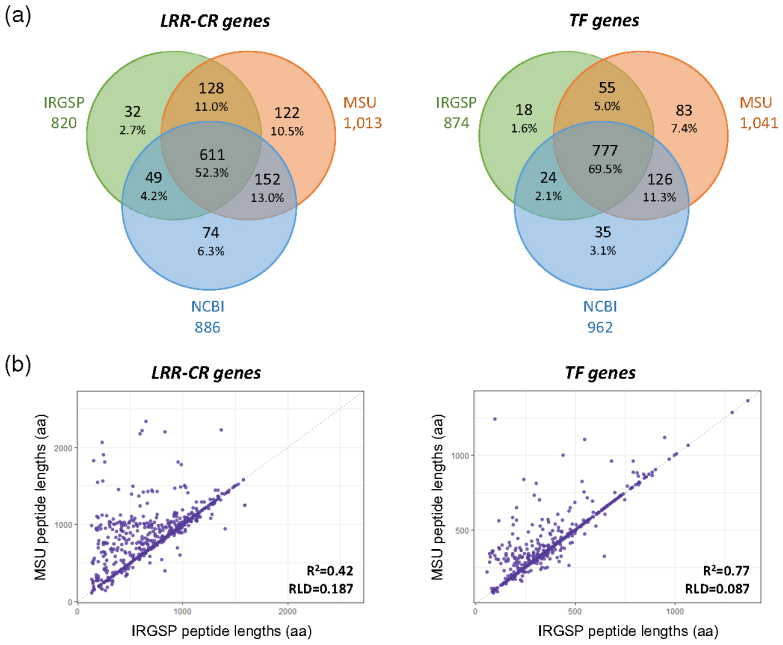


Figure 2. Comparison of publicly available MSU, IRGSP and NCBI annotations for the Nipponbare rice reference genome for two types of genes: LRR-containing receptors (LRR-CRs) and transcription factors (TFs).
(a) Venn diagrams representing the number of overlapping gene models for LRR-CRs and TFs among the MSU, IRGSP and NCBI annotations. To be considered as overlapping, gene models from two (or three) different annotations should have at least one nucleotide in common (overlapping loci). The total number of genes in each annotation does not correspond to the total number in Table 1 because of the complex relationships between loci: for instance, a single NCBI gene can overlap with a gene in IRGSP and another in MSU, whereas these IRGSP and MSU genes do not overlap.
(b) Dot plots representing the polypeptide length in amino acids (aa) for genes predicted by both IRGSP and MSU annotations. On the left, LRR-CRs and on the right, TFs. The R^2 and the average relative length difference (*RLD*) values are given at the bottom right for each gene family. For all pairwise comparisons among IRGSP, MSU and NCBI, see Figure S3.

predicted protein lengths did not guarantee that the gene models were identical. A comparison of predicted protein lengths for all LRR-CR gene pairs located at the same locus but predicted by two different annotations is presented in Figure 2b and Figure S3. Here, again, the number of genes with a difference in the predicted protein length highlighted a substantial annotation discrepancy. This difference was greater for LRR-CR genes than for TF genes. As an example, when IRGSP and MSU were compared, the average difference between the predicted product sizes was 18.7% for LRR-CR loci (with an R^2 of only 0.42), whereas this difference was only 8.6% for TF loci (with a much higher R^2 of 0.77) (Figure 2b). These results highlight the extent to which annotations generally differ, but more particularly for LRR-CR gene subfamilies. These comparisons also showed that LRR-CR genes predicted by IRGSP were generally shorter than those predicted by MSU or NCBI at the same locus (Figure 2b and Figure S3).

Manual re-annotation of LRR-CR-encoding loci in the Nipponbare rice genome

Here we provide a brief description of the procedure that we followed to manually curate LRR-CR annotations (Figure 3a). First, note that for the sake of traceability the procedure retained one of the three proposed gene models as much as possible. For a given locus, we first selected one of the gene models among the available annotations based on the completeness of the predicted protein. We then applied our expertise to the selected gene model by combining protein and nucleotide data. At the protein level, we checked that all of the expected domains for each subfamily were present (e.g. LRRs and TM for LRR-RLPs and LRR-RLKs, kinase for LRR-RLKs or NB-ARC for NLRs) in the right order, with the expected length and interdomain intervals. Protein domain information was particularly useful for detecting potential gene fusion and fission. At the nucleotide level, we examined: (i) whether the gene models had the expected intron/exon structure (e.g. introns, when present, are often found at the same exact position); (ii) whether nearby open reading frames (ORFs) belonging to LRR-CR-encoding sequences were present; and (iii) whether the gene models included suspicious introns, such as short introns, enabling the gene to sidestep stop codons or frameshifts, especially when they were never found in homologs (Figure 3b). Any structural annotation containing an in-frame stop codon or a frameshift (i.e. any gap in coding sequence that was not an intron but that changed the translation phase), lacking a start codon or a terminal stop codon, or presenting an unexpected splicing site (different from the GT-AG and GC-AG donor/acceptor canonical splicing sites) was called 'non-canonical'. This careful inspection was facilitated by viewing the sequence annotations with the ARTEMIS editor (Carver et al., 2012).

In a last step, we also looked for LRR-containing sequences that would have been missed by the three publicly available annotations. The Nipponbare reference genome was split into 1-kb segments with overlapping 100-bp borders, translated into amino acid sequences in the six reading frames (as performed by Steuernagel et al., 2020), and domains (LRR, kinase, NB-ARC, etc.) were searched

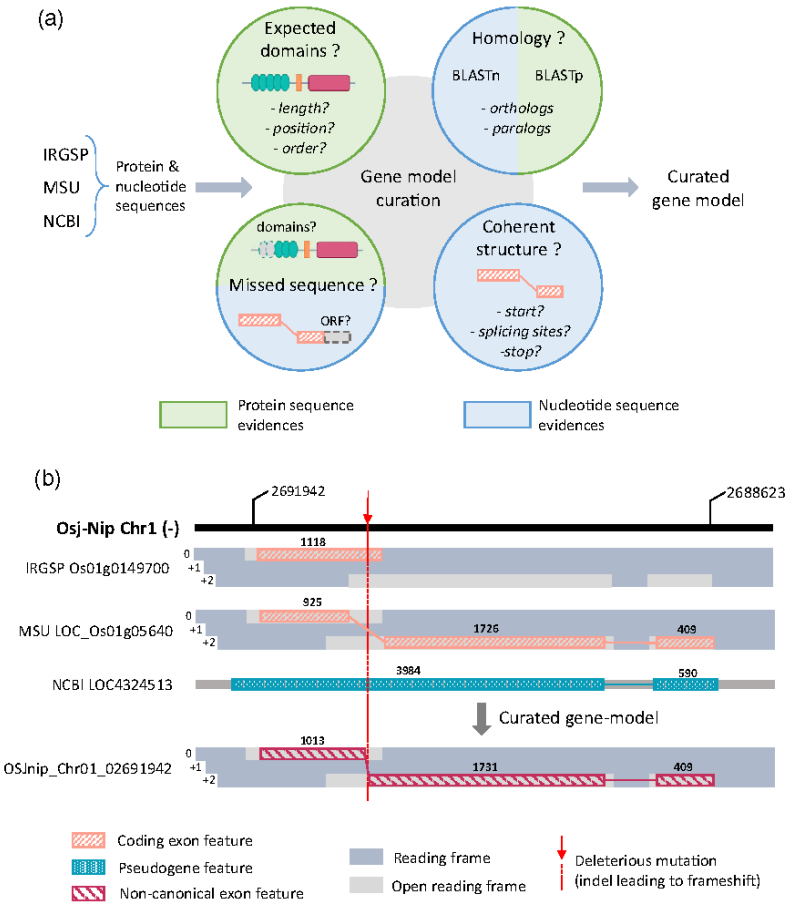


Figure 3. Manual curation of LRR-CR gene model strategy and example of annotation inconsistencies.
(a) Schematic representation of the strategy used to curate Nipponabre LRR-CR gene models. An initial gene model was selected from the three public annotations. This gene model was gradually modified based on protein and nucleotide sequence evidence. The curated model was then classified as canonical or non-canonical.
(b) Schematic representation of an example of inconsistency between gene models from publicly available annotations and how the curation was performed. The gene is an LRR-RLK located on chromosome 1 of the Nipponbare genome. The numbers above the boxes indicate the length of the feature. In this example, an indel mutation caused a frameshift in the first exon of the gene. The IRGSP annotation retrieved the first part of the coding sequence, stopping at the first stop codon on frame 0. The MSU annotation retrieved a longer coding sequence but sidestepped the indel mutation by introducing a 'dubious' intron in order to reach the open reading frame (ORF) on the +2 frame. This 'dubious' intron was abnormally short and contained a sequence highly homologous to the coding sequence in other paralogous gene copies. The NCBI annotation gave a pseudogene feature, i.e. a feature from which a protein sequence could not be deduced: the cDNA sequence is available but would not allow protein translation as it would be in the wrong reading frame after the mutation. The curation took advantage of the three annotations. It retried a cDNA sequence that overlapped the complete former coding sequence in the two successive correct reading frames via the identification of the indel mutation. The identification of the indel mutation was clear cut as the gene was tagged as 'non-canonical' with the presence of a frameshift, but it allowed a complete protein sequence to be deduced and used for sequence comparison and alignment.

with HMMSEARCH. All the results were concatenated and filtered for redundancies. The retained new sequences of interest had to contain at least three LRR motifs in tandem. If another domain was detected at less than 5 kb from these LRR motifs, the sequence of interest was enlarged to also include these domains. These sequences, not overlapping known LRR-CR exons, were compared with other plant genomes using BLAST to screen for the potential presence of a gene model in the region under consideration.

The final set of manually validated LRR-CR loci on the Nipponbare genome consisted of 1058 genes (350 LRR-RLKs, 147 LRR-RLPs, 503 NLRs and 58 UCs) (Data S2; Table 2). Among these 1058 genes, eight (one LRR-RLK, three LRR-RLP and four UC) were located at loci for which none of the three publicly available annotations detected a gene. The LRR-RLK was a canonical full-length sequence on the forward strand of chromosome 2 (from 6 831 702 to 6 834 761). Note that this sequence is actually present in GenBank under accession number EAZ22278.1 and is located on the reverse strand in a non-coding region of the *Os02g0222500* gene. The other seven are non-canonical truncated genes. In addition, for seven of these 1058

The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2021), doi: 10.1111/tpj.15456

Table 2 Number of LRR-CR proteins in the predicted proteomes from our curated annotations for the Nipponbare rice reference genome. Sequences were identified and classified into subfamilies using the LRRPROFILER pipeline

	Total	LRR-RLK	LRR-RLP	NLR	UC
LRR-CR loci[a]	1058 (8)	350 (1)	147 (3)	503 (0)	58 (4)
Modified loci (%[b])	328 (31.0%)	56 (16%)	55 (37.4%)	197 (39.2%)	20 (34.5%)
Non-canonical loci (%)	306 (28.9%)	53 (15.1%)	48 (32.7%)	183 (36.4%)	22 (37.9%)
Modified and non-canonical (%)	274 (25.9%)	43 (12.3%)	43 (29.3%)	170 (33.8%)	18 (31.0%)

[a]Numbers in parentheses are newly identified LRR-CR genes.
[b]Percentages were calculated based on the number of manually curated genes, i.e. the total number of genes minus the number of newly identified genes.

validated genes, the LRRPROFILER pipeline did not detect any further LRR motifs in the predicted protein. LRR motifs were initially detected for these genes, but at the threshold limit when using HMM profiles built on the basis of the initial data set (for details, see the LRRPROFILER pipeline section in the Experimental procedures). When using the slightly different HMM profiles obtained with the final data set, the same LRR motifs were no longer detected as they did not surpass the threshold. However, a careful manual inspection showed that the LRR domain was present but contained divergent LRR motifs, thereby complicating the automatic detection. Consequently, these genes were kept and classified according to the presence of the other domains (kinase or NB-ARC). These seven genes included one LRR-RLK and six NLRs.

Among these 1058 LRR-CR genes, 328 (197 NLR, 56 LRR-RLK, 55 LRR-RLP and 20 UC) were manually modified because none of the three publicly available annotations had a satisfactory gene model based on the previously defined criteria (Data S2; Figure 3a). The overall proportion of modified loci was 31.0% (328/1058), and varied markedly according to the gene subfamily considered. Only 16% of LRR-RLK loci were modified, whereas 37.4% of the LRR-RLP loci and 39.2% of the NLR loci were modified (Table 2). Among these 1058 LRR-CR genes, 306 (28.9%) were non-canonical. Again, the different gene subfamilies did not contain the same proportion of non-canonical gene models. Very similar to what was observed regarding the proportion of modified gene models according to gene subfamily, non-canonical gene models concerned only 15.1% of the LRR-RLKs, compared with 32.7 and 36.4% of the LRR-RLPs and NLRs, respectively. Thus, 274 genes were both non-canonical and modified, representing 83.5% of the total modified loci (274 over 328) and 89.5% of the non-canonical loci (274 over 306) (Table S2). The remaining 32 non-canonical genes were either unreported by any of the annotations (seven) or were reported by the NCBI as pseudogene or gene models having putative errors in the genomic sequence (25, see below).

One way to assess the relevance of our expert LRR-CR annotation is to compare the number of functional domains (TMs, NB-ARCs, kinases and LRRs) found in LRR-CR proteins derived from the reference annotations to the number of functional domains found in the proteins derived from our expert annotation (Figure S4). These comparisons revealed that quite a few more LRR-CR domains were found in our manual annotation as compared with the publicly available annotations. For example, when compared with the reference IRGSP annotation, our expert annotation highlighted 29% more TM, 42% more NB-ARC, 33% more kinase and 20% more LRR motifs.

Annotation of the LRR-CR genes in the rice cultivar Kitaake

Kitaake is another *O. sativa* ssp. *japonica* variety for which a complete genomic sequence is available (Jain et al., 2019). In order to compare the LRR-CR repertoire between Nipponbare and Kitaake and limit the need for manual curation in the re-annotation of this closely related rice cultivar, we developed a strategy to transfer our expert annotations from the Nipponbare to the Kitaake genome.

The strategy summarized in Figure 4 starts by identifying Kitaake genome regions that are homologous to Nipponbare LRR-CR sequences. Then it successively takes into account three levels of annotation transfer, depending mostly on the level of sequence identity of each considered region with the LRR-CR gene that identified it. At each locus, our strategy strives to retrieve the most probable gene model with the idea that, if possible, it should be canonical. At the end of the process, LRR-CR gene models that are found to be non-canonical or having a dubious protein structure in Kitaake are manually checked and corrected if needed. At this step, the transfer allowed us to identify 1046 LRR-CR genes in the Kitaake genome.

As carried out for Nipponbare, the Kitaake genome was finally scanned with LRR HMM profiles using HMMSEARCH for new LRR-CR identifications. This procedure allowed us to annotate 18 additional genes, thereby leading to a total of 1064 LRR-CR genes in the Kitaake genome.

The LRRPROFILER pipeline was used on the 1064 predicted Kitaake proteins and allowed us to detect LRR in 1053 of them; 999 were further classified into a LRR-CR subfamily and 54 remained in the UC group. The automatic detection of LRR failed for 11 genes. As carried out for Nipponbare,

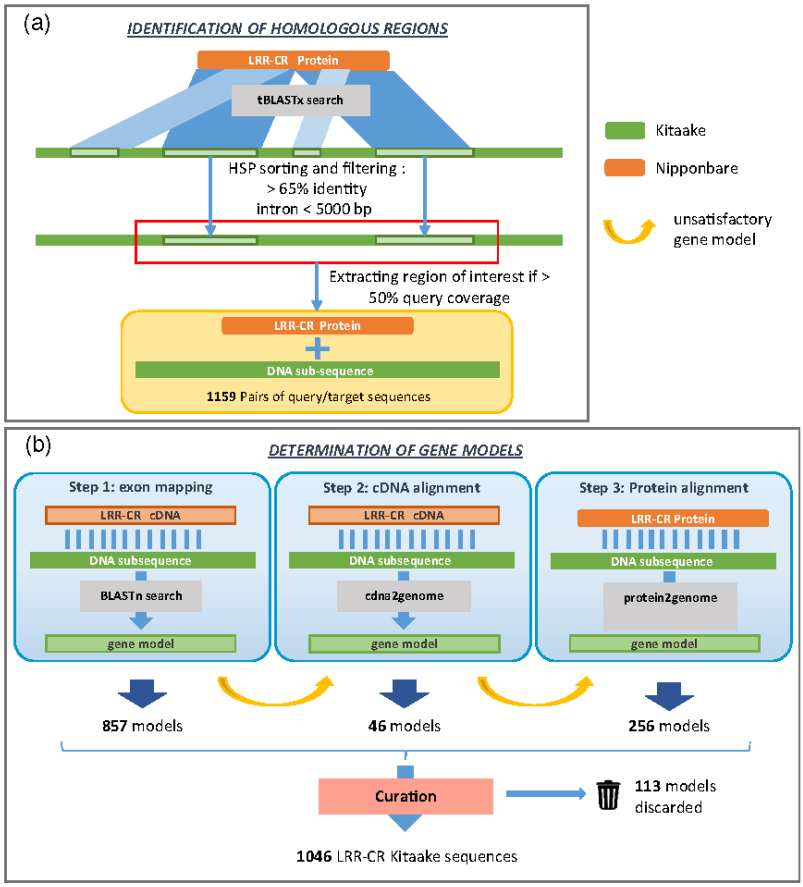


Figure 4. Schematic representation of the annotation transfer strategy between closely related genomes.
(a) Identification of LRR-CR homologous regions. Nipponbare LRR-CR proteins were used to search regions of interest in the Kitaake genome using tBLASTx. BLAST hits with over 65% identity were ranked in the LRR-CR query protein sequence order and used to define the region boundaries. If the filtered BLAST hits within a Kitaake region covered more than 50% of the query LRR-CR sequence, then the Kitaake sequence of this region was extracted and linked to the Nipponbare LRR-CR query protein.
(b) Determination of gene models. The process strives to give a gene model for each region of interest identified in the Kitaake genome. The annotation is attempted in three consecutive steps. If the model from one step is unsatisfactory, i.e. gives an alignment of poor quality with the Nipponbare query protein, the process goes to the next step for this region. At the end of the third step, gene models that remained unsatisfactory were manually checked. This process allowed us to annotate 1046 genes in the Kitaake genome.

at this step, manual validation of the protein annotations confirmed the presence of an LRR domain of the expected size and located at the expected positions. These 11 genes were therefore kept in the final data set. The gene subfamilies for these 11 loci were determined based on other functional domains and a homology search against other LRR-CR protein sequences. Finally, the LRR-CR gene set from Kitaake was composed of 360 LRR-RLKs, 140 LRR-RLPs, 510 NLRs and 54 UCs (Data S3; Figure 5). These numbers were very similar to those obtained for Nipponbare, i.e. 350 LRR-RLKs, 147 LRR-RLPs, 503 NLRs and 58 UCs.

We then tagged all of these Kitaake LRR-CR gene models as either canonical or non-canonical. We obtained 742 (69.7%) canonical genes and 322 (30.3%) non-canonical genes. Again, the proportions of canonical and non-canonical genes per subfamily for Kitaake were very similar to those obtained for Nipponbare (Figure 5).

A notable result is that our strategy enabled us to identify 114 LRR-CR genes (48 of which were canonical) that were not present in the publicly available annotation of the Kitaake genome: 17 LRR-RLKs, 24 LRR-RLPs, 50 NLRs and 23 UCs.

All LRR-CR loci annotation and sequence data for the Nipponbare and Kitaake genomes can be viewed and downloaded on the dedicated website (https://rice-genome-hub.southgreen.fr/content/geloc).

Comparison of LRR-CR allelic pairs between Nipponbare and Kitaake

Nipponbare and Kitaake are two varieties of the same subspecies: *O. sativa* ssp. *japonica*. As such, for the majority of the genes found in Nipponbare, an allele (i.e. a version of the same gene located at the same chromosomal location) was expected to be found in Kitaake. By using SYNMAP (Lyons and Freeling, 2008), we identified 1002 allelic pairs (representing 90.5% of the total number of loci) between Nipponbare and Kitaake (Data S4). In addition, we noticed that for three NLR gene pairs located close to each other on chromosome 9 in the Nipponbare genome, three consecutive genes on chromosome 3 of the Kitaake genome were found with 100% identity with regards to their predicted coding sequence. The intergenic sequences of these two regions also had a high level of identity (99% over 40.5 kb), suggesting that these three genes are located in a translocated region of the genome.

First, to assess the impact of re-annotations on the number of LRR motifs in the alleles, the number of LRR motifs

The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2021), doi: 10.1111/tpj.15456

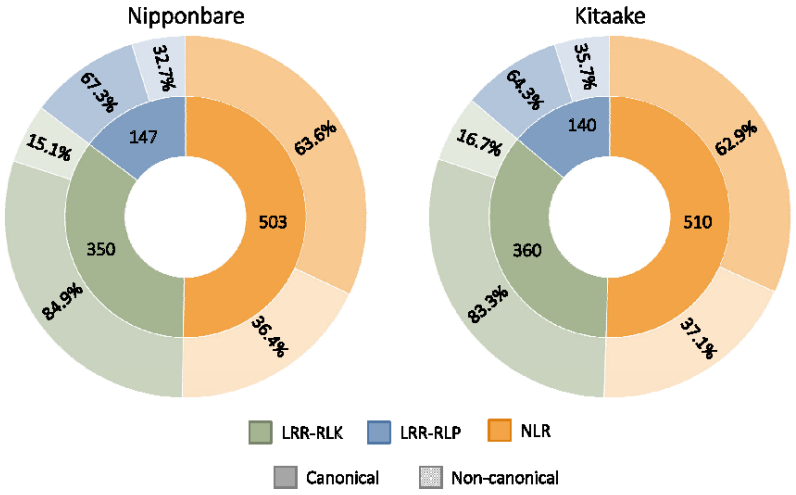


Figure 5. Proportion of canonical and non-canonical loci per gene subfamily in our Nipponbare and Kitaake expert annotations. Percentages were calculated per gene subfamily. The inner circle provides the number of loci per family, with a different color for each. The outer circle shows a lighter/darker version of the loci family color to represent the fraction of the non-canonical/canonical members, respectively, within this gene family.

predicted in Nipponbare was compared with the number of LRR motifs predicted in Kitaake for each pair of allelic sequences. To obtain a precise annotation of the LRR motifs in each protein, we used the LRRPROFILER pipeline. The same procedure was also applied on allelic pairs identified between the publicly available annotations of Nipponbare and Kitaake. We observed a mean difference in LRR number per protein of 3.58 when comparing the publicly available annotations (IRGSP for Nipponbare and the only one that exists for Kitaake) (Figure 6 and Figure S5). This difference fell to 0.6 when our re-annotated data were compared. Using our curated annotations hence led to LRR number predictions that were much more consistent between Nipponbare and Kitaake alleles, and this trend was observed for all LRR-CR gene subfamilies. Moreover, the mean difference in LRR number still varied between LRR-CR gene subfamilies, with greater conservation of LRR motif numbers between LRR-RLK and LRR-RLP alleles than between NLR alleles.

Second, we analyzed the re-annotated allelic pairs related to their canonical or non-canonical status. Among the 1005 pairs (1002 allelic plus three translocated pairs), 688 (68.5%) were pairs of canonical gene models, 269 (26.8%) were pairs of non-canonical gene models and 48 (4.8%) were pairs of genes found to be canonical in only one of the two cultivars. Interestingly, 83.1% of the LRR-RLK pairs were canonical in both cultivars, compared with only 63.9% of the LRR-RLP pairs and 60.3% of the NLR pairs (Table 3).

To go further into this comparison, for each of the 1005 pairs of LRR-CR alleles, the fraction of exact matches along the cDNA pairwise global alignment (i.e. their percentage of identity) was computed. This cDNA identity was about 98.6% on average. The highest identity rate (99.3%) was obtained for alleles belonging to the LRR-RLK subfamily, followed by the NLR (98.2%) and LRR-RLP (97.9%) subfamilies. On average, non-canonical conserved gene pairs (NC/NC category in Table 3) had a lower identity level (97.9%) than conserved canonical gene pairs (99.4%). The lowest level of sequence identity (91.2%) was noted between gene pairs with one cultivar having a canonical form and the other cultivar having a non-canonical form (categories C/NC and NC/C in Table 3). Only 25 pairs of alleles (three LRR-RLK, four LRR-RLP, 17 NLR and one UC) shared less than 80% cDNA identity as a result of both deletions (up to 1.7 kb) and high sequence divergence. Two of these NLRs are located in the RGA5 and Pik clusters that both hold resistance genes to rice blast disease (Table S3) (Li et al., 2007; Okuyama et al., 2011).

Genotype-specific LRR-CR genes in Nipponbare and Kitaake genomes

This gene presence absence variation (PAV) analysis revealed that 48 LRR-CR genes were present only in Nipponbare and 58 LRR-CR genes were present only in Kitaake, of which 30 (six LRR-RLK, nine LRR-RLP, 13 NLR and two UC) and 34 (11 LRR-RLK, three LRR-RLP and 20 NLR), respectively, were canonical. Note that among the 11 LRR-RLK Kitaake-specific genes are the two *Xa21* transgenes introduced into the KitaakeX sequenced genome (Jain et al., 2019). The *Xa21* gene was initially cloned from the wild rice species *Oryza longistaminata* (Song et al., 1995). We indeed identified these two transgenes at positions 28 161 378 and 28 165 947 on chromosome 6, in accordance with published data (Jain et al., 2019). Among the Nipponbare-specific genes, two LRR-RLK (OsLP2 and RLCK354), three NLR (RPR1, STA260 and Osh359-3) and one UC (Bph33) have been named previously (Hu et al., 2018) (Sakamoto et al., 1999; Thilmony et al., 2009; Yao et al., 2018).

The genotype-specific genes were not evenly distributed on the genomes. Most of them, 72.9% (35/48) and 60.3%

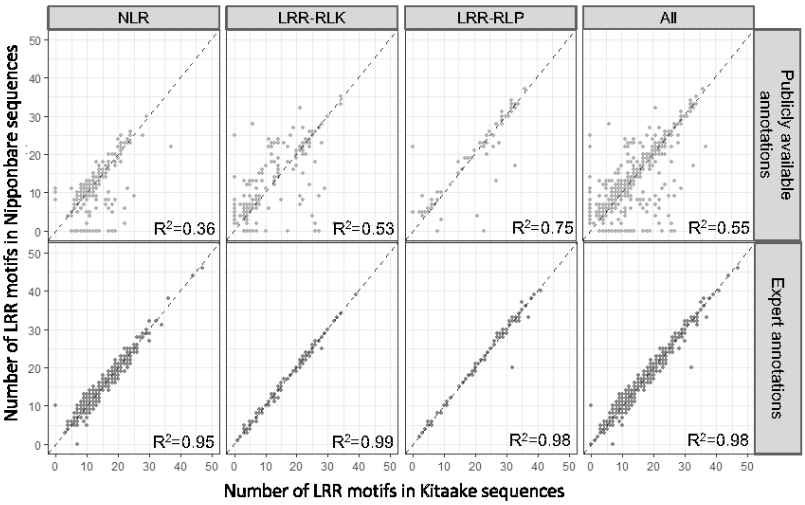


Figure 6. Comparison of LRR motif numbers between Nipponbare and Kitaake LRR-CR alleles, according to the annotations used. In green, comparison between two publicly available annotations of Nipponbare and Kitaake using IRGSP reference data for Nipponbare. In pink, comparison between our Nipponbare and Kitaake expert annotations.

Table 3 Number of allelic pairs between Nipponbare and Kitaake cultivars according to categories and subfamilies

Allele categories	Total	LRR-RLK	LRR-RLP	NLR	UC
C/C[a]	688 (68.5%)	285 (83.1%)	85 (63.9%)	289 (60.3%)	29 (58.0%)
NC/NC[a]	269 (26.8%)	47 (13.7%)	41 (30.8%)	163 (34.0%)	18 (36.0%)
C/NC[a]	29 (2.9%)	7 (2.0%)	4 (3.0%)	15 (3.1%)	3 (6.0%)
NC/C[a]	19 (1.9%)	4 (1.2%)	3 (2.3%)	12 (2.5%)	0 (0%)
Total	1005	343	133	479	50

[a]The four categories partitioned the loci according to whether they were canonical (C) or non-canonical (NC) in Nipponbare/Kitaake. Numbers in parentheses are the percentages per subfamily.

(35/58) of the Nipponbare- and Kitaake-specific loci, respectively, were located on chromosomes 2, 11 and 12. On these chromosomes, some gene clusters were entirely composed of genotype-specific genes (Figure 7a). Other genotype-specific genes were found dispersed in regions containing conserved allelic pairs (Figure 7b). Chromosome 11, which also contained about a fifth of all LRR-CR genes, hosted 43 of the 106 (40.6%) cultivar-specific loci.

Moreover, for more than half of these canonical genes (38 out of 64) the highest homology found in the Nipponbare or the Kitaake proteome is <80% of identity. Note that among these 38 genes, five Kitaake genes and seven Nipponbare genes have more than 95% of identity with indica cultivar proteins. Thus, the divergence of these genotype-specific proteins, related or not to the breeding histories of these varieties, highlights the variability of the LRR-CR repertoires between these two closely related accessions.

Finally, we took advantage of having the LRR-CR repertoire for both Nipponbare and a second rice genome to quantify putative sequence errors in the Nipponbare assembly. Among the 306 Nipponbare non-canonical genes, 241 (78.8%) presented at least one nonsense mutation also found within the Kitaake allele. These mutations are then assumed to be real. The remaining 65 non-canonical genes were manually checked and four different cases were identified: (i) Nipponbare-specific genes (18 genes, 27.7%); (ii) Kitaake allelic canonical genes (19 genes, 29.2%); (iii) Nipponbare genes that have been classified as non-canonical for a different reason than the non-canonical Kitaake allele (six genes, 9.2%); and (iv) genes for which the 'RefSeq' data of NCBI reported a potential sequence error in Nipponbare (22 genes, 33.8%). The sequence of these 65 genes was compared using BLASTn with a full-length complementary DNA (FLcDNA) clone library (Rice Full-Length cDNA Consortium, 2003) and with 14 Illumina sequence read archives (SRAs) of Nipponbare (Wang et al., 2018). The mutations observed in 18 and 40 genes (89.2%) were validated by the FLcDNA library and SRA, respectively. For four genes, no hits were obtained (6.2%). For the three remaining genes, a genomic sequence error was detected. The first one contained an 'N' that generated a frameshift, the second one had a small inversion of 24 bases that turned out to be erroneous and the third one had a wrong indel. These three genes, belonging to the NLR subfamily, were tagged in the data sets. These genes have not yet been described in the literature.

DISCUSSION

In recent years, in the wake of the gigantic volume of genome sequenced data available, it was exciting to undertake

The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2021), doi: 10.1111/tpj.15456

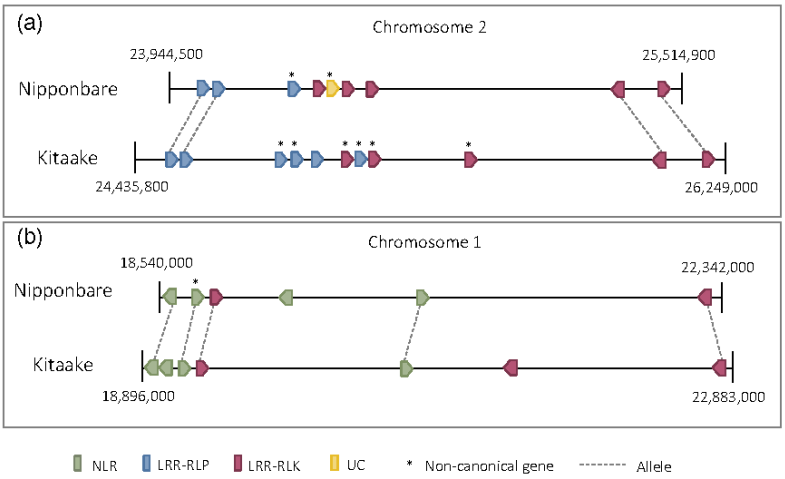


Figure 7. Schematic representation of two large loci on chromosomes 1 and 2 containing cultivar-specific LRR-CR genes.
(a) Representation of an unconserved cluster between Nipponbare and Kitaake on chromosome 2. Five and seven genes in Nipponbare and Kittake, respectively, were cultivar specific. The unconserved region was framed by four conserved genes, i.e. two LRR-RLPs and two LRR-RLKs.
(b) Representation of a conserved region between Nipponbare and Kitaake on chromosome 1 hosting cultivar-specific genes. The Nipponbare region hosted a cultivar-specific NLR, whereas the corresponding Kitaake region hosted two cultivar-specific genes, i.e. an NLR and an LRR-RLK.

evolutionary studies of gene families. We have been part of this collective enthusiasm, and like many others have based our research conclusions on perfectible versions of automatic structural and functional gene annotations (Dufayard et al., 2017; Fischer et al., 2016). Although previous genome-wide phylogenetic approaches on LRR-CR gene families enhanced our knowledge on their evolution, they almost never included a data curation step. Indeed, the manual re-annotation of gene families is a laborious time-consuming task, especially when dealing with large complex gene families, such as LRR-CR, and even more so when dealing with many plant genomes. Despite that automatic annotation tools are continuously improving, and remain essential at the genome level, human expertise is clearly still needed to achieve the level of annotation accuracy suitable for finer and deeper analyses. Today, we have finally undertaken this re-annotation work because we are convinced that these curated data are required to produce new reliable results on the evolution of these gene families, especially in the current pangenomic era. Here, we describe a new so-called 'comprehensive' annotation strategy. We hope that this annotation process will gain its place alongside the structural and functional annotations in use so far.

Automatic annotations give inconsistent gene models on complex multigenic families

Our comparison of the three publicly available annotations for the Nipponbare rice reference genome showed major discrepancies regarding the total number of LRR-CR genes, the number of LRR-CRs assigned to each subfamily and the gene models (Figure 2 and Figure S3; Table 1). These differences were greater for LRR-CR genes compared with other genes such as TFs. Automatic annotation pipelines appeared to be suitable overall for many gene families, but they led to a large proportion of inconsistent gene models when annotating complex multigenic families like LRR-CR. The annotation of fast-evolving multigene families is especially challenging for automatic approaches because a high duplication rate is often accompanied by a loss-of-function process (pseudogeneization) for many copies through, for instance, mutations like nucleotide substitutions, which may introduce premature stop codons or indels, in turn generating frameshifts. This can lead to the presence of several gene copies sharing high sequence similarity even though some of them may contain nonsense mutations, frameshifts or be truncated. The annotation of gene copies harboring nonsense mutations is problematic compared with the initial unaltered copy. Some pipelines will be able to detect the entire coding phase but will introduce false introns to sidestep stop codons or frameshifts in order to retrieve a putatively translatable CDS (Figure 3b). Indeed, we noticed that the MSU automatic annotation tended to sidestep nonsense and frameshift mutations by introducing short introns. Such errors were observed previously by Meyers when re-annotating the *A. thaliana* NLR gene family (Meyers et al., 2003). Two arguments strengthen the assertion that such introns are false: (i) such introns are never found in more than one copy, whereas the intron positions are known to be well preserved between closely related copies; and (ii) sequence comparisons performed against recent paralogs (or orthologs from close relative species) have shown that the sequences of these wrongly annotated introns are always clearly homologous to coding sequences in other gene copies. Among the intron gain mechanisms, intronization (i.e. the process by which an exonic sequence is changed into an intron by mutation accumulation) is a complex process that is not yet very well documented or understood (Roy, 2016; Yenerall and Zhou, 2012). If it really occurs in genomes, it implies that a sufficiently long period of time must have passed for these mutations to occur and generate novel splicing sites. It is

thus very unlikely that so many new introns arose in such a short period of time, as revealed by the low level of divergence between these genes and their paralog and/or ortholog counterparts. Other annotation pipelines, such as that of the IRGSP consortium, are more conservative in the sense that they give gene models with a more biologically meaningful expected structure, e.g. truncated proteins, in accordance with the presence of the first premature stop codons (either in-frame or caused by a frameshift; Figure 3b). This conservative choice could likely explain why we found more sequences classified as LRR-RLP from the IRGSP annotation than from the two other annotations (Table 1). Indeed, any LRR-RLK with a premature stop codon somewhere before the kinase domain would be considered as LRR-RLP (Figure 1). The annotation inconsistencies that we pinpointed here were observed for LRR-CR genes and did not question the overall quality of the three available rice genome annotations. They highlighted the limits of automatic annotation pipelines to annotate such complex multigene families and the consequences that given pipeline decision rules may have when drawing evolutionary conclusions.

The comparisons of the different gene models proposed by the three Nipponbare annotations led us to undertake a manual curation of the LRR-CR gene family. We are not the first to get involved in this painstaking but necessary work. Several high-quality studies have been based on re-annotated data, particularly in *A. thaliana* (Meyers et al., 2003; Van de Weyer et al., 2019). Expert annotations could also contain errors, of course, but expert curation limits their number. Opting exclusively for automated annotation should be avoided or otherwise operators should be aware that these annotations may contain errors induced by gene family specificities. These biases must be known and understood to avoid drawing misleading conclusions.

The expert Nipponbare rice genome contains more than 1000 LRR-CR loci, of which 30% have a non-canonical gene model

We curated LRR-CR loci in the reference Nipponbare rice genome by first comparing the three publicly available annotations at each locus: IRGSP, MSU and NCBI. Our aim was to retrieve LRR-CR genes in their entirety and account for the coding sequences as they probably stood before mutation accumulation. We obtained evidence that the sequence portions that we included in our gene models were not random genomic sequences but instead parts of the original gene CDS, as shown by the recovery of protein domains belonging to LRR-CR genes (kinase, NB-ARC, TM; Figure S4). Man et al. (2020) reported seven cases of missed domains through probable annotation errors in rice. We have also identified and corrected these seven genes and recovered the same domains. However, because our search for annotation errors was exhaustive, we recovered a higher number of missed domains.

When a gene had a nonsense mutation (in-frame stop codon or frameshift), an unexpected splicing site, or no terminal stop or start codon, we tagged it as non-canonical. This canonical versus non-canonical classification was based solely on features observed in gene models and did not imply any judgements on gene functionality. Genes tagged as non-canonical spanned a wide variety of cases, some of which could very likely not be translated into a functional protein while others may have had a function. As a first example, mutations inducing a premature stop codon could lead to a shorter protein that might sometimes perform the same function. Yet in many other cases shorter proteins might not perform the same function, if able to perform any function at all. Another example concerns the loss of the expected stop codon. When screening a sequence, a stop codon will eventually be encountered, but determining the functional consequences of this additional amino acid stretch would be impossible *in silico*. The same holds when the start codon is lost. Determining the criteria by which an alternative start codon (if any) may become the new start codon is a hazardous task. These few examples highlight the extent to which sorting out different functional scenarios is challenging. Moreover, mRNA molecules may play a regulating role, even if they cannot be translated as such, thereby justifying the need to annotate them. These reflections led us to voluntarily disregard such interpretations in our re-annotation process.

We observed that a third of the LRR-CR genes were non-canonical, but their proportion varied according to the gene subfamily (Table 2). A lower proportion of LRR-RLK genes were non-canonical (15%), compared with LRR-RLP (33%) and NLR (36%). The LRR-RLK subfamily could be divided into 15 20 subgroups based on phylogenetic study findings, and the duplication rate was shown to be quite variable according to the subgroup considered (Fischer et al., 2016; Tang et al., 2010). Some subgroups, the genes of which have been described as mostly involved in developmental processes, have had a more stable copy number over the course of angiosperm evolution (Fischer et al., 2016). These genes are less prone to duplication and thus are less likely to generate copies accumulating nonsense mutations, thereby lowering the proportion of non-canonical genes when the entire LRR-RLK subfamily is considered. The higher proportion of non-canonical genes obtained for NLR and LRR-RLP suggests that these subfamilies generally have higher birth and death rates. A quarter of the LRR-CR genes required manual curation and were non-canonical (representing 83.5% of the curated loci and 89.5% of the non-canonical loci; Table S2). In fact, manual curation was conducted mainly when none of the three annotations gave satisfactory gene models (such as the example presented in Figure 3b). The high correlation

The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2021), doi: 10.1111/tpj.15456

between non-canonical and curated loci was thus likely caused by the presence of mutations introducing ambiguities, which are overcome to different extents by the three annotation pipelines. A step forward would be to improve the annotation tools so that they deal differently with these nonsense mutations, e.g. including them in the sequences and indicating their presence without sidestepping them. To this end, annotation tools would have to predict non-canonical structures and tag them accordingly. To process such complex data, machine learning approaches are very promising (Mahood et al., 2020), but this implies having a significant learning corpus that has yet to be built.

We stress that this categorization of loci into canonical or non-canonical models could be impacted by the genomic sequence quality. Errors in the reference genome sequence could introduce errors in the gene models. In our curated data set, 27 non-canonical genes were tagged in NCBI data as harboring a difference between the RefSeq transcript sequence or protein and the Nipponbare reference sequence. The mutations jeopardizing the expected gene structure corresponded exactly to the positions where inconsistencies had been highlighted between the genomic and the RefSeq data. In order to appreciate the impact of errors when genes were categorized as non-canonical, we checked 65 of them in Nipponbare using both expression data and genome resequencing data. Only three probable errors were detected (one containing an 'N'), and four could not be validated. Moreover, among the 27 genes for which NCBI reported a potential error in the genomic sequence, 25 actually contained the identified nonsense mutation. Redundancies in LRR-CR gene sequences can give rise to ambiguities during both genome sequence assembly and expression data mapping, thus leading to errors (Torresen et al., 2019). Access to more specific re-sequencing data will resolve those potential inconsistencies. In the current state of the data, the reference genome errors identified concern less than 1% of the non-canonical genes.

LRR-CR repertoire in Kitaake, and comparison with Nipponbare

We propose a modular strategy to transfer our manually curated annotations to other rice genomes. We applied this strategy to annotate LRR-CR genes from the genome of the Kitaake cultivar, which also belongs to the *O. sativa* japonica subspecies. A comparison of the Nipponbare and Kitaake LRR-CR repertoires revealed an equivalent number of loci. The distributions of LRR-CR loci per gene subfamily, chromosome and category (canonical or not) were also consistent between these two cultivars (Figure 5).

In the Nipponbare genome, eight new LRR-CRs (one LRR-RLK, three LRR-RLPs and four UCs) were identified. These genes had not been previously annotated in any of the three publicly available annotations. In Kitaake, the same strategy enabled us to identify 114 new LRR-CR genes (48 of which were canonical). The higher number of unannotated LRR-CR genes in the Kitaake genome compared with the Nipponbare genome (114 versus eight) suggested that annotation inaccuracies had a greater impact on recently sequenced genomes that have not benefited from as much annotation investment as reference genomes.

A comparison of the LRR motif number for all allelic pairs between Nipponbare and Kitaake revealed a much greater difference in LRR number between alleles for publicly available annotations (ranging from 2.68 to 3.58), in comparison with our manually curated annotations (0.58), when the three subfamilies were all considered (Figure 6 and Figure S5). When publicly available annotations are considered, some rare allelic pairs harboring a very different number of LRRs may be truly different functional alleles. For instance, between two LRR-RLP alleles, one may contain a premature stop codon leading to the loss of a few motifs, but it may still have a biological function. It is important to identify such a pair. In our expert annotation alleles may share an identical number of LRRs, but such allelic pairs would be clearly identifiable because one of the alleles would be tagged as non-canonical whereas the other would be tagged as canonical. Moreover, in non-canonical alleles, the causal mutation, its position and impact on the gene (i.e. frameshift or premature stop codon) could be identified.

The difference in LRR motif number observed between allele pairs was greater for NLR than for the other subfamilies (Figure 6). This might be explained by the fact that NLR motifs are more variable and hence harder to detect (Ng et al., 2011), which could lead to apparent variations in the number of motifs in the two alleles. LRR motifs that have been found in NLRs differed from the common 'plant-specific' LRR consensus sequence, and were more irregular in terms of both length and residue conservation (Kajava, 1998; Kuang et al., 2004; Matsushima and Miyashita, 2012; Sela et al., 2012). Although we enhanced the LRR detection accuracy through the development of a new LRR HMM profile for NLR, it is still not exhaustive. This also suggests that the number of LRR motifs varies more in NLR than in other LRR-CR subfamilies.

High sequence similarity was observed between Nipponbare and Kitaake alleles (98.9% identity for cDNA) (Table S3), which was consistent with previous comparisons (Jain et al., 2019). However, some allelic pairs showed a lower identity level with a more ancient coalescent history between the two genomes. This heterogeneity may have been the consequence of the breeding programs from which these varieties were derived. The breeding process involves crosses with more or less closely related genotypes, sometimes from different subspecies, and may generate mosaic genomes (Santos et al., 2019). No allelic

pairs were found for 106 genes: i.e. 48 were specific to Nipponbare and 58 were specific to Kitaake. A majority of those genes were located in clusters on chromosomes 2, 11 and 12, which have already been described as containing a large number of LRR-CRs (Mizuno et al., 2020; Zhou et al., 2004) (Figure 7). Some clusters have also been shown to be less conserved (Mizuno et al., 2020). More than half of these genes were classified as canonical.

The methods we developed allowed us to undertake an exhaustive comparison of the LRR-CR repertoire between Nipponbare and Kitaake. Allelic pairs, including those hosting nonsense mutations in either or both genotypes, were described (Data S4). Genotype-specific genes were also identified and localized, again along with information related to the potential presence of nonsense mutations (Figure 7). These results were achieved through a combination of an expert annotation and its transfer to a second genotype for which a high quality *de novo* genome assembly was available. Validation of the LRR-CR annotations of Kitaake was not very time consuming compared with the initial work in Nipponbare, where each gene was investigated individually. Our study highlighted that investment in a combination of technologies would guarantee high-quality assemblies and annotations, especially when the discovery of allelic diversity is targeted (Zhou et al., 2020).

The tools and curated data sets that we generated in this study are available from: https://rice-genome-hub.southgreen.fr/content/geloc (data) and https://github.com/cgottin/LRRprofiler (tools). Note that we focused on developing a website where stop codons and frameshifts are easily identified. We believe that evolutionary studies and allele discovery initiatives for LRR-CRs would be more accurate and reliable when using our manually curated comprehensive annotations for these genes. Moreover, we feel that this comprehensive annotation approach should be widely adopted by the community in the light of the major potential benefits it provides.

EXPERIMENTAL PROCEDURES

Genomes and annotation files

Reference genomic sequences of Nipponbare (Kawahara et al., 2013) and Kitaake (Jain et al., 2019) *O. sativa* ssp. *japonica* cultivars were downloaded from the Rice Annotation Project Database (RAP-DB) website (https://rapdb.dna.affrc.go.jp) and the Phytozome website (https://phytozome.jgi.doe.gov/pz/portal.html). The general feature format (GFF) and fasta files with coding DNA sequences (CDSs) and protein sequences for Nipponbare were downloaded for three different annotation projects: (i) the MSU 7.0 annotation was downloaded from the Rice Genome Annotation Project FTP server (http://rice.plantbiology.msu.edu); (ii) the IRGSP annotation files were downloaded from the RAP-DB website (https://rapdb.dna.affrc.go.jp); and (iii) the NCBI annotation (release 102) annotated by the NCBI Eukaryotic Genome Annotation Pipeline was downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov). The IRGSP annotation consists of two gene sets ('genes supported by FL-cDNAs, ESTs or proteins' and 'computationally predicted genes') that were concatenated for the analyses (Sakai et al., 2013).

LRRPROFILER implementation

The LRRPROFILER pipeline was implemented in two steps (Figure S1). The first step involved the iterative refinement of LRR HMM profiles specific to a gene subfamily (LRR-RLK or NLR) and proteome (Figure S1; inspired by Ng et al., 2011). Only LRR-RLK and NLR were considered for profile refinement because they contain a specific domain (i.e. kinase and NB-ARC domains, respectively), thereby allowing the clear identification of the subfamily to which they belong. A set of candidate protein sequences was identified from a given proteome to refine the specific LRR profiles. This set was composed of either LRR-RLKs identified with iTAK (Zheng et al., 2016) or NLRs identified with the PF00931 Pfam NB-ARC profile. A first round of LRR motif detection was performed in either of the candidate protein sets using HMMSEARCH (HMMER; Eddy, 2011) with the SM00370 LRR profile from the SMART database. Motifs of 20–26 amino acids in length were extracted, aligned with MAFFT (Katoh and Standley, 2013) with default parameters and a new profile was built from the alignment using HMMBUILD (HMMER; Eddy, 2011). This process was repeated using the HMM LRR profile built at the previous iteration to search again for LRR motifs in the considered protein candidate set. At each iteration, the sum of the amino acid lengths of the detected LRR motifs was calculated. The process stopped when three iterations (not necessarily consecutive) resulted in a decrease of the statistics. Finally, the process retrieved the HMM LRR profile identifying the maximum number of LRR motifs in the candidate protein set.

The second step of the LRRPROFILER pipeline consisted of the identification of LRR-CR proteins present in a given proteome, the annotation of their functional domains as well as their classification into a gene subfamily: LRR-RLK, LRR-RLP, NLR or UC (Figure S1). Six publicly available LRR HMM profiles from the SMART database, i.e. SM00364 (LRR_BAC), SM00365 (LRR_SD22), SM00367 (LRR_CC), SM00368 (LRR_RI), SM00369 (LRR_TYP) and SM00370 (LRR), in addition to the newly built LRR profiles obtained in the first step were used to detect LRR motifs in the complete proteome under consideration using HMMSEARCH. An annotation of the LRR domains, containing the start and end positions of each LRR motif, was part of the output. The annotation of each protein was then supplemented using publicly available profiles for other functional domains: TIR (PF01582), TIR_2 (PF13676), Malectin (PF11721), Malectin-like (PF12819), RPW8 (PF05659), Cys-Pairs (Dievart and Clark, 2003; Dufayard et al., 2017), F-box (PF00646) and FBD (PF08387). NB-ARC and kinase domain annotations were retrieved from the first step, whereas transmembrane domains (TMs) were detected with TMHMM 2.0c, with default parameters (Sonnhammer et al., 1998). The subfamily assignment of each identified LRR-containing protein was deduced from its domain structure. Proteins were classified into the LRR-RLK subfamily if they contained at least one LRR motif and a kinase domain, and sometimes other domains such as the malectin, malectin-like, Cys-pair and TM domains. Proteins were classified in the NLR subfamily if they included an NB-ARC domain and at least one LRR motif, sometimes with a TIR or an RPW8 domain. The LRR-RLP subfamily included proteins with LRRs plus a TM, malectin, malectin-like and/or Cys-pair, or LRR-only structures when at least 13 plant-specific LRR repeats were detected. Proteins containing an F-box or an FBD domain in addition to LRRs were classified as F-box-LRR. All other LRR-containing proteins were ranked in the UC group, and for these we performed a

BLASTp search with default parameters against the other gene sets (LRR-RLP, LRR-RLK, NLR and F-box) to estimate their probable membership of one of these gene subfamilies. F-box proteins were removed from our data sets and not considered further in the analyses. We ended up with four gene sets: LRR-RLP, LRR-RLK, NLR and UC.

At the end of the construction phase, the complete LRRPROFILER pipeline was tested on the manually reviewed *A. thaliana* protein data set downloaded from the Swiss-Prot section (https://www.uniprot.org) (Data S1; Figure S2; Methods S1; Table S1) (Boutet et al., 2007) of the UniProt databank (The UniProt Consortium, 2019). This set was composed of 15 818 sequences. Domain and repeat information was also extracted from the database, in particular the number of LRR motifs per sequence and the gene subfamily to which it belonged (LRR-RLP, LRR-RLK, NLR, etc.).

Rice transcription factor data set

Transcription factor genes (TFs) were identified in the proteome predicted from the three publicly available annotations of the Nipponbare rice reference genome using ITAK (Zheng et al., 2016). Nine subfamilies were considered: C2H2, FAR1, MYB-related, WRKY, NAC, AP2/ERF-ERF, bHLH, bZIP and MYB.

Annotation transfer from Nipponbare to Kitaake

The first phase consisted of locating regions of interest in the Kitaake genome, i.e. regions homologous to Nipponbare LRR-CR loci (Figure 4a). Nipponbare LRR-CR protein sequences from our expert annotations were aligned with the Kitaake genome using tBLASTn (Altschul et al., 1990). Only high scoring pairs (HSPs) with more than 65% identity with Nipponbare LRR-CR protein fragments and spanning at least 50% of the Nipponbare query protein were retained. To define coherent candidate regions in Kitaake, HSPs from the same Nipponbare query protein had to be located less than 5000 bp apart, except when the Nipponbare homologous gene queried had a longer intron. In that case, the Nipponbare intron length plus 500 bp was used as the upper bound for the distance separating Kitaake HSPs. Multiple regions of interest could be found for a single Nipponbare protein. This allowed us to annotate genes duplicated in the Kitaake genome even if a single gene copy was present in the Nipponbare genome. In a second phase, gene model determination was attempted in three consecutive steps for each region of interest (Figure 4b). Only regions that could not be successfully annotated at a given step passed to the next step. In the first step, the Nipponbare query exons are mapped to the target Kitaake region of interest with BLASTn. A gene model was then reconstructed based on ordered HSPs. The gene model reconstruction quality was checked by comparing the predicted protein with that of Nipponbare using BLASTp. The gene model was retained if all expected exons were present and the Kitaake protein sequence had more than 90% identity with the Nipponbare protein sequence. Otherwise, the annotation of this region was delegated to the second step. In the second step, the EXONERATE cdna2genome model (Slater and Birney, 2005) was run independently for every remaining query/target pair. The EXONERATE output GFF file was parsed to construct the target gene model and to document putative frameshift positions. Again, the Kitaake predicted protein was compared with the Nipponbare query sequence with BLASTp and retained if the coverage and identity were above 90 and 75%, respectively. Otherwise, the annotation of this target region was delegated to the third step. In the third step, the remaining loci were reconstructed with the EXONERATE protein2genome model. This model is better at finding the correct reading frames when the target and model loci are more divergent, but it fails to correctly annotate type-1 and -2 splicing sites (intron/exon junction falling inside a codon). This problem arises because it uses the same reading frame to translate the whole genomic sequence (the six reading frames are tested, but each resulting translation just uses one of them). To overcome this issue, intron junctions are then corrected with a PYTHON script that looks for canonical splicing sites in a range of two nucleotides before and after the current junctions. Finally, gene models highly divergent from the Nipponbare query sequence, with multiple premature stop codons or without start or terminal stop codons, and overlapping frameshifts are tagged to be checked manually.

Identification of alleles between Nipponbare and Kitaake

We used SYNMAP (Lyons and Freeling, 2008) to identify LRR-CR allelic pairs, i.e. genes with the exact same chromosomic position in Nipponbare and Kitaake. SYNMAP was developed to identify orthologous genes between different species based on microcolinearity conservation, and it identifies blocks of genes of conserved order and position. It retrieves a list of relationships between genic repertoires of two genomes. We identified alleles by first selecting genes for which SYNMAP found a reciprocal relationship, i.e. a relationship found in both Nipponbare–Kitaake and Kitaake–Nipponbare comparisons. Genes for which allelic relationships could not be unambiguously resolved by SYNMAP were manually resolved, when possible, using VISTA (Mayor et al., 2000) and ARTEMIS (Carver et al., 2012).

ACKNOWLEDGMENTS

This work was partly supported by the CGIAR Research Program on Rice (CRP RICE) (to AD) and by a PhD fellowship from Institut Agro and CIRAD (to CG). Computational resources were provided by the South Green bioinformatics platform. The authors would like to thank Dr A. Cenci for a critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

CG, NC, AD, CP and VR designed the research. CG, NC and AD performed the research. CG, NC, AD, VR, MS and GD contributed to new analytic and computational tools. CG, NC and AD analyzed the data. CG, NC, AD and VR wrote the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest associated with this work.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Schematic representation of the LRRPROFILER pipeline.

Figure S2. Comparison of expected and predicted LRR motifs in protein sequences from the Swiss-Prot *Arabidopsis thaliana* data set using publicly available and refined HMM profiles.

Figure S3. Comparison of predicted peptide lengths between Nipponbare publicly available annotations (IRGSP, MSU and NCBI) for LRR-CR and TF loci.

Figure S4. Number of domains and motifs identified with LRRPROFILER for the Nipponbare proteomes predicted by publicly available and manually curated annotations.

Figure S5. LRR motif number conservation between Nipponbare and Kitaake LRR-CR loci, depending on the annotation compared.

Table S1. Performance of publicly available and refined LRR HMM profiles in the Swiss-Prot *Arabidopsis thaliana* data set.

Table S2. Contingency table of canonical/non-canonical and modified/not modified LRR-CR loci from the Nipponbare manually curated annotation.

Table S3. Percentage of cDNA identity between Nipponbare and Kitaake alleles according to gene subfamilies and categories.

Methods S1. Validation of the LRRPROFILER pipeline.

Data S1. LRRPROFILER results in the Swiss-Prot *Arabidopsis thaliana* data set.

Data S2. LRR-CR loci from the Nipponbare rice reference genome.

Data S3. LRR-CR loci from the rice KitaakeX genome.

Data S4. Allelic relationship and cDNA identity between Nipponbare and Kitaake LRR-CR loci.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at the detailed link as follows: https://doi.org/10.5281/zenodo.5110015

DATA AVAILABILITY STATEMENT

All of the data files (gff and fasta files) are available from the dedicated website (https://rice-genome-hub.southgreen.fr/content/geloc) and from the open data repository Zenodo (https://doi.org/10.5281/zenodo.5110015).

A new identifier was allocated to each LRR-CR gene unraveled by this procedure. These identifiers use the <OSJnip_ChrXX_00000000> or <OSJkit_ChrXX_00000000> pattern for Nipponbare and Kitaake loci, respectively, with XX being the chromosome number followed by the start codon position of the coding sequence (CDS) on the chromosome (Data S2 and S3).

According to the Multiple Alignment of Coding Sequences (MACSE) convention (Ranwez et al., 2018; Ranwez et al., 2011), indels causing frameshift mutations have been pinpointed by the presence of one or two '!' characters in the nucleotide sequences of non-canonical genes and are available in an additional specific data set.

REFERENCES

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S. et al. (2016) The Ensembl gene annotation system. *Database*, **2016**, baw093.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Andersen, E.J., Nepal, M.P., Purintun, J.M., Nelson, D., Mermigka, G. & Sarris, P.F. (2020) Wheat disease resistance genes and their diversification through integrated domain fusions. *Frontiers in Genetics*, 11, 898.

Bailey-Serres, J., Parker, J.E., Ainsworth, E.A., Oldroyd, G.E.D. & Schroeder, J.I. (2019) Genetic strategies for improving crop yields. *Nature*, **575**, 109–118.

Bayer, P.E., Edwards, D. & Batley, J. (2018) Bias in resistance gene prediction due to repeat masking. *Nature Plants*, **4**, 762–765.

Bella, J., Hindle, K.L., McEwan, P.A. & Lovell, S.C. (2008) The leucine-rich repeat structure. *Cellular and Molecular Life Sciences*, **65**, 2307–2333.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. (2007) UniProtKB/Swiss-Prot. *Methods in Molecular Biology*, **406**, 89–112.

Boutrot, F. & Zipfel, C. (2017) Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. *Annual review of Phytopathology*, **55**, 257–286.

Burdett, H., Bentham, A.R., Williams, S.J., Dodds, P.N., Anderson, P.A., Banfield, M.J. et al. (2019) The plant "Resistosome": structural Insights into Immune Signaling. *Cell Host & Microbe*, **26**, 193–201.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J. & McQuillan, J.A. (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, **28**, 464–469.

Couto, D. & Zipfel, C. (2016) Regulation of pattern recognition receptor signalling in plants. *Nature Reviews Immunology*, **16**, 537–552.

Dievart, A. & Clark, S.E. (2003) Using mutant alleles to determine the structure and function of leucine-rich repeat receptor-like kinases. *Current Opinion in Plant Biology*, **6**, 507–516.

Dufayard, J.F., Bettembourg, M., Fischer, I., Droc, G., Guiderdoni, E., Perin, C. et al. (2017) New insights on leucine-rich repeats receptor-like kinase orthologous relationships in angiosperms. *Frontiers in Plant Science*, **8**, 381.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195.

FAO. (2018) The future of food and agriculture—Alternative pathways to 2050 Rome.

Fawal, N., Li, Q., Mathe, C. & Dunand, C. (2014) Automatic multigenic family annotation: risks and solutions. *Trends in Genetics*, **30**, 323–325.

Fischer, I., Dievart, A., Droc, G., Dufayard, J.F. & Chantret, N. (2016) Evolutionary dynamics of the leucine-rich repeat receptor-like kinase (LRR-RLK) subfamily in angiosperms. *Plant Physiology*, **170**, 1595–1610.

Fritz-Laylin, L.K., Krishnamurthy, N., Tor, M., Sjolander, K.V. & Jones, J.D. (2005) Phylogenomic analysis of the receptor-like proteins of rice and Arabidopsis. *Plant Physiology*, **138**, 611–623.

Furumizu, C. & Sawa, S. (2021) Insight into early diversification of leucine-rich repeat receptor-like kinases provided by the sequenced moss and hornwort genomes. *Plant Molecular Biology*. Online ahead of print. https://doi.org/10.1007/s11103-020-01100-0

Han, G.Z. (2019) Origin and evolution of the plant immune system. *New Phytologist*, **222**, 70–83.

Hosseini, S., Schmidt, E.D.L. & Bakker, F.T. (2020) Leucine-rich repeat receptor-like kinase II phylogenetics reveals five main clades throughout the plant kingdom. *The Plant Journal*, **103**, 547–560.

Hu, J., Chang, X., Zou, L., Tang, W. & Wu, W. (2018) Identification and fine mapping of Bph33, a new brown planthopper resistance gene in rice (Oryza sativa L.). *Rice*, **11**, 55.

Hwang, S.G., Kim, D.S. & Jang, C.S. (2011) Comparative analysis of evolutionary dynamics of genes encoding leucine-rich repeat receptor-like kinase between rice and Arabidopsis. *Genetica*, **139**, 1023–1032.

Innan, H. & Kondrashov, F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, **11**, 97–108.

Jain, R., Jenkins, J., Shu, S., Chern, M., Martin, J.A., Copetti, D. et al. (2019) Genome sequence of the model rice variety KitaakeX. *BMC Genomics*, **20**, 905.

Jones, D.A. & Jones, J.D.G. (1997) The role of leucine-rich repeat proteins in plant defences. In *Advances in Botanical Research* (Andrews, J.H., Tommerup, I.C. & Callow, J.A., eds). Academic Press, pp. 89–167.

Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G.J. et al. (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *The Plant Journal*, **76**, 530–544.

Kajava, A.V. (1998) Structural diversity of leucine-rich repeat proteins. *Journal of Molecular Biology*, **277**, 519–527.

Kajava, A.V. (2012) Tandem repeats in proteins: from sequence to structure. *Journal of Structural Biology*, **179**, 279–288.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S. et al. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.

Kuang, H., Woo, S.S., Meyers, B.C., Nevo, E. & Michelmore, R.W. (2004) Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *The Plant Cell*, **16**, 2870–2894.

Lai, X., Chahtane, H., Martin-Arevalillo, R., Zubieta, C. & Parcy, F. (2020) Contrasted evolutionary trajectories of plant transcription factors. *Current Opinion in Plant Biology*, **54**, 101–107.

Lee, H.Y., Mang, H., Choi, E., Seo, Y.E., Kim, M.S., Oh, S. et al. (2021) Genome-wide functional analysis of hot pepper immune receptors reveals an autonomous NLR clade in seed plants. *New Phytologist*, **229**, 532–547.

Leister, D. (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends in Genetics*, **20**, 116–122.

Li, J., Ding, J., Zhang, W., Zhang, Y., Tang, P., Chen, J.Q. et al. (2010) Unique evolutionary pattern of numbers of gramineous NBS-LRR genes. *Molecular Genetics and Genomics*, **283**, 427–438.

Li, L.-Y., Wang, L., Jing, J.-X., Li, Z.-Q., Lin, F., Huang, L.-F. et al. (2007) The Pikm gene, conferring stable resistance to isolates of Magnaporthe oryzae, was finely mapped in a crossover-cold region on rice chromosome 11. *Molecular Breeding*, **20**, 179–188.

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S. & You, F.M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics*, **17**, 852.

Lyons, E. & Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, **53**, 661–673.

Mahood, E.H., Kruse, L.H. & Moghe, G.D. (2020) Machine learning: a powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, **8**, e11376.

Man, J., Gallagher, J.P. & Bartlett, M. (2020) Structural evolution drives diversification of the large LRR-RLK gene family. *New Phytologist*, **226**, 1492–1505.

Martin, E.C., Sukarta, O.C.A., Spiridon, L., Grigore, L.G., Constantinescu, V., Tacutu, R. et al. (2020) LRRpredictor-A new LRR motif detection method for irregular motifs of plant NLR proteins using an ensemble of classifiers. *Genes*, **11**, 286.

Matsushima, N. & Miyashita, H. (2012) Leucine-rich repeat (LRR) domains containing intervening motifs in plants. *Biomolecules*, **2**, 288–311.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A. et al. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046–1047.

McDowell, J.M. & Simon, S.A. (2006) Recent insights into R gene evolution. *Molecular Plant Pathology*, **7**, 437–448.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R.W. (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *The Plant Cell*, **15**, 809–834.

Michelmore, R.W. & Meyers, B.C. (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research*, **8**, 1113–1130.

Mizuno, H., Katagiri, S., Kanamori, H., Mukai, Y., Sasaki, T., Matsumoto, T. et al. (2020) Evolutionary dynamics and impacts of chromosome regions carrying R-gene clusters in rice. *Scientific Reports*, **10**, 872.

Nei, M. & Rooney, A.P. (2005) Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, **39**, 121–152.

Ng, A.C., Eisenberg, J.M., Heath, R.J., Huett, A., Robinson, C.M., Nau, G.J. et al. (2011) Human leucine-rich repeat proteins: a genome-wide bioinformatic categorization and functional analysis in innate immunity. *Proceedings of the National Academy of Sciences U S A*, **108**(Suppl 1), 4631–4638.

Okuyama, Y., Kanzaki, H., Abe, A., Yoshida, K., Tamiru, M., Saitoh, H. et al. (2011) A multifaceted genomics approach allows the isolation of the rice Pia-blast resistance gene consisting of two adjacent NBS-LRR protein genes. *The Plant Journal*, **66**, 467–479.

Prigozhin, D.M. & Krasileva, K.V. (2021) Analysis of intraspecies diversity reveals a subset of highly variable plant immune receptors and predicts their binding sites. *The Plant Cell*, **33**, 998–1015.

Ranwez, V., Douzery, E.J.P., Cambon, C., Chantret, N. & Delsuc, F. (2018) MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, **35**, 2582–2584.

Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E.J. (2011) MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One*, **6**, e22594.

Rice Full-Length cDNA Consortium, National Institute of Agrobiological Sciences Rice Full-Length cDNA Project Team; Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K. et al. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from Japonica rice. *Science*, **301**, 376–379.

Richter, T.E. & Ronald, P.C. (2000) The evolution of disease resistance genes. *Plant Molecular Biology*, **42**, 195–204.

Roy, S.W. (2016) How common is parallel intron gain? Rapid evolution versus independent creation in recently created introns in daphnia. *Molecular Biology and Evolution*, **33**, 1902–1906.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y. et al. (2013) Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant and Cell Physiology*, **54**, e6.

Sakamoto, K., Tada, Y., Yokozeki, Y., Akagi, H., Hayashi, N., Fujimura, T. et al. (1999) Chemical induction of disease resistance in rice is correlated with the expression of a gene encoding a nucleotide binding site and leucine-rich repeats. *Plant Molecular Biology*, **40**, 847–855.

Santos, J.D., Chebotarov, D., McNally, K.L., Bartholome, J., Droc, G., Billot, C. et al. (2019) Fine scale genomic signals of admixture and alien introgression among Asian rice landraces. *Genome Biology and Evolution*, **11**, 1358–1373.

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N. & Nelson, A. (2019) The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, **3**, 430–439.

Sekhwal, M.K., Li, P., Lam, I., Wang, X., Cloutier, S. & You, F.M. (2015) Disease resistance gene analogs (RGAs) in plants. *International Journal of Molecular Sciences*, **16**, 19248–19290.

Sela, H., Spiridon, L.N., Petrescu, A.J., Akerman, M., Mandel-Gutfreund, Y., Nevo, E. et al. (2012) Ancient diversity of splicing motifs and protein surfaces in the wild emmer wheat (Triticum dicoccoides) LR10 coiled coil (CC) and leucine-rich repeat (LRR) domains. *Molecular Plant Pathology*, **13**, 276–287.

Shao, Z.Q., Wang, B. & Chen, J.Q. (2016) Tracking ancestral lineages and recent expansions of NBS-LRR genes in angiosperms. *Plant Signaling & Behavior*, **11**, e1197470.

Shiu, S.H. & Bleecker, A.B. (2001a) Plant receptor-like kinase gene family: diversity, function, and signaling. *Science STKE*, **2001**, re22.

Shiu, S.H. & Bleecker, A.B. (2001b) Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. *Proceedings of the National Academy of Sciences U S A*, **98**, 10763–10768.

Slater, G.S. & Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Song, W.Y., Wang, G.L., Chen, L.L., Kim, H.S., Pi, L.Y., Holsten, T. et al. (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science*, **270**, 1804–1806.

Sonnhammer, E.L., von Heijne, G. & Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, **6**, 175–182.

Stanke, M. & Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**(Suppl 2), ii215–ii225.

Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C. et al. (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nature Genetics*, **50**, 285–296.

Steuernagel, B., Witek, K., Krattinger, S.G., Ramirez-Gonzalez, R.H., Schoonbeek, H.J., Yu, G. et al. (2020) The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiology*, **183**, 468–482.

Sun, X. & Wang, G.L. (2011) Genome-wide identification, characterization and phylogenetic analysis of the rice LRR-kinases. *PLoS One*, **6**, e16079.

Tamborski, J. & Krasileva, K.V. (2020) Evolution of plant NLRs: from natural history to precise modifications. *Annual Review of Plant Biology*, **71**, 355–378.

Tang, P., Zhang, Y., Sun, X., Tian, D., Yang, S. & Ding, J. (2010) Disease resistance signature of the leucine-rich repeat receptor-like kinase genes in four plant species. *Plant Science*, **179**, 399–406.

The UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**, D506–D515.

Thilmony, R., Guttman, M., Thomson, J.G. & Blechl, A.E. (2009) The LP2 leucine-rich repeat receptor kinase gene promoter directs organ-specific,

light-responsive expression in transgenic rice. *Plant Biotechnology Journal*, **7**, 867–882.

Torresen, O.K., Star, B., Mier, P., Andrade-Navarro, M.A., Bateman, A., Jarnot, P. et al. (2019) Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, **47**, 10994–11006.

Van de Weyer, A.L., Monteiro, F., Furzer, O.J., Nishimura, M.T., Cevik, V., Witek, K. et al. (2019) A species-wide inventory of NLR genes and alleles in Arabidopsis thaliana. *Cell*, **178**, 1260–1272.e14

van der Burgh, A.M. & Joosten, M. (2019) Plant immunity: thinking outside and inside the box. *Trends in Plant Science*, **24**, 587–601.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z. et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.

Wilming, L. & Harrow, J. (2009) Gene annotation methods. In *Bioinformatics* (Edwards, D., Stajich, J. & Hansen, D., eds). New York, NY: Springer.

Xiong, Y., Han, Z. & Chai, J. (2020) Resistosome and inflammasome: platforms mediating innate immunity. *Current Opinion in Plant Biology*, **56**, 47–55.

Yao, W., Li, G., Yu, Y. & Ouyang, Y. (2018) funRiceGenes dataset for comprehensive understanding and application of rice functional genes. *Gigascience*, **7**, 1–9.

Yenerall, P. & Zhou, L. (2012) Identifying the mechanisms of intron gain: progress and trends. *Biology Direct*, **7**, 29.

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R. et al. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Research*, **31**, 229–233.

Zheng, Y., Jiao, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P. et al. (2016) iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Molecular Plant*, **9**, 1667–1670.

Zhou, T., Wang, Y., Chen, J.Q., Araki, H., Jing, Z., Jiang, K. et al. (2004) Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Molecular Genetics and Genomics*, **271**, 402–415.

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S. et al. (2020) A platinum standard pan-genome resource that represents the population structure of Asian rice. *Scientific Data*, **7**, 113.

The Plant Journal, (2021), doi: 10.1111/tpj.15456

Chapitre 3

Chapitre 3 : Développements pour la valorisation des outils et des données d'annotation

Table des matières

A. Introduction

Au cours de cette thèse, j'ai développé un pipeline, LRRprofiler, permettant d'identifier et d'annoter *de novo* les séquences protéiques de récepteurs à LRR à partir d'un protéome (cf. chapitre 1). J'ai également effectué un travail de ré-annotation des modèles de gène des récepteurs LRR dans le génome de Nipponbare. Afin de transférer cette nouvelle annotation vers d'autres génomes de riz, dont KitaakeX, j'ai développé un pipeline de transfert d'annotation : LRRtransfer (cf. chapitre 2). Par ailleurs, cette ré-annotation ayant la particularité de contenir des gènes dit non-canoniques, il est difficile d'envisager sa concaténation avec des données publiques existantes. Pour permettre la mise à disposition des données et valoriser ce travail, nous avons collaboré avec Marilyne Summo (CIRAD, UMR AGAP Insitut) pour développer un site web dédié. Le site « Geloc » a été intégré au *Rice Genome Hub* (https://rice-genome-hub.southgreen.fr/) qui est développé et maintenu par le CIRAD. Ce site a été pensé pour permettre la visualisation, l'exploration et la comparaison des annotations des récepteurs LRR chez le riz.

Ce chapitre sera consacré aux stratégies et outils mis en œuvre pour faciliter l'utilisation et l'exploitation des pipelines et des données obtenues lors de cette thèse.

Dans une première partie seront présentées les stratégies utilisées pour faciliter l'utilisation et améliorer la reproductibilité, l'adaptabilité et la portabilité des deux pipelines développés au cours de cette thèse : LRRprofiler et LRRtransfer. Une partie de ces développements sur le pipeline LRRtransfer a fait l'objet d'un stage de master 2 bio-informatique réalisé par Thibaud Vicat, que j'ai co-encadré pendant 6 mois au cours de ma dernière année de thèse.

Dans une seconde partie, nous présentons les possibilités offertes par le site Geloc à travers l'exploration d'un cluster étudié récemment par Mizuno *et al.* (Mizuno et al., 2020). Ce cluster est localisé sur la fin de bras long du chromosome 11. Il comporte 26 NLR et 12 LRR-RLK englobant les gènes de résistance NLR appelés Pikm1 et Pikm2 (Ashikawa et al., 2008), et LRR-RLK appelé Xa3/Xa26 (Sun et al., 2004).

Ces développements s'inscrivent dans une démarche incontournable d'accès libre aux données et aux outils sachant que les analyses de données en génomique, et en biologie de façon générale, manquent souvent de reproductibilité (Goodman et al., 2016; Tommaso et al., 2017).

B. Des outils d'annotation libres, transparents et réutilisables

B.1. Un code libre disponible sous Git

Le code des deux pipelines principaux développés au cours de cette thèse est librement accessible sur GitHub (https://github.com/cgottin/). Git est un gestionnaire de version libre, facile d'utilisation et permettant le travail collaboratif sur un même projet. Les gestionnaires de version permettent de sauvegarder et de suivre les modifications de scripts effectuées sur un projet. Cela offre une traçabilité du travail au cours du temps, en conservant en mémoire les versions antérieures du projet et les auteurs de chaque modification dans le cadre d'un projet mené en collaboration. GitHub est un site web servant de dépôt permettant de stocker et partager le code versionné via Git.

B.2. Des pipelines génériques et faciles à prendre en main

B.2.1. Un gestionnaire de workflow pour l'adaptabilité et l'automatisation

Les gestionnaires de « flux de données » (appelé *workflow manager* en anglais) permettent l'exécution autonome et reproductible des analyses de données. L'utilisation des gestionnaires de workflow facilite la division des analyses en différentes tâches indépendantes appelées « processus » offrant ainsi une plus grande lisibilité des scripts. Ils ont été développés pour répondre spécifiquement aux problèmes de reproductibilité, d'adaptabilité et de performances des analyses qui sont particulièrement présents en biologie. SnakeMake et Nextflow sont les deux gestionnaires les plus utilisés en bio-informatique et offrent des fonctionnalités très proches. Le développement de pipelines avec l'un de ces gestionnaires apporte plusieurs avantages. Premièrement, le gestionnaire de workflow gère automatiquement la parallélisation des processus qui ne sont pas interdépendants. Ensuite, il garde en mémoire les intermédiaires des différents processus et permet de ré-exécuter, de façon automatique, une analyse depuis la dernière étape sans erreur. Enfin, il s'adapte facilement à différents environnements de travail grâce à un fichier de configuration simple permettant ainsi de passer facilement de l'utilisation du pipeline sur un PC de bureau, à son déploiement sur différents clusters de calcul.

Les outils développés lors de la thèse n'utilisaient initialement pas de gestionnaire de workflow. Les besoins de maintenabilité et de reproductibilité nous ont amenés à faire évoluer en priorité le pipeline de transfert d'annotation, développé en chapitre 2 de cette thèse

(Gottin et al., 2021). L'optimisation de l'outil était l'objectif du stage de Thibaud Vicat (master 2 bio-informatique). Pour ce stage, différentes améliorations étaient envisagées afin d'optimiser le temps d'exécution du pipeline et de faciliter sa distribution. Afin de mener à bien ces objectifs, nous avons adapté ce pipeline de transfert d'annotation pour qu'il fonctionne avec Nextflow. Ainsi, l'architecture du pipeline a été revue et séparée en différents processus indépendants gérés par un script *Nextflow* (Figure 25). Par ailleurs, le temps nécessaire à la recherche de régions homologues à nos gènes d'intérêt, initialement réalisée avec l'outil *tblastn* (Altschul et al., 1990), était particulièrement long (entre 5 et 8 heures) et constituait l'étape temporellement limitante. Une étude bibliographique nous a permis d'identifier plusieurs outils alternatifs à BLAST. Une phase de test sur nos données a confirmé que l'outil MMseqs2 (Steinegger and Soding, 2017) pouvait avantageusement remplacer *tblastn* dans notre pipeline puisqu'il conduisait à des transferts d'annotation au moins aussi bons avec des temps de calcul bien plus faibles (Vicat, 2021).

Une dernière limitation au déploiement du pipeline concerne les outils externes nécessaires à son exécution. Chaque déploiement sur un nouvel environnement nécessitait l'installation et la configuration de plusieurs outils annexes, réduisant la portabilité et la reproductibilité de notre pipeline.

B.2.2. Des containers pour la portabilité et la reproductibilité

Les technologies de containerisation permettent de faciliter la portabilité des outils et la reproductibilité des analyses. Ces technologies reposent sur le principe des Machines Virtuelles (VM). Elles consistent à créer un environnement informatique complet (OS et dépendances) et indépendant de la machine réelle sur laquelle l'analyse doit être exécutée. Les containeurs permettent alors de faciliter la diffusion des outils en s'affranchissant des dépendances des programmes. Le développeur recrée l'environnement de développement en intégrant l'ensemble des scripts et programmes, dans les bonnes versions, nécessaires au fonctionnement de l'outil. Par ailleurs, l'utilisation d'un containeur aide à la reproductibilité des analyses en stabilisant l'environnement d'exécution entre les différents utilisateurs et assure la stabilité des outils dans le temps en s'affranchissant des éventuelles modifications de scripts et mises à jour de programmes.

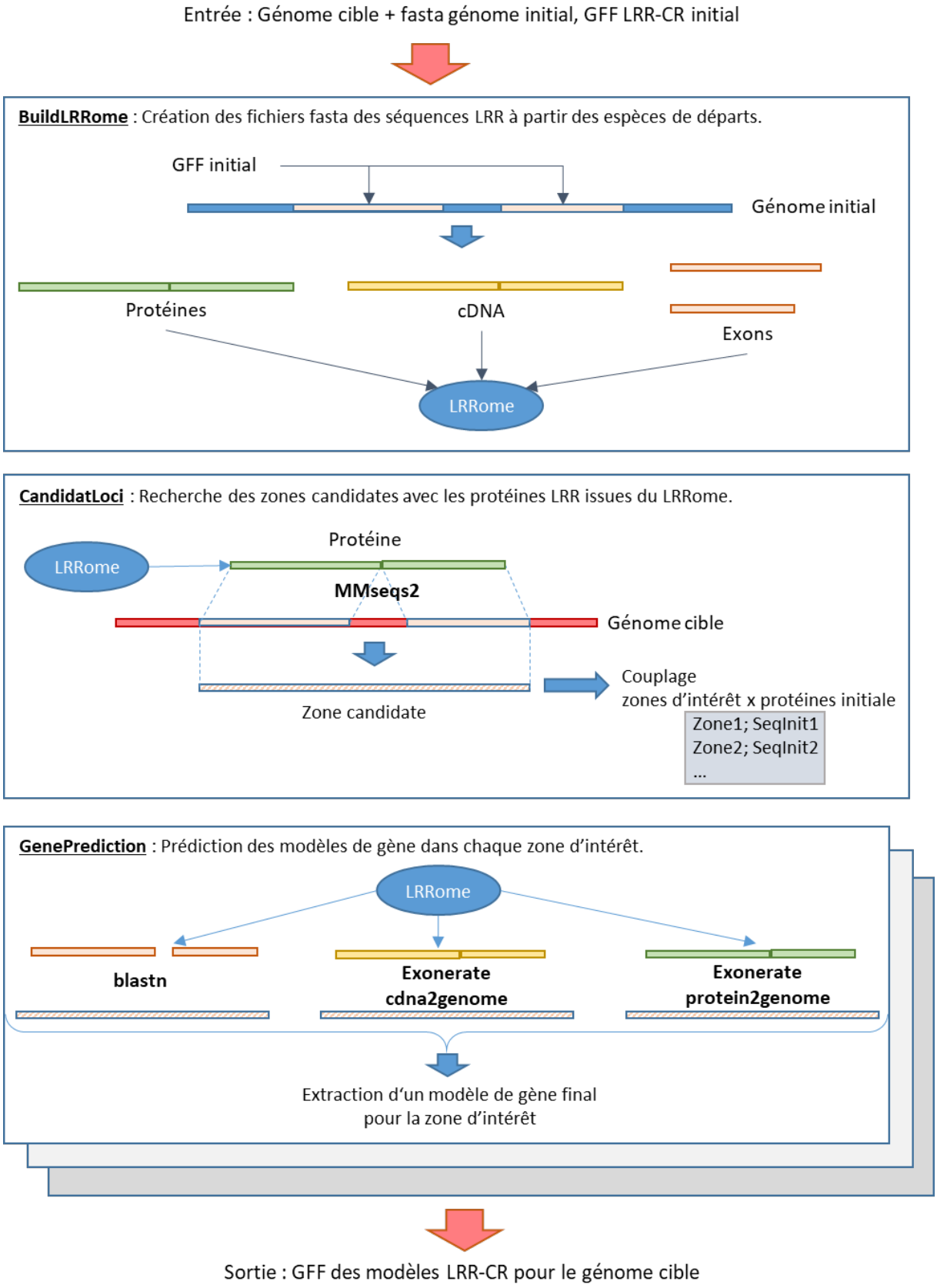


Figure 25 : Représentation schématique du pipeline de transfert d'annotation intégré à Nextflow.
Le pipeline est divisé en trois processus dont le dernier (GenePrediction) est parallélisé pour chaque région d'intérêt de façon autonome par Nextflow.

Il existe différentes technologies de containerisation, dont les deux principales sont Docker et Singularity. Pour la diffusion de nos outils, le pipeline « LRRprofiler » (cf. chapitre 1) et le pipeline de transfert d'annotation « LRRtransfer » (cf. chapitre 2), nous avons choisi de les mettre à disposition sous forme de containeur Singularity sur le cloud Sylabs (https://cloud.sylabs.io/library/cgottin).

B.2.4. Paramètres obligatoires et documentation

Afin de faciliter la prise en mains des outils, les pipelines sont rendus accessibles sur GitHub avec une documentation et des fichiers d'exemples. La documentation guide l'utilisateur pour la récupération des pipelines et des fichiers, et fournit les commandes pour leur exécution. Un descriptif des paramètres obligatoires et facultatifs est également donné. Ces paramètres ont été pensés pour être limités au maximum pour assurer une utilisation facile des pipelines développés au cours de la thèse. Le pipeline LRRprofiler prend deux paramètres obligatoires : le fasta du protéome à annoter et un nom pour le process. Pour le pipeline de transfert d'annotation, les paramètres ont également été revus pendant le stage pour ne prendre que trois paramètres obligatoires (contre 6 initialement) : le fichier fasta du génome à annoter, le fichier fasta du génome initial et le fichier GFF contenant les annotations des LRR du génome initial. Pour limiter ce nombre de paramètres initiaux, une procédure a été ajoutée en amont du pipeline pour générer automatiquement les 3 autres fichiers fasta (protéines, CDS et exons du LRRome à transférer) qui devaient initialement être fournis par l'utilisateur.

B.3. Perspectives

B.3.1. LRRprofiler

Le pipeline LRRprofiler produit une annotation des LRR à partir d'un ensemble de HMM publics (de la base de données SMART) et de deux HMM reconstruits spécifiquement pour s'adapter aux données de chaque accession. Cette stratégie pourrait introduire des biais dans le cadre d'une analyse comparative des annotations de plusieurs accessions (chaque accession étant annotée en partie avec ses propres HMM). Nous n'avons malheureusement pas eu l'occasion de tester si l'utilisation de profils différents pouvait impacter des analyses comparatives. Pour éviter ce problème, on pourrait envisager l'utilisation d'un même ensemble de profils HMM qui seraient ajustés globalement à l'ensemble des accessions étudiées. Nous voulons proposer

alors la possibilité pour l'utilisateur de fournir des HMM personnels et d'esquiver la procédure itérative de reconstruction des HMM spécifiques.

Nous envisageons également d'intégrer la détection de domaines et motifs supplémentaires, notamment les domaines Coiled-Coil (CC) et les motifs LRR-NT et LRR-CT. Les motifs LRR-NT et LRR-CT sont trouvés chez les récepteurs LRR-RLK et LRR-RLP. Ils correspondent à des motifs LRR légèrement divergents par rapport aux motifs LRR classiques et sont présents en amont (LRR-NT) et en aval (LRR-CT) du domaine LRR. D'un point de vue structural, ces motifs forment des « chapeaux » au début et à la fin du domaine LRR. Le domaine CC est trouvé en amont du domaine NB-ARC de certains récepteurs NLR. Différents outils dédiés à la détection des domaines CC existent. Nous avons testé COILS (Lupas et al., 1991) mais son utilisation ralentissait fortement l'exécution du pipeline. Nous envisageons de tester d'autres outils et de comparer leurs performances pour la détection des domaines CC chez les NLR. Une mise à jour du pipeline pourra alors être réalisée pour intégrer la détection de ce domaine.

B.3.2. Transfert d'annotation

Le pipeline de transfert d'annotation réalise une prédiction des modèles de gène dans le génome cible selon 3 stratégies différentes : (i) un *mapping* des exons via blastn, (ii) une prédiction via exonerate *cdna2genome* et (iii) une prédiction via exonerate *protein2genome*. Initialement, le pipeline renvoyait la première prédiction qui passait certains seuils d'identité et de couverture avec la protéine initiale. Pendant le stage réalisé sur le pipeline, le mode de prédiction a été revu pour renvoyer le meilleur modèle parmi les trois stratégies de prédiction. Nous envisageons d'ajouter un mode de prédiction permettant de combiner les trois modèles fournis par les trois stratégies lorsque ces modèles présentent des variations. Cette stratégie nécessite encore des réflexions sur la façon de combiner efficacement différentes annotations pour un même locus.

C. Un outil convivial dédié à l'exploration et la visualisation des LRRome

C.1. Visualisation de la répartition des gènes LRR

Les récepteurs LRR évoluent par duplication, principalement en tandem, et forment des clusters pouvant comporter plusieurs dizaines de gènes. Le site Geloc implémente une visualisation du caryotype des génomes d'intérêt avec la densité des gènes LRR. Cette densité

est représentée par un gradient de couleur permettant d'identifier rapidement les régions concentrant de grands clusters et d'appréhender l'hétérogénéité de répartition des gènes d'intérêt dans le génome du riz (Figure 26a). Les trois familles de gènes sont identifiées par des marqueurs de couleur différentes permettant d'avoir une vision rapide de la composition des clusters, en particulier des clusters mixtes. On peut voir, par exemple, une concentration très importante de gènes LRR au niveau du chromosome 11, avec un nombre très important de NLR. Ce chromosome porte notamment le cluster d'intérêt étudié par Mizuno *et al.*

C.2. Visualisation des annotations des gènes

Les trois familles d'intérêt présentent des structures protéiques différentes avec des domaines spécifiques. Geloc implémente une visualisation combinée de l'annotation des domaines et des structures intron-exon pour chaque gène. La sélection d'un chromosome ouvre une visualisation spécifique horizontale du chromosome avec une fenêtre de sélection mobile permettant de visualiser les annotations des gènes de cette zone (Figure 26b). Cette visualisation s'ouvre pour l'ensemble des locus annotés de la zone sélectionnée permettant par exemple de s'intéresser de manière interactive à la conservation (ou variabilité) des structures observées pour les gènes de cette zone. La vue des locus est accompagnée de plusieurs informations (Figure 27) : l'identifiant du locus, la famille et la classe du gène (canonique ou non), le sens de lecture du gène, i.e. brin sens (+) ou brin anti-sens (-). La présence de mutations non-sens de type stop et *frameshift* est indiquée de manière visuelle par une étoile rouge et une flèche orange respectivement.

Sur l'ensemble du cluster étudié, Mizuno *et al.* identifient 10 gènes non-canoniques (8 NLR et 2 LRR-RLK). La correspondance entre les annotations fournies dans ce papier et les données disponibles sur Geloc est donnée en annexe 3. Dans les données expertisées à partir de la référence Nipponbare, on identifie 20 gènes non-canonique dans le même cluster (12 NLR et 8 LRR-RLK). Notamment, le gène Xa3/Xa26 (Sun et al., 2004) est donné canonique dans la publication, mais sa structure est non-canonique dans notre ré-annotation du génome de référence de Nipponbare. On peut observer la présence d'un *frameshift* avant le domaine LRR et d'un codon stop au début de dernier exon. Par ailleurs, la structure intron-exon de ce locus diffère des structures observées pour les autres loci du cluster. On observe un intron entre le domaine TM et le domaine kinase qui n'est pas présent chez les autres LRR-RLK (Figure 27).

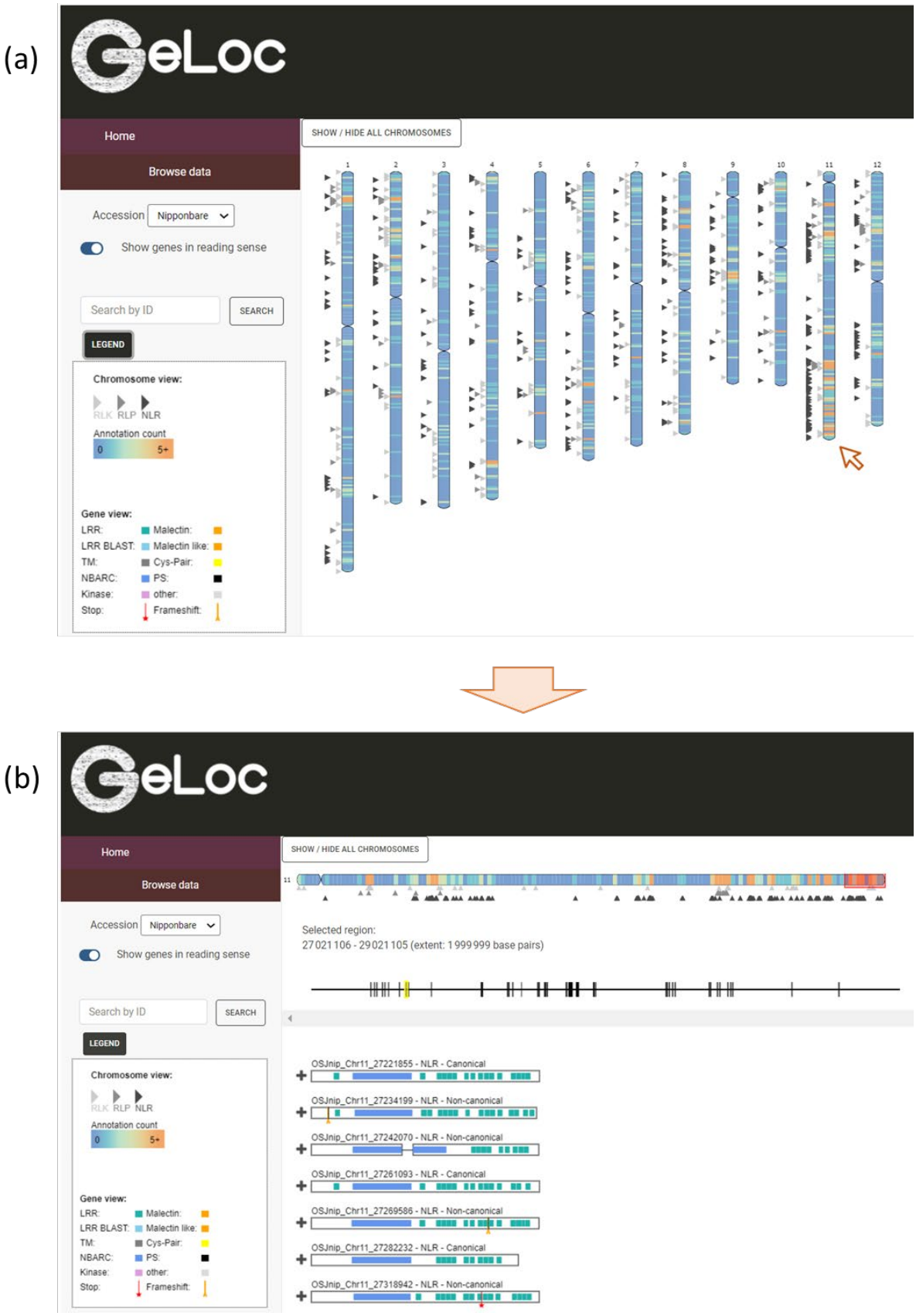


Figure 26 : Visualisation du génome de Nipponbare dans Geloc et localisation du cluster de gènes d'intérêt.
(a) Vue globale du génome permettant la sélection d'un chromosome par simple clic. (b) Vue centrée sur un chromosome dont une zone peut être sélectionnée pour obtenir une visualisation détaillée de la structure des gènes d'intérêt qu'elle contient.

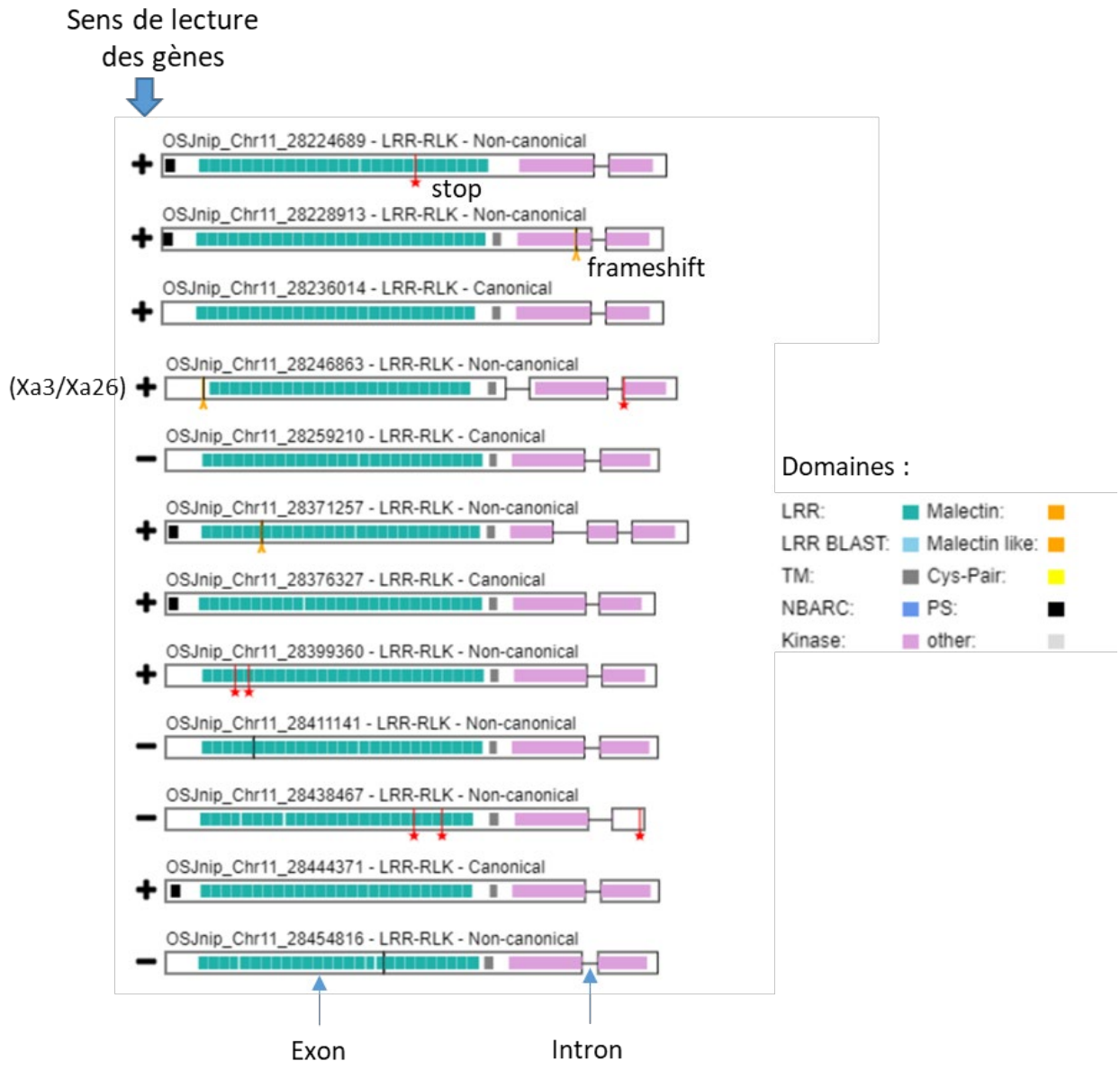


Figure 27 : Visualisation du cluster Xa3/Xa26 comportant 12 LRR-RLK chez Nipponbare.

Ces différences entre la séquence publique de Nipponbare et les données séquencées par Mizuno *et al.* pourraient être dues au fait que leur séquence diffère de celle de la référence publique. Or, un alignement BLAST réalisé entre les deux séquences nucléotidiques montrent qu'elles sont 100% identiques au niveau du locus, invalidant cette hypothèse. La classification du gène en canonique ou non-canonique provient donc uniquement d'une divergence d'annotation. Mizuno *et al*. utilisent le pipeline d'annotation développé par l'IRGSP. Leur annotation semble sous-estimer le nombre de modèles non-canoniques (10 dans la zone étudiée, contre 20 pour nous). En se basant sur la comparaison de ce cluster entre trois couples d'espèce cultivées et sauvages, ils concluent que la domestication a augmenté le

nombre de gènes de résistance dans les cultivars. Si la plupart des copies sont effectivement non-canoniques (comme le suggère notre annotation pour Nipponbare), l'interprétation biologique de Mizuno *et al.* peut être questionnée car leur conclusion repose directement sur l'augmentation du nombre de gènes (fonctionnels) de résistance dans les formes cultivées par rapport aux sauvages.

C.3. Identification des gènes orthologues

Pour chaque locus, Geloc fourni un ensemble d'information (« *gene card* ») sur le bandeau gauche du site lorsqu'un gène est sélectionné (Figure 28a). La *gene card* fournie les identifiants du locus et une correspondance avec les identifiants des annotations publiques (IRGSP, MSU et NCBI pour Nipponbare). Ainsi le locus OSJnip_Chr11_28246863 correspondant au gène Xa3/Xa26 a été annoté par MSU (LOC_Os11g47000) et le NCBI (LOC4351164) mais n'est pas annoté par IRGSP. La *gene card* fournit également les relations d'orthologie entre Nipponbare et KitaakeX identifiées lors du transfert d'annotation sur lequel s'appuie notre ré-annotation de KitaakeX. Un lien cliquable au niveau de l'identifiant permet d'accéder directement à l'annotation de l'orthologue dans l'autre génome (Figure 28b).

L'orthologue du gène Xa3/Xa26 chez KitaakeX est également non-canonique. En revanche, les mutations non-sens responsables ne sont pas les mêmes que celles observées pour Nipponbare, à l'exception du *frameshift* identifié au début du gène. L'orthologue de KitaakeX présente un *frameshift* supplémentaire dans le dernier exon et on ne retrouve pas le codon stop au début du dernier exon. Par ailleurs, l'intron entre le domaine TM et le domaine kinase n'est pas présent non plus.

C.4. Récupération des LRRomes selon différents points de vues

Les données ré-annotées sont fournies librement sous différents formats standards. Les annotations des loci sont données au format gff3. Les séquences des loci sont téléchargeables au format fasta. Trois types de séquences sont accessibles : (i) la séquence des gènes complets avec exons et introns sans interruption ; (ii) les séquences « cDNA » correspondant aux séquences nucléotidiques après élimination des introns ; (iii) les séquences peptidiques issues de la traduction des séquences CDS.

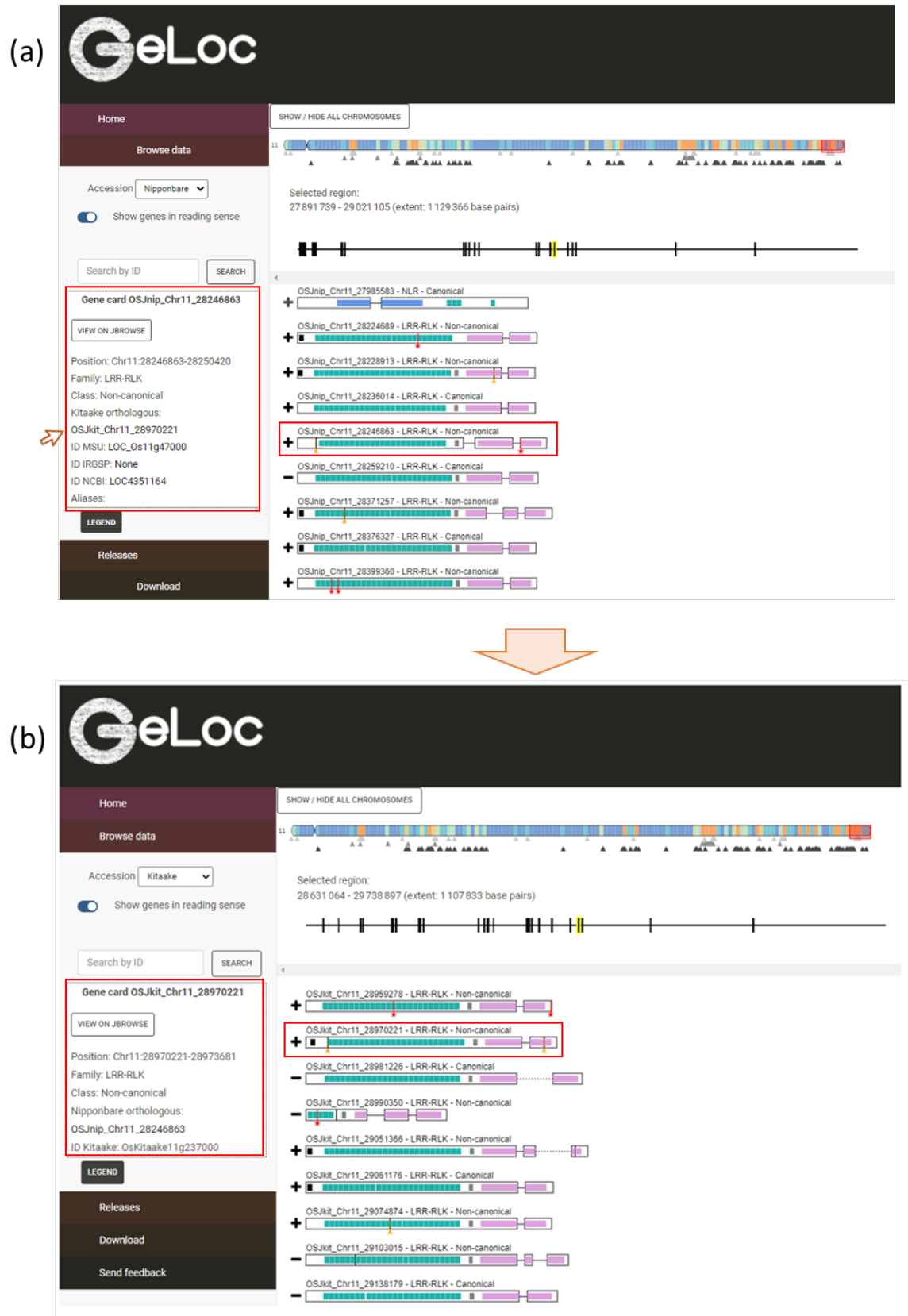


Figure 28 : Visualisation des gènes d'intérêt dans Geloc, de leurs informations (« gene card ») et des relations d'orthologie entre les différentes accessions.

(a) Cliquer sur la structure d'un gène d'intérêt permet d'afficher l'ensemble des informations disponibles pour ce gène dans le bandeau gauche de site. (b) Cliquer sur l'identifiant du gène orthologue fournit par la gene card permet de passer à la visualisation du gène correspondant chez l'autre accession.

Certaines séquences sont impactées par des mutations non-sens de type *frameshift*. Ces mutations reflètent la présence de codons incomplets détectables par la présence d'indels non multiple de 3 dans l'alignement de cette séquence avec une séquence homologue canonique proche. Si un *frameshift* existe dans une séquence alors la traduction directe de cette séquence codante conduira à une séquence peptidique erronée à partir du premier *frameshift*. En effet, à partir du *frameshift*, la traduction ne sera plus faite dans le bon cadre de lecture. Pour documenter ces évènements, nous avons utilisé la convention du logiciel MACSE (Ranwez et al., 2011; Ranwez et al., 2018) qui code les *gaps* au sein de codons incomplets par un caractère particulier, le '!'. Selon l'objectif, on peut alors choisir de considérer la séquence brute (nommée cDNA pour '*coding DNA'*, dont la traduction conduira à une séquence erronée à partir du premier *frameshift*) (Figure 29a), une version incluant en plus les caractères '!' permettant de conserver le cadre de lecture (nommées **'**cDNA_wFrameshift' permettant de repérer la position des *frameshifts* dans la séquence) (Figure 29b) et une version alternative excluant les zones non-homologues contenant des *frameshifts* (CDS, équivalente à 'cDNA_wFrameshift', mais sans le(s) codon(s) possédant les caractères '!') (Figure 29c). Ainsi, nous proposons de télécharger trois versions distinctes des séquences codantes pour chaque gène contenant au moins un *frameshift*.

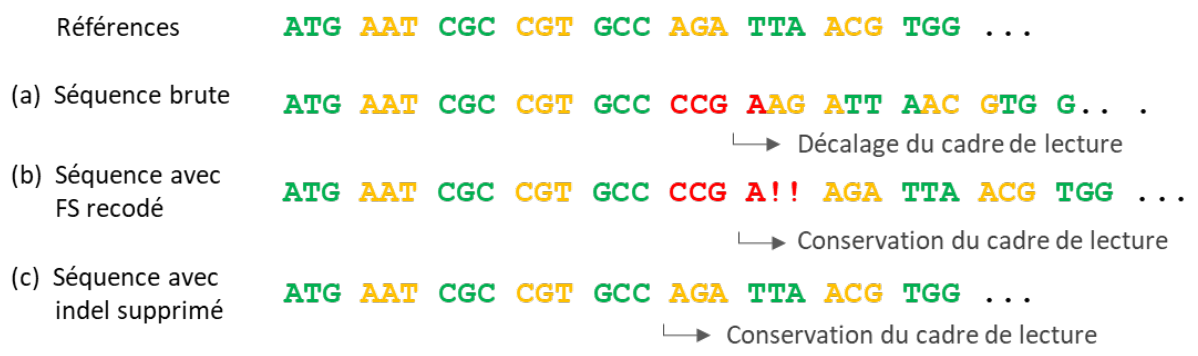


Figure 29 : Les différentes façon de représenter une séquence codante affectée par un frameshift et leurs conséquences sur la traduction en protéine.
(a) Représentation de la séquence brute sans modification telle que présente dans le génome. Sa traduction conduit à une séquence protéique erronée à partir du frameshift. (b) Représentation de la séquence dans laquelle sont insérés un ou deux ' !' selon la convention MACSE, pour rétablir le cadre de lecture initial. La traduction de la séquence conduit à une insertion dont l'un des acides aminés est ambiguë (codé par un « X »). (c) Représentation de la séquence dont l'indel a été exclu permettant de conserver le cadre de lecture.

C.5. Perspectives

Plusieurs perspectives de mises à jour du site Geloc sont en cours de réflexion. La première concerne la possibilité d'extraire les données à partir d'un système de requête. Ce système permettrait aux utilisateurs de récupérer une partie des données répondant à certains critères d'intérêt. La requête devrait s'effectuer sur les critères suivants : l'accession, les chromosomes, la position sur les chromosomes, les familles de gène (LRR-RLP, LRR-RLK et NLR) et les classes de gène (canoniques ou non). Le système devrait également permettre d'extraire les orthologues des gènes sélectionnés chez les autres espèces disponibles, ou de récupérer les différentes séquences associées à chaque gène (nucléique, protéique, avec ou sans *frameshifts*).

La seconde perspective d'amélioration concerne la visualisation comparative des orthologues. Cette visualisation devrait s'effectuer à deux échelles : (i) au niveau chromosomique permettant d'étudier la colinéarité et la présence/absence des copies entre espèces, et (ii) au niveau des gènes permettant d'étudier la conservation des structures (introns/exons et domaines). Par ailleurs, nous avons travaillé à la ré-annotation des récepteurs LRR pour l'espèce *O. rufipogon* (W1943), correspondant à la forme sauvage dont sont issus les cultivars *japonica*. Ces données seront prochainement ajoutées sur le site Geloc en complément de celles de Nipponbare et KitaakeX.

Chapitre 4

Chapitre 4 :
Origine et diversité des *Receptor-Like Kinases* chez les plantes

Table des matières

A. Introduction

Chez les plantes, les gènes *Receptor-like Kinase* (RLK) constituent la principale famille de récepteurs transmembranaires avec notamment plus de 600 membres identifiés chez *Arabidopsis thaliana* (Shiu and Bleecker, 2001b, a). Ces récepteurs sont composés d'un domaine kinase sur leur section intracellulaire qu'on retrouve également chez les bactéries et les animaux. L'analyse phylogénétique des domaines kinases montre que les kinases des plantes forment un groupe monophylétique. Par ailleurs, ces domaines kinases des plantes présentent des homologies avec les gènes Pelle (Drosophile) (Hanks and Hunter, 1995) et IRAK (mammifères) (Shiu and Bleecker, 2001a) supposant une origine des kinases antérieure à la séparation entre les plantes et les animaux (Hanks et al., 1988; Hanks and Hunter, 1995; Shiu and Bleecker, 2001a). En 2001, Shiu et Bleecker regroupent ces gènes en un seul clade nommé « RLK/Pelle » (Shiu and Bleecker, 2001a). Leurs comparaisons du nombre de gènes kinase de la famille RLK/Pelle entre plantes et animaux suggèrent une expansion massive des gènes kinase chez les plantes, postérieure à la séparation entre plantes et métazoaires (Shiu and Bleecker, 2001a, 2003). Les RLK des plantes présentent une grande diversité de domaines associés au domaine kinase. La forte homologie du domaine kinase entre ces différents récepteurs suggère que les associations entre les domaines extracellulaires et le domaine kinase ont eu lieux à partir d'une kinase ancestrale (Shiu and Bleecker, 2003). La formation de nouvelles structures protéiques via un réarrangement des domaines est un mécanisme connu de l'évolution des protéines et de l'émergence de nouvelles fonctions (Apic et al., 2001; Bjorklund et al., 2005).

Le travail de revue qui est présenté dans ce chapitre, publié dans le journal *Annual review of Plant Biology* en 2020 et dont je suis co-premier auteur (Dievart et al., 2020), est consacré à l'étude de ces arrangements de domaines avec le domaine kinase. Plus particulièrement, nous avons adressé différentes questions : (i) A quel moment au cours de l'évolution des plantes les fusions de domaines sont-elles apparues ? (ii) Les associations observées chez les plantes ont-elles eu lieu une seule fois ou plusieurs fois au cours de l'évolution ?

Pour y répondre, nous avons combiné une revue bibliographique et des résultats originaux obtenus sur 194 génomes de plantes. Les génomes ont été analysés avec InterProScan (Jones et al., 2014). L'ensemble des prédictions de domaines pour une protéine a été concaténé et filtré pour éliminer les redondances. Cette étape est nécessaire car InterProScan utilise

différents outils de prédiction en parallèle (SMART, Pfam, Phobius, TMHMM, PANTHER, CATH-gene3D, etc.) pouvant générer des prédictions multiples pour un même locus. De plus, certaines prédictions ont dû être renommées afin d'uniformiser les noms de domaines. Par exemple, le peptide signal apparait sous différents identifiant dans les sorties brutes d'InterProScan : SIGNAL_PEPTIDE, SignalP-TM, SignalP-noTM. Ensuite, chaque architecture unique a été recensée et dénombrée pour l'ensemble des espèces. Les résultats ont finalement été agrégés par architecture et par groupe phylogénétique. Nous avons identifié, par exemple, 42 467 structures 'LRR+kinase' dans les 194 génomes étudiés, dont 10 011 chez les monocotylédones.

B. Résumé des résultats

- Les fusions entre les différents domaines fonctionnels et la kinase sont apparues à différents moments de l'histoire évolutive des plantes et peuvent être associés à des étapes marquantes de leur évolution comme par exemple le passage du milieu aquatique au milieu terrestre. Il est néanmoins difficile de dire si ces nouvelles associations, en créant de nouvelles fonctions, ont participé ou non à l'adaptation des plantes.
- L'association des LRR avec le domaine kinase aurait été présente chez l'ancêtre des streptophytes et serait apparue il y a plus de 1 000 Ma (million d'années). Les associations des LRR avec les domaines malectin et malectine-like sont plus tardives et auraient émergé il y a plus de 420 Ma et plus de 600 Ma respectivement.
- Les différentes associations auraient émergé une seule fois chacune au cours de l'évolution des plantes. Mais pour certaines associations peu représentées dans les protéomes (C-LEC-kinase, CRINKLY4-kinase notamment), on observe une histoire complexe qui peut s'expliquer de deux façons différentes : (i) il y a eu plusieurs émergences de la même association dans différentes branches ou (ii) il y a eu une seule émergence ancestrale suivie d'une perte de la structure dans certaines branches. Les ambiguïtés étant présentes pour des espèces peu étudiées, il n'est pas possible de dire si les absences de certaines associations sont effectives ou dues à des erreurs d'annotation. Il sera nécessaire d'obtenir et d'étudier les génomes de plantes positionnées à des points charnières de la phylogénie pour répondre à cette question.

C. Article de revue : *'Origin and diversity of plant receptor-like kinases'*

ANNUAL REVIEWS

Annual Review of Plant Biology

Origin and Diversity of Plant Receptor-Like Kinases

Anne Dievart,[1,2,*] Céline Gottin,[1,2,*] Christophe Périn,[1,2] Vincent Ranwez,[2] and Nathalie Chantret[2]

[1]CIRAD, UMR AGAP, F-34398 Montpellier, France; email: anne.dievart@cirad.fr

[2]AGAP, Université de Montpellier, CIRAD, INRA, Montpellier SupAgro, F-34060 Montpellier, France

Annu. Rev. Plant Biol. 2020. 71:131–56

First published as a Review in Advance on March 18, 2020

The *Annual Review of Plant Biology* is online at plant.annualreviews.org

https://doi.org/10.1146/annurev-arplant-073019-025927

*These authors contributed equally to this article

Keywords

receptor, kinase, plant, diversity, origin, domain, RLK, RK

Abstract

Because of their high level of diversity and complex evolutionary histories, most studies on plant receptor-like kinase subfamilies have focused on their kinase domains. With the large amount of genome sequence data available today, particularly on basal land plants and Charophyta, more attention should be paid to primary events that shaped the diversity of the RLK gene family. We thus focus on the motifs and domains found in association with kinase domains to illustrate their origin, organization, and evolutionary dynamics. We discuss when these different domain associations first occurred and how they evolved, based on a literature review complemented by some of our unpublished results.

Contents

1. INTRODUCTION

Protein phosphorylation is a biochemical posttranslational modification involved in all signaling pathways and cellular activities in living organisms. It influences cell signaling networks in response to cellular or environmental stimulation, e.g., by the reversible regulation of protein functions through their activation or inactivation, the formation of protein complexes, and the arrangement of subcellular protein locations (31). The phosphorylation process, involving the transfer of a phosphoryl group from adenosine triphosphate (ATP) (or other nucleoside phosphates) to an acceptor hydroxyl residue of the protein substrate, is one of the roles ascribed to well-conserved protein kinase domains (48). Note that serine/threonine (Ser/Thr) phosphorylation is the most common phosphorylation event. Besides their role in this catalytic function, kinase domains also facilitate the binding and orientation of the ATP phosphate donor and protein substrate.

ATP: adenosine triphosphate

Ser/Thr: serine/threonine

RTK: receptor tyrosine kinase

RLK: receptor-like kinase

RK: receptor kinase

Networks of protein kinases involved in substrate phosphorylation are highly complex and consist of hundreds of proteins, making them one of the largest protein families (95). In 2002, Manning and coauthors (97) proposed the term kinome to encompass the complete set of protein kinases encoded in a genome. The kinome of the first plant genome sequenced in 2000 (*Arabidopsis thaliana*) included more than 1,000 proteins (2). Several years later, with more genomes available, the same analysis revealed that the plant kinome superfamily represented ~1–4% of the protein coding genes in 16 genomes (77). In plants, among the variety of protein kinases, the largest subfamily, besides those with functional homologs in other kingdoms (~400 proteins), consists of singular proteins with structural organization resembling that of receptor tyrosine kinases (RTKs) in animals. In 1993, Walker (153) introduced the term receptor-like kinases (RLKs) to account for some of the new genes he had cloned. This terminology is now used widely to define the whole subfamily of these plant receptors. Recent publications have also begun to name those that have a well-defined receptor function receptor kinases (RKs).

The first to promote the concept of receptor in the late nineteenth and early twentieth centuries were the immunologist Ehrlich and the physiologist Langley (94). Working on toxins and ferments, Ehrlich suggested the existence of molecules in body cells—initially referred to as side-chain and then receptor molecules—that need to fix substances in order to be biologically active (126). In parallel, another receptor theory introduced by Langley to explain antagonistic drug action in the body proposed the existence of receptive substances with chemical affinities for specific compounds (93). The questioning and controversy about the receptor theory ended in 1948 when the pharmacologist Ahlquist (1) published a paper describing two types of adrenergic receptors. The fundamental nature and structure of these receptors remained hypothetical until the 1980s. At this time, the first structurally related receptors from the RTK family were cloned. They are specific cell surface receptors for several growth factors, such as epidermal growth factor (EGF) or insulin, and their protein kinase activity was found to be intrinsic to receptor polypeptides and not part of a separate effector system (39, 145, 146). All of these proteins contained a large glycosylated extracellular ligand-binding domain (ECD), a hydrophobic domain that appeared to constitute a typical transmembrane (TM) domain, and an internal kinase catalytic domain (118). In plants, the first RLK, which is structurally similar to these RTKs, was discovered only 5 years later in 1990 (154).

EGF: epidermal growth factor

ECD: extracellular domain

TM: transmembrane

cDNA: complementary DNA

AA: amino acid

SP: signal peptide

2. A SHORT HISTORY OF THE DISCOVERY AND DIVERSITY OF RECEPTOR-LIKE KINASE SUBFAMILIES

It was assumed for decades that most cell-to-cell communication in plants would occur via cytoplasmic bridges called plasmodesmata because of the presence of cell walls separating plant cells from one another. Publication of the paper by Walker & Zhang (154) was thus a landmark event since it introduced a new paradigm for the mechanism by which cells communicate in plants. Although structurally resembling animal RTKs, plant RLK subfamilies underwent highly complex evolutionary histories due to considerable gene expansions (an increased number of genes per genome) coupled with diversifications (an increased number of specific combinations of domains) (57, 77, 79). The large number of domains found associated with kinase domains in these RLKs reflects the wide array of all cell communications that plants must handle within themselves and with the outside world, e.g., as pattern recognition receptors during immunity (26, 51, 139).

2.1. Receptor-Like Kinases with Diverse Extracellular Domains

In this section, we pay tribute to the pioneering researchers who discovered the first members of each of the new structural organizations observed in these receptors. Through them, the golden age for the discovery of hundreds of other RLKs in plants and the elucidation of their functions began (**Figure 1**).

2.1.1. S-domain (G-type lectin, B lectin). The first plant RLK was described in 1990, when Walker & Zhang (154) amplified and cloned, from a complementary DNA (cDNA) library of maize roots, the *ZmPK1* gene. This open reading frame, representing an almost complete transcript, encoded an 817–amino acid (AA) protein. This protein contained two hydrophobic regions, an N-terminal signal peptide (SP) of 28 AAs and a TM domain from residues 473 to 498, followed by a stop transfer sequence responsible for the interruption of the translocation process through the membranes. A comparison with the protein kinase known at this time in animals (48), and particularly the presence in ZmPK1 of certain key residues that are highly conserved in the protein kinase catalytic domain, suggested that the ZmPK1 receptor could be classified as a

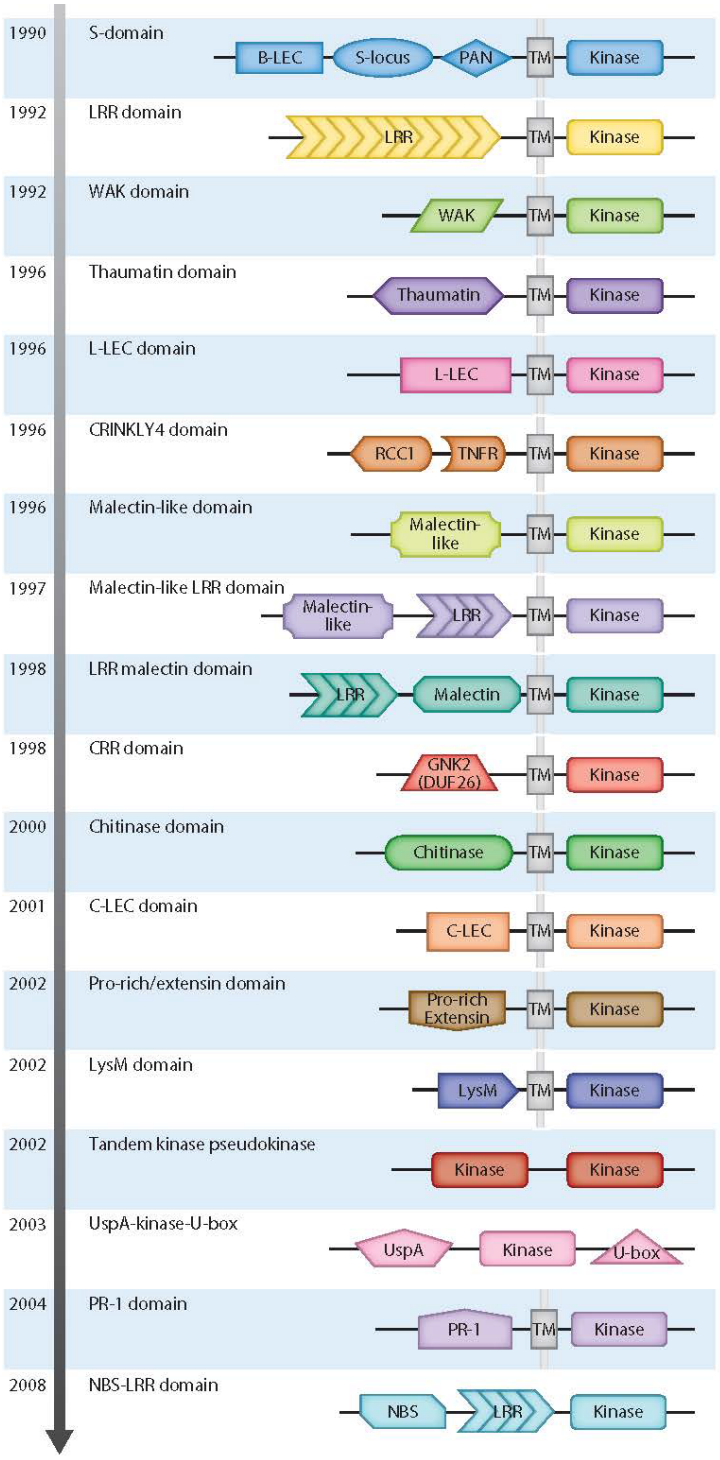


Figure 1

Chronological overview of the major structural motif and domain organizations discovered in plant RLK proteins. Some of these domain organizations are found only in association with specific kinase subgroups. The relationships between these domain organizations and the group types classified according to the kinase phylogeny proposed by Lehti-Shiu & Shiu (77) are provided in **Supplemental Figure 1.** Abbreviations: B-LEC, bulb-type lectin; C-LEC, calcium-dependent lectin; CRR, cysteine-rich repeat; DUF26, DOMAIN OF UNKNOWN FUNCTION 26; GNK2, ginkbilobin2; LEC, lectin; L-LEC, legume lectin; LRR, leucine-rich repeat; LysM, lysin motif; NBS, nucleotide-binding site; PAN, plasminogen/hepatocyte growth factor family/apple domains of the plasma prekallikrein/coagulation factor XI family/domains of various nematode proteins (142); PR-1, pathogenesis-related protein 1; RCC1, regulator of chromosome condensation 1; RLK, receptor-like kinase; TM, transmembrane; TNFR, tumor necrosis factor receptor; UspA, universal stress protein A; WAK, wall-associated kinase.

Supplemental Material >

Ser/Thr kinase. This kinase domain would be intracellular and located downstream of the TM domain. Moreover, in the Ser/Thr kinase subfamily, the closest relative to ZmPK1 would be the human c-raf1 kinase (11), a proto-oncogene related to rapidly accelerated fibrosarcoma (RAF) Ser/Thr kinases, which were initially encountered in viruses (114, 165). The S-domain, in the extracellular part of the protein, has been subdivided into three subdomains: The first, the bulb-type lectin (B-LEC) domain, contains a highly conserved 14-residue stretch, and the third, the cysteine-rich (or PAN) domain, comprises 10–13 cysteine residues clustered near the TM domain (112, 142). Unexpectedly, besides these remarkable residues, the whole ECD showed 52% identity with the ECD of an S-locus-specific glycoprotein (SLG) of *Brassica* (104). These *SLG* genes were known to be among the products of the S-locus that controls the pollen-stigma interaction of self-incompatibility in the Brassicaceae mustard family. Moreover, even though the maize and *Brassica* genes shared some sequence similarities, data showed that they were expressed in different tissues, suggesting that ZmPK1 would not play a role in pollen-stigma recognition.

SLG: S-locus-specific glycoprotein

SRK: S receptor kinase

LRR: leucine-rich repeat

Screening cDNA libraries by using protein kinase genes as hybridization probes under low-stringency conditions was widely used in the 1990s to reveal serendipitously many new RLKs in several plant species. In 1991, Nasrallah's group (134) described the cloning of a new S-domain receptor kinase. The ECD displayed genotype-specific sequence polymorphisms that paralleled those of SLGs and S receptor kinase (SRK) transcripts were detected only in reproductive organs, suggesting that self-recognition between pollen and stigma during pollination in *Brassica oleracea* is mediated by receptor-ligand interactions between pollen and pistil components, leading to pollen acceptance or rejection. In 1992, the same team described the receptor ARK1 (141). Just like ZmPK1, ARK1 was expressed mainly in leaves, ruling out a role in self-incompatibility. Since then, several S-domain receptor kinases have been characterized, and besides their role in reproduction, some play a role in innate immunity, like mediating low-complexity bacterial metabolite sensing in *Arabidopsis* (71), in plant-mycorrhizal interaction (73), or in environmental and developmental processes (8, 109, 110, 117). These G-type lectin receptors [for *Galanthus nivalis* agglutinin (GNA)] belong to the large lectin (carbohydrate-binding) domain RLK family with other subfamilies, such as the L-type (L-LEC), C-type (C-LEC), malectin-like (also known as CrRLK1L), cysteine-rich repeat (CRR), and lysin motif (LysM) RLKs described below.

2.1.2. Leucine-rich repeat domain. In 1991, Bleecker (9) reported the cloning of an *A. thaliana* putative TM protein kinase. One year later, the *TRANSMEMBRANE KINASE 1* (*TMK1*) gene was described in detail by Chang and coauthors (21). In contrast with the previously cloned receptors, the ECD of TMK1 was composed of 11 copies of a 22-AA leucine-rich repeat (LRR) unit, a protein structural motif rich in leucine arranged in tandem repeats. In animals, the Toll receptor, involved in the establishment of dorsoventral polarity in *Drosophila* embryos, most resembled this new plant LRR receptor (50). The structural difference between these two receptors was the presence of the kinase domain in the cytoplasmic region of the plant receptor, which is absent in the Toll receptor [as it contains instead a Toll/interleukin-1 receptor (TIR) domain]. Since the 1990s, there has been enormous progress in determining the functions and mechanisms of activation of some of these receptors. Many functions have been attributed to these receptors, from plant growth and development to symbiosis and immunity. Activation of these receptors is generally based on peptide or hormonal ligand-dependent interactions of LRR receptor heterodimers containing one LRR-RLK with a few LRR units playing the role of coreceptor, such as BRASSINOSTEROID INSENSITIVE1-ASSOCIATED KINASE1 (BAK1) or SUPPRESSOR OF BIR1-1 (SOBIR) (46, 57, 58, 131, 160). However, because the LRR-containing receptor kinase subfamily contains the largest number of genes in all plant genomes studied so far (**Table 1**), many of these genes are still uncharacterized, despite efforts to uncover biologically

Table 1 Average number of proteins per species in monocots and dicots (when present), and percentage of species in which these domains are found

Domains	Average number of proteins per species in monocots (% of species)	Average number of proteins per species in dicots (% of species)
S-domain kinase	51 (96%)	54 (100%)
LRR kinase	189 (100%)	227 (100%)
WAK kinase	62 (100%)	42 (100%)
L-LEC kinase	44 (100%)	36 (100%)
CRINKLY4 kinase	1 (89%)	2 (97%)
Malectin-like kinase (CrRLK1L)	22 (96%)	30 (100%)
Malectin-like LRR kinase	9 (87%)	10 (100%)
LRR malectin kinase	3 (89%)	9 (97%)
CRR kinase	34 (96%)	37 (100%)
Pro-rich/extensin kinase	20 (96%)	14 (100%)
LysM kinase	10 (94%)	5 (100%)
C-LEC kinase	1 (91%)	1 (89%)
UspA-kinase-U-box	6 (96%)	9 (100%)
Thaumatin	2 (64%)	1 (27%)
Chitinase	1 (60%)	2 (62%)
Kinase pseudokinase	6 (92%)	8 (96%)
PR-1 kinase	1 (30%)	1 (5%)

Abbreviations: C-LEC, calcium-dependent lectin; CRR, cysteine-rich repeat; CrRLK1L, *Catharanthus roseus* receptor-like kinase 1-like; L-LEC, L-type lectin; LRR, leucine-rich repeat; LysM, lysin motif; PR-1, pathogenesis-related protein 1; UspA, universal stress protein A; WAK, wall-associated kinase.

relevant interactions within LRR receptors containing a kinase domain or not, the latter being defined as receptor-like protein (RLP) (46, 82, 86, 128).

2.1.3. Wall-associated kinase domain. In 1992, Kohorn and coauthors (65) described a third type of RLK containing several EGF repeats in its ECD. This new protein, PRO25, was renamed WAK1 (for WALL-ASSOCIATED KINASE 1) a few years later (53). Recent discoveries have shown that these receptors are involved in cell wall integrity maintenance mechanisms through the perception of pectin or pectin-derived molecules not only during developmental cell wall extensions but also during stress- or pathogen-induced cell wall damage (5, 64).

2.1.4. Legume lectin domain. In 1996, Herve, Lescure, and coauthors (55) described a protein, *Arabidopsis thaliana* lectin-receptor kinase 1 (Ath.lecRK1), containing an ECD homologous to the carbohydrate-binding proteins of the legume lectin family. Since 1889, extracts of castor beans or many other leguminous seeds were considered capable of agglutinating animal red blood cells via some remarkable proteins called lectins (from the Latin word *legere* for choose) (87, 135). However, despite this homology, all AA residues forming the putative monosaccharide-binding site of the lectin domain of Ath.lecRK1 were different from those usually found in legume lectins, suggesting that these receptors might not possess true lectin activity (14). To date, most studied L-LEC-RLKs have roles in plant immunity, with recent data suggesting that these receptors could bind metabolites such as extracellular adenosine triphosphate (eATP) (24) or extracellular nicotinamide

RLP: receptor-like protein

adenine dinucleotide (eNAD) (155). This latter molecule could be a key component of systemic acquired resistance in *Arabidopsis*, and the binding of eNAD to the L-LEC-RLK requires that this receptor forms a complex with the LRR-RLK BAK1 coreceptor (155, 157).

2.1.5. CRINKLY4 (and tumor necrosis factor receptor-like) domain. The 1995 Cold Spring Harbor Symposium review of Dangl, Preuss, and Schroeder (29) reports prime scientific discoveries including the cloning and functional analyses by Becraft, Stinard, and McCarty (7) of the maize receptor CRINKLY4. The ECD of this RLK was described as containing a novel 37-AA domain repeated seven times (the crinkly domain related to a GTPase-activating protein found in all eukaryotes) and a 26-AA region with similarities to the tumor necrosis factor receptor (TNFR) cysteine-rich region. The fact that this TNFR motif was known to contact the TNF suggests that the ligand for the maize protein might be a peptide. Even if a peptide of the CLE family has been suggested to be a ligand for the ARABIDOPSIS CRINKLY4 (ACR4) receptor in complex with the CLAVATA1 receptor belonging to the LRR-RLK subfamily, no biochemical data supporting this suggestion are available at the moment (28). It has to be noted that, contrary to previously described receptors that are present at several tens of copies per genomes, the CRINKLY4 subfamily is only represented by one or two copies on average per genome (**Table 1**).

2.1.6. Malectin-like (CrRLK1-like) domain. Schulze-Muth, Schröder, and coauthors (119) cloned the first member of the *Catharanthus roseus* receptor-like kinase 1-like (CrRLK1L) subfamily in 1996 during a screen of a library from a Madagascar periwinkle cell culture. None of the previously described ECDs showed similarity to the 450-AA residues composing the N-terminal part of this receptor. In 2011, Boisson-Dernier, Kessler, and Grossniklaus (10) showed that the ECD of these CrRLK1L receptors is homologous to a carbohydrate-binding domain, the malectin-like domain, which seems however to lack residues important for carbohydrate-rich ligand binding (37, 103). Various members of this RLK subfamily, in complex with other CrRLK1L receptors and glycosylphosphatidylinositol-anchored proteins chaperons/coreceptors, are involved in developmental or immune cell-wall-sensing responses induced by RAPID ALKALINIZATION FACTOR (RALF) peptides (43, 107, 161).

2.1.7. Malectin-like leucine-rich repeat domain. A malectin-like domain has also been found in the extracellular part of the receptors in association with LRRs. The light-repressible receptor protein kinase (LRRPK) is the first receptor of this type cloned in 1997 by Deeken & Kaldenhoff (33). The presence of a malectin-like domain followed by a few LRRs in the ECD of an RLK was described for the first time for the IMPAIRED OOMYCETE SUSCEPTIBILITY1 (IOS1) receptor in 2011 by Hok, Keller, and coauthors (59). It has recently been established that the CrRLK1L FERONIA and IOS1 function as scaffolds to regulate LRR-RLK immune receptor complexes (133, 164). Another well-studied receptor with this organization is the SYMBIOSIS RECEPTOR-LIKE KINASE (SYMRK), which was cloned in 2002 by Stracke, Parniske, and coauthors (136). For this receptor, named DOES NOT MAKE INFECTIONS2 (DMI2) in *Medicago truncatula*, the protein level paced by protection from proteasome-mediated degradation could be a regulator of bacteria–plant and fungi–plant symbioses (108, 151).

2.1.8. Leucine-rich repeat malectin domain. In the LRR malectin domain, LRRs are located upstream in the extracellular part of the receptor, and the first gene of this type was described in 1998 by Takahashi, Chua, and coauthors (137). The extracellular part of receptor-like kinase

in flowers 1 (RKF1) possesses 13 LRRs. The presence of the malectin domain, located between LRRs and the TM domain in this receptor, was discovered later (38). The primary sequence difference between the malectin-like and malectin domains suggests that these two domains could have distinct ligand affinities. To our knowledge, scarce data are available on these LRR malectin RLKs in the literature, except that one of these receptors conferred enhanced resistance to the fungal pathogen *Magnaporthe oryzae* when overexpressed in rice (84).

2.1.9. Cysteine-rich repeat domain. Takahashi et al. (137) cloned a second gene, *RKF2*. In this 617-AA protein, the ECD has two copies of a C-X8-C-X2-C motif, distinct from the cysteine-rich region of SLGs and SRKs but highly similar to fungal lectins, suggesting either a common origin or convergent evolution of these domains (147). A few years later, the receptors possessing these CRR motifs were named CRKs, for CRR RLKs (22). These motifs are also named GINKBILOBIN2 (GNK2) [e.g., in the InterPro database (32, 102)] or DOMAIN OF UNKNOWN FUNCTION 26 (DUF26) (in the literature). The latest research on this receptor subfamily suggests that they play roles in defense and oxidative stress responses, programmed cell death, and development (13, 75, 111, 147).

2.1.10. Pro-rich/extensin domain. In 2002, Silva & Goring (125) described and characterized a new gene from *Brassica napus*: *PROLINE EXTENSIN-LIKE RECEPTOR KINASE1* (*PERK1*). This gene possesses a proline-rich ECD with sequence similarity to the extensin family of proteins, and *PERK1* gene expression was rapidly induced by wounding, suggesting a role in sensing cell wall modifications following cell wall damage or pathogen responses, similar to the WAKs. Even though characterized *PERK* genes suggest that they have functions in development and virus infection, up to now, ligands for these receptors have not been described (12).

2.1.11. Lysin motif domain. In the search for receptors involved in the perception of Nod factors during the early steps of nodulation following rhizobial infection of legumes, a hypothesis on the presence of two receptors was put forward in 1994 (3). However, it took almost 10 years before two different teams, led by Stougaard (92, 113) and Geurts (88), published their results on the cloning of a yet-undescribed type of receptor containing LysMs in their ECD. The receptors analyzed in these papers, i.e., NOD FACTOR RECEPTOR1 (NFR1), NFR5, and SYM2 [now known as LysM domain-containing receptor-like kinases (LYKs)], contained two or three LysM modules that occur in bacterial peptidoglycan-binding proteins or in chitinases from yeast and alga (20). Based on studies conducted over the last decade, it is now clear that all LysM-RLK receptors characterized to date, in complex with LysM-RLP or with a kinase domain devoid of kinase activity [called non-RD kinases (30) and representing 20% of all kinases (72)], bind to chitooligosaccharides ligands (89). Their dual role in plant immunity and in the establishment of the arbuscular mycorrhizal and rhizobium-legume symbioses is also well-defined (19).

2.1.12. Calcium-dependent lectin domain. The association between the calcium-dependent (or C-type) lectin (C-LEC) motif and a kinase domain in plants was noted for the first time in *Arabidopsis* by Shiu & Bleecker (121) in 2001. Interestingly, this association was observed in a single receptor. Contrary to the abundance of other types of lectin receptor kinase in plants, the very low number of receptors with this C-LEC kinase structural organization has now been confirmed in several genomes (one or two copies) (8, 124, 148, 163). However, we still do not know their functions. In animals, C-LEC receptors are numerous and have been extensively studied because of their decisive role in pathogen recognition and immunity (16).

2.2. Other Extracellular Motifs with Few Representatives or Uncommon Associations

PR: pathogenesis-related

Several new domain associations have been discovered in the search for Resistance genes (R genes). The RLKs described in this section are representative of domain associations that are sometimes species-specific and present in only a few copies in plant genomes. Consequently, for many of them, more extensive investigations are needed to uncover their real functions and activation mechanisms.

2.2.1. Chitinase (glycoside hydrolase)-type domain. In 2000, Kim, Pai, and coauthors (62) cloned the *CHITINASE-RELATED RLK1* (*CHRK1*) gene. This new type of receptor contains a domain closely related to chitinase in its extracellular part, but lacking the essential residue required for chitinase activity. Co-suppression of the endogenous tobacco gene showed pleiotropic developmental phenotypes, and mRNA accumulation was strongly stimulated by fungi and viruses (76).

2.2.2. Thaumatin domain. In 1996, Wang, Lawton, and coauthors (156) published a paper stating that "The PR5K Receptor Protein Kinase from *Arabidopsis thaliana* Is Structurally Related to a Family of Plant Defense Proteins." Indeed, the extracellular part of this pathogenesis-related (PR)-5-like RLK contains a thaumatin-like motif, which shares sequence similarities with thaumatin, a sweet-tasting protein originally found in an African berry (150). Thaumatin RLKs are transcriptionally induced by pathogenic and environmental stress, and some of these receptors play a specific role in abscisic acid–dependent drought-stress signaling (6, 144).

2.2.3. Tandem kinase-pseudokinase. In 2002, Brueggeman and coauthors (18) cloned the stem rust-resistance gene *Rpg1* in barley. This gene had homology to receptor kinases, but the domain organization was unique: Two kinase domains were fused into one protein, with the second one being a pseudokinase. Because this structure has been found in a wide range of plant taxa, it has recently been termed tandem kinase-pseudokinase (TKP) (63).

2.2.4. Nucleotide-binding site leucine-rich repeat kinase. In 2008, Brueggeman and coauthors (17) described the *Rpg5* gene, which is also involved in stem rust resistance. This time, this new R gene was found to code for a protein with a nucleotide-binding site (NBS), LRRs, and protein kinase domains. This organization is now defined as integrated domain (ID) nucleotide-binding leucine-rich repeat (NLR) proteins (NLR-IDs). NLR-IDs have been found in many plant species, and up to 10% of NLRs contain IDs (68, 115). The kinase association, as for other IDs, could be a strategy to trap kinase-targeting pathogen effectors (66; see also 138).

2.2.5. Pathogenesis-related protein 1 domain. In 2004, a new structural organization was described by Shiu and colleagues (123) in a paper comparing *Arabidopsis* and rice RLKs. The domain pathogenesis-related protein 1 (PR-1) was found to be fused to a kinase domain related to the CRR subfamily. Several years later, in 2013, the same ECD configuration was detected in two genes in the *Theobroma cacao* genome (140). The PR-1 domain was found to be fused to the TM and kinase domains related to the L-LEC subfamily, suggesting that these domain associations had occurred at least twice. The origin of this fusion in cacao seemed to involve retrotransposition since one of the genes was surrounded by retrotransposition marks and repetitive elements.

2.3. Undefined Motifs in Extracellular Domain or No Associated Domain: RKF3 and LRK10-Type

Takahashi et al. (137) described a third receptor, RKF3. It possessed all the features of a canonical receptor kinase, namely an SP, an ECD, a TM domain, and a kinase domain. However, in the 200 AAs forming the ECD, no distinct motifs were found. This was also the case for the LRK10 receptor cloned in wheat by Feuillet and coauthors (41). This receptor is part of a gene family, wheat leaf rust kinase (WLRK), containing three conserved regions separated by three variable regions in their ECD.

2.4. Cytoplasmic or Membrane-Associated Receptor Kinases

Besides these receptor structural configurations, some kinase proteins lacking an apparent TM domain but closely related to these RLKs were also discovered.

2.4.1. Kinase only. For example, the *Arabidopsis* APK1 protein, which was described as phosphorylating tyrosine, serine, and threonine, was cloned by Hirayama & Oka (56) in 1992. This protein is composed of 410 AAs and possesses only a kinase domain with a short N-terminal domain containing an N-myristoylation site, but no TM domain or ECD was identified. Moreover, in 1993, Martin, Tanksley, and coauthors (99) reported cloning the first kinase domain-containing protein conferring bacterial speck resistance in tomato. The cloned *Pto* gene encoded a 321-AA hydrophilic protein that lacked the TM domains and ECDs. As for APK1, the presence of a putative myristoylation site suggested that the Pto kinase could be membrane-associated but not an integral membrane protein. In 1999, Snyders & Kohorn (129) revealed other interesting *Arabidopsis* receptors associated with energy transduction. These thylakoid-associated kinase RLKs (TAK RLKs), with no ECD but with a single N-terminal hydrophobic region, were the first RLKs found to be involved in the chloroplast phosphorylation network. Up to now, the kinase-only proteins that have been studied function in complex with S-domain, LRR, malectin-like, and LysM RLKs. They are phospho-dependent transducers acting downstream of RLKs, and they intervene in all signaling pathways and functions fulfilled by RLKs (85).

2.4.2. Universal stress protein A-kinase-U-box domains. Finally, in 2003, Kerk, Gribskov, and coauthors (61) described a new structural configuration consisting of an N-terminal universal stress protein A (UspA) domain of *Escherichia coli*, a kinase domain related to RLKs, and a C-terminal U-box domain. The association of a kinase domain followed by a U-box domain had already been reported in 2001 in an article discussing the characteristics of a new protein family called plant U-box (PUB) proteins (4). To our knowledge, none of these proteins have been studied so far, but, with the ID model in mind, these proteins could also play the role of a trap in the many RLK signaling pathways in which PUBs and ubiquitination have been involved (27, 40). Alternatively, this could simply be a way to fuse two proteins that normally work subsequentially.

This Prévert-style inventory of researchers and articles describing the discoveries of new RLKs in the last decade of the twentieth century has laid the foundations for showcasing that (*a*) these receptors are involved in a myriad of different functions such as developmental processes, disease resistance, symbiosis, or self-incompatibility, (*b*) they are present in many plant species, and (*c*) they are all phylogenetically related.

3. THE ORIGIN OF THE RECEPTOR-LIKE KINASE DOMAIN

The high diversity of the domain combinations in the RLKs we have described, especially in their ECD, raises several questions regarding their origin: How are RLK domains related to kinase

domains in other kingdoms? When did each of these combinations appear? Did they appear around the same time or at very distinct periods? Did each of them appear only once, or did some of them appear multiple times over the course of evolution? A journey into the past is warranted to answer these questions.

Hanks, Quinn, and Hunter (48) were the first to initiate, with the data set of protein kinases available in 1988, a classification based on similarities in their AA sequences. They thus aligned the 250–300 AAs of 65 protein kinase catalytic domains from vertebrates and invertebrates. This allowed them to precisely describe, for the first time, the 11 major conserved subdomains of the catalytic domains. Additionally, their phylogenetic analysis revealed five major clades that pooled protein kinases with similar modes of regulation or substrate specificities. One of the clades grouped protein tyrosine kinase sequences closely related to the Ser/Thr kinase Raf protein. In 1995, Hanks & Hunter (47) proposed a new classification of a larger set of 400 protein kinases, this time representative of all kingdoms. Among the 99 protein kinase catalytic domains used to infer the kinase phylogeny based on the maximum parsimony principle, 41 plant protein kinases were included in this tree, 6 of which were RLKs, 3 containing an S-domain [ZmPK1 (154), SRK2 and SRK6 (134)], 1 an LRR domain [TMK1 (21)], 1 a WAK domain [Pro25 (WAK1) (65)], and 1 with no associated domain [APK1 (56)]. The resulting tree presented in that paper revealed that the plant RLK sequences clustered together, suggesting a monophyletic origin. They were classified as "other protein kinase families (not falling in major groups)." Moreover, the closest animal kinase to this group was the *Drosophila* Pelle kinase, indicative of a possible common origin of plant RLKs and animal Pelle kinases. The fruit fly kinase was also classified in the "other protein kinase families (not falling in major groups)" category, but in a different subgroup. Another important result from this analysis is the fact that, in the S-domain group, ZmPK1 and SRKs were not grouped together in the inferred tree, even though they both contain an S-domain associated with the kinase domain.

The second attempt to build a phylogenetic tree combining RLK and animal kinase domains was published in 1997 by Clark, Williams, and Meyerowitz (25). In their phylogenetic tree, based on the minimum evolution principle, all plant RLKs analyzed [S-domain: ZmPK1, RLK1 and RLK4 (153), and SRK; LRR domain: CLAVATA1 (CLV1) (25), RLK5 (153), TMK1, ERECTA (ER) (143), and Xa21 (132); CRINKLY4 domain: CR4 (7); and no associated domain: PTO (99) and FEN (91)] formed separate lineages distinct from animal kinases. Moreover, the four S-domain receptors were again not clustered together, with SRK being outside of the RLK1, RLK4, and ZmPK1 clade, thus confirming previous results (47). Another observation was that the no associated domain receptors PTO and FEN clustered together and were close relatives to an LRR domain receptor clade, including CLV1, RLK5, and ER. However, some other LRR domain sequences, i.e., Xa21 and TMK1, were not included in this group, and formed two disjointed clades, also distinct from the CRINKLY4 domain CR4 clade. In 1999, Hardie (49) published a review in which the phylogenetic relationships of 89 kinase domains from *A. thaliana* were considered, including 18 RLKs. This analysis showed that these protein kinases clustered into 12 major subfamilies, one being RLKs. However, as noted in the article, the kinase domain sequences of RLKs were more variable than those of the other subfamilies. Moreover, these sequences did not cluster in exactly the same way, depending on whether the analyses were done using their kinase or ECD sequences. Four subclasses emerged in the phylogenetic tree inferred using the ECDs of RLKs. The first subclass is the S-domain containing RLK1, RLK4, and ARK1 (141). The second is the LRR domain containing CLV1, ER, RLK5, TMK1, TMKL1 (149), RKF1 (137), and the BRASSINOSTEROID-INSENSITIVE1 (BRI1) receptor (83). The third is the WAK domain containing Pro25 (WAK1) and WAK4 (52). The fourth is the L-LEC domain containing LecRK1 (55). The accumulation of *Arabidopsis* sequences available in databases by

Mya: million years ago

2000 (101), and the recent release of the complete sequence of the *Arabidopsis* genome, revealed that the *Arabidopsis* genome contains nearly 1,000 genes encoding Ser/Thr protein kinases, several hundred of which are RLKs (2). Comparative genome analysis between *Arabidopsis* and other sequenced genomes again highlighted that RLKs were highly similar to the *Drosophila* Pelle protein kinase and the mammalian INTERLEUKIN1 RECEPTOR-ASSOCIATED KINASE (IRAK).

Overall, these results provided a first indication that (*a*) RLKs have a monophyletic origin and are close relatives to animal Pelle kinases and IRAKs, (*b*) RLK kinase domains evolve more rapidly than kinase domains of other protein kinases, (*c*) some ECDs found in plants are unfamiliar in animals, and (*d*) although the ECDs are similar, kinase domains of some of these RLK receptors do not cluster together in the phylogenetic trees, suggesting that several associations between these ECDs (e.g., S-domain or LRR) and different kinase domains could have occurred independently. These results paved the way for Shiu & Bleecker (121) to publish a first remarkable paper in 2001. All of this evidence supports the monophyletic origin of the RLK/Pelle subfamily [renamed by Shiu & Bleeker (121)], and the large-scale expansion of this gene family in plants (few genes in animals versus hundreds in plants) (123). However, the time of divergence of this clade from other kinases has yet to be precisely established. It seems clear from these data that the origin of the RLK/Pelle clade could predate the divergence of plants and animals (121), i.e., around 1,580 ± 90 million years ago (Mya) (54), but the presence of RLK/Pelle in the Alveolata phylum [e.g., *Plasmodium* (122), *Perkinsus*, and *Toxoplasma* (79)] suggests that it may be even older, i.e., between 1,580 and 1,840 Mya (54) (**Figure 2**). Analyses of new genome sequences revealed, however, that RLK/Pelle orthologs are not found in the excavates *Leishmania major* and *Giardia intestinalis* (96), in *Dictyostelium*, the model organism of Amoebozoa (44), in fungal genomes, or in *Monosiga brevicollis*, a choanoflagellate close relative to animals (79, 121, 122), therefore suggesting massive losses in these lineages under this hypothesis. A search for RLK/Pelle sequences in many other species in these phyla should thus help to complete the evolutionary history of these kinase domains, which could be more complex than anticipated [e.g., involving horizontal transfers (79) or convergent evolution].

4. AT THE ORIGIN OF EXTRACELLULAR LIGAND-BINDING DOMAIN-KINASE DOMAIN ASSOCIATION

Two approaches are found in the RLK-related literature that are focused on either a specific RLK subfamily or a specific taxon. The first approach analyzes a given RLK subfamily, and it is studied among several species in order to reconstruct its evolutionary history. This approach and the resulting conclusions depend on the genomes available at the time of the analyses. The second approach focuses on a given species, and it aims to inventory all of its RLKs, generating a very precise picture of a given step in the evolution of the RLK family. Below we summarize findings on RLK evolution derived from the literature as well as our studies using these two approaches.

The seminal extensive genome-wide analyses conducted by Shiu & Bleecker (121, 122) provide an evolutionary framework that is highly useful for the whole scientific community working on RLKs. Moreover, additional later studies by Shiu and collaborators (77–79, 123) have also greatly helped gain further insight into the evolutionary dynamics of this gene family in plants. Indeed, these were the first reported large-scale analyses of RLK gene families comparing several species. Phylogenetic analyses based on kinase domains combined with annotations of ECD structural domains led to the first attempt to classify this large family of plant receptors. The RLK family was in turn subdivided into more than 50 different subfamilies, around 20 of which are LRR receptor kinase subfamilies and 20 are receptor-like cytoplasmic kinases (RLCKs) that have been defined as receptor kinases with no apparent SP or TM domain (grouping the kinase-only and the UspA kinase U-box) (121). These different subfamilies tend to have similar structural organization in

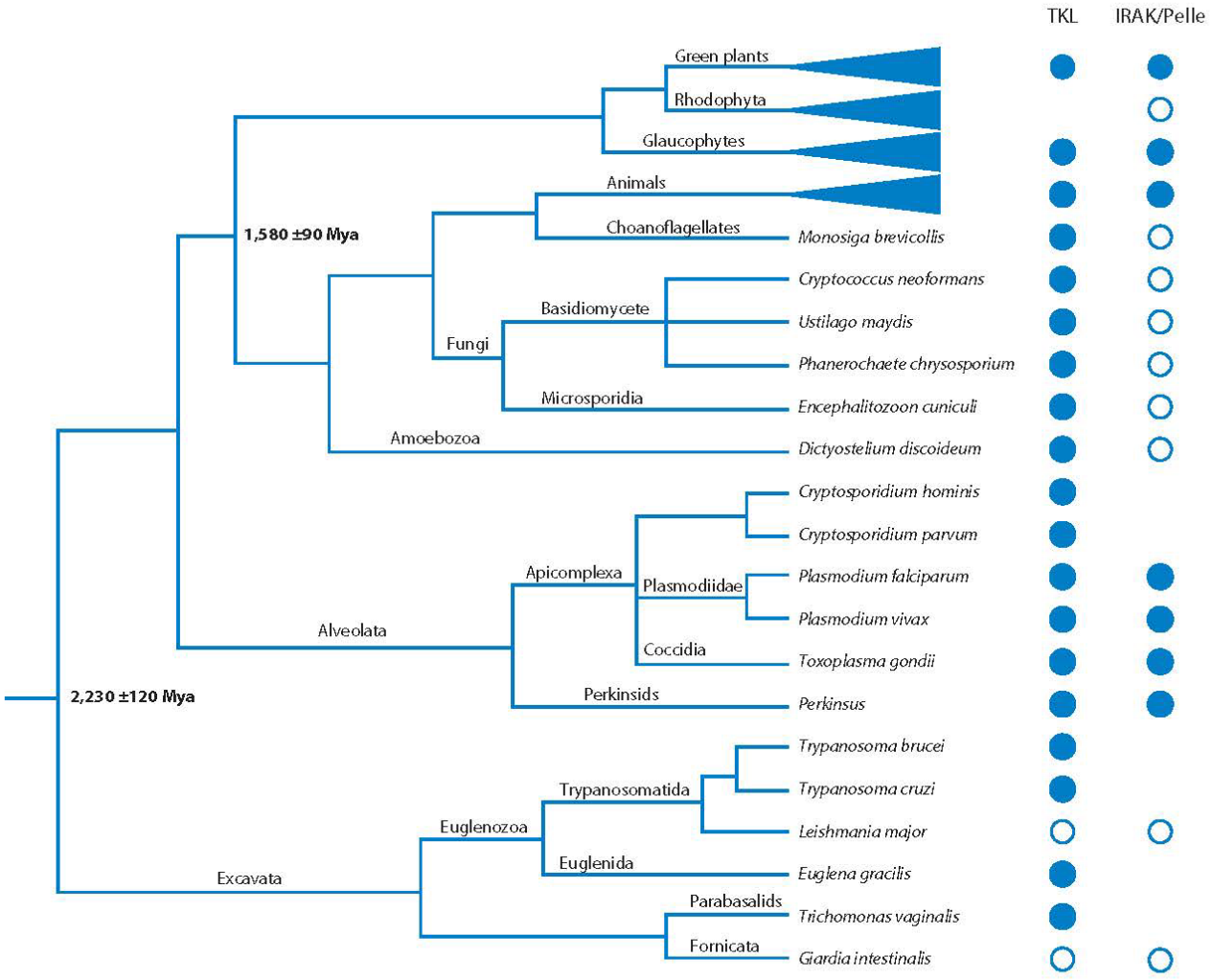


Figure 2

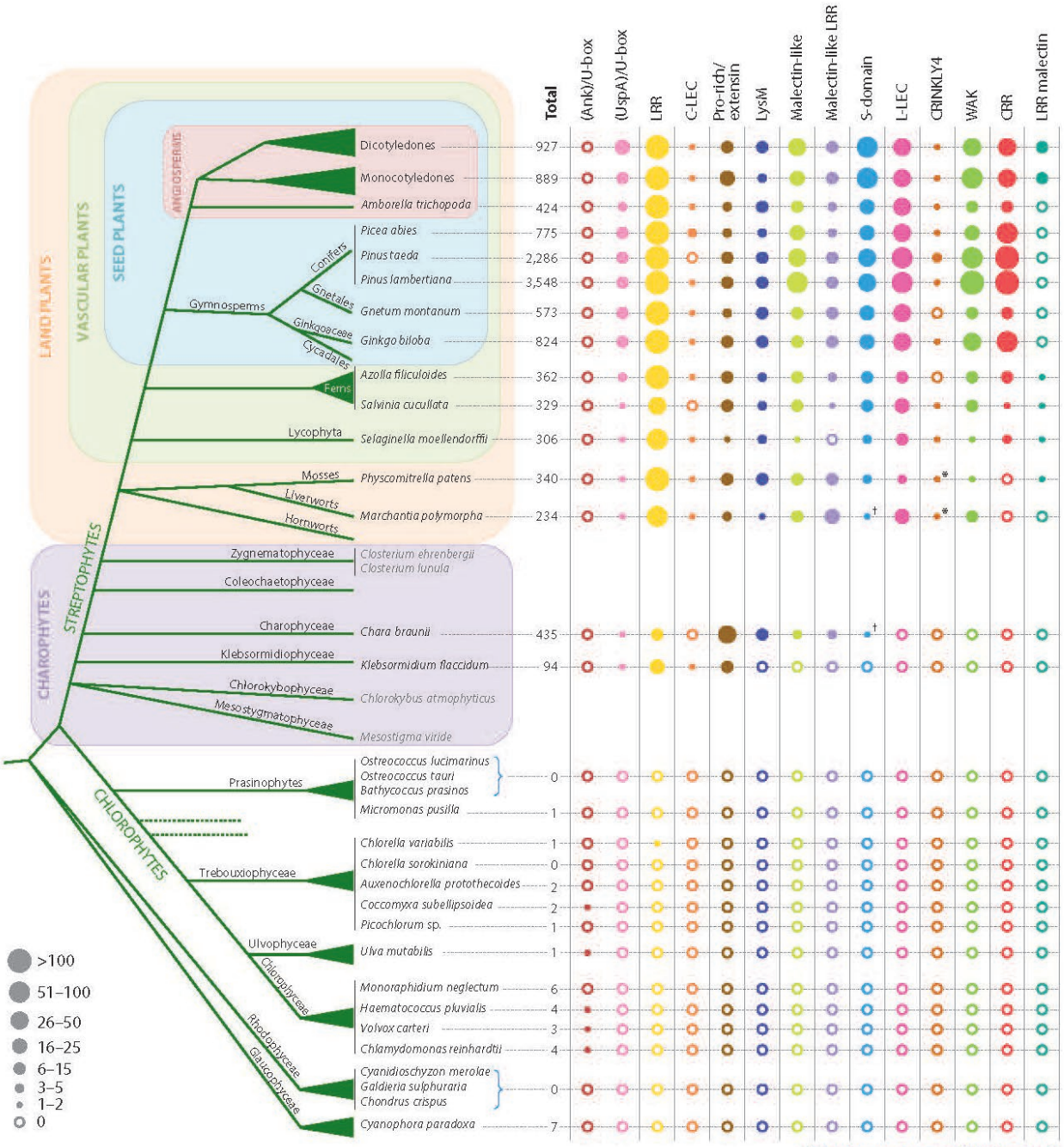
The presence/absence of tyrosine kinase-like (TKL) proteins and inter leukin 1 receptor-associated kinase (IRAK)/Pelle kinase families in the tree of life, showing a schematic representation of phyla and species in which TKL and/or IRAK/Pelle kinases have been found (*filled circle*), searched but not found (*open circle*), or not searched yet (*blank*). The tree is based on the Tree of Life Web Project (**http://tolweb.org/tree/**) and Reference 54. Note that based on Hanks classification, RLKs are part of the tyrosine kinase-like (TKL) IRAK/Pelle subgroup (45, 69, 98).

their ECDs (**Supplemental Figure 1**). This large body of literature has shed considerable light on the extraordinary gene expansion, i.e., the increasing number of genes in some subfamilies, as the consequence of duplication events, including whole genome, segmental, or tandem duplications.

Supplemental Material >

In the streptophytes, the number of proteins within RLK subfamilies expanded, but also the number of these different subfamilies increased (resulting in diversification) (70). An assessment of the presence and structural architecture of RLK ECDs in the different lineages is required to trace the origins of RLK diversification, i.e., the increasing number of specific domain combinations. In the early twenty-first century, whole plant genome sequences have started to be released, and a wealth of data is now available (100). For example, the kinomes of almost 200 genomes of plants are stored in the iTAK database (166), and they are classified based on the latest Shiu

classification proposed in 2012 (77). Moreover, the recent availability of many new genome sequences in Charophyta and basal land plant lineages has led to a more precise description of the fusion events that have shaped RLK subfamilies in angiosperms. These data from the literature and our unpublished studies are described below and summarized in **Figure 3**.



(Caption appears on following page)

Figure 3 *(Figure appears on preceding page)*

The presence and frequencies of RLK domains within major green lineages. The color code is the same as in **Figure 1**; a filled circle depicts the presence of the concerned structure, and its size is proportional to the number of sequences found; an open circle means that no gene has been found. For a brief description of the current knowledge on green plant lineage (e.g., Viridiplantae) phylogeny, readers are referred to **Supplemental Data 1**. There are two areas of discrepancy between published results and ours. The first, shown by an asterisk (*), is the appearance of the CRINKLY4 subfamily in *Marchantia* (15). Our data suggest that the fusion between the CRINKLY4 domain and a kinase appeared in the *Physcomitrella patens* genome (74) and not in *Marchantia*. Previous phylogenetic analysis of the CRINKLY4 subfamily showed that only one protein was a true CRINKLY4 family member containing all of the characteristic domains in the *P. patens* genome (105). This discrepancy could be due to the different levels of importance given to the presence of structural domains in the CRINKLY4 RLK. These results should then be taken with caution. The second discrepancy is shown by a dagger (†). Our data suggest that S-domain RLKs could be a new putative receptor configuration that appeared before the emergence of the Phragmoplastophyta clade (that includes three charophyte lineages: Charophyceae, Coleochaetophyceae, and Zygnematophyceae, together with land plants). However, these observations refer to a sole sequence that, in our study, would have been lost in *Marchantia*. Based on published data, the S-domain RLK subfamily only appeared in land plants (15, 162). Gray text indicates a species that is included in this article, but for which a complete proteome is not available. Abbreviations: ANK, ankyrin repeat; C-LEC, calcium-dependent lectin; CRR, cysteine-rich repeat; L-LEC, L-type lectin; LRR, leucine-rich repeat; LysM, lysin motif; RLK, receptor-like kinase; UspA, universal protein A; WAK, wall-associated kinase.

Supplemental Material

4.1. Receptor-Like Kinases in Angiosperms

To supplement the data harvested in the literature, we selected 176,020 protein kinases classified as RLK/Pelle in the iTAK database (166) and scanned them to annotate their structural domains (**Supplemental Table 1**). On average, there are approximately 900 RLKs per genome in the 127 dicot and 53 monocot species analyzed in accordance with previous results (123, 167). One third consist only of a kinase domain (62,642 proteins). Some of these proteins are RLCKs. In 12% of the analyzed receptor configurations, only a TM and kinase domain were detected without any other motifs in the ECD (13,591 proteins). However, in many of these proteins, the ECD was very long (several hundred AAs) (**Supplemental Figure 2**). New motifs or domains may ultimately be discovered in these receptors.

More than 2,000 unique domain organizations were found in the protein data set. Among them, 1,730 were found in less than 10 proteins in all of the 180 genomes (**Supplemental Table 1**). Moreover, many variations per domain have been noted in multidomain ECDs (162) (see also **Supplemental Figure 3**). These variations shed light on the multiple domain organizations found in plant species, and show the trial by error dynamics within genome evolution. These diverse organizations also reveal that these domains have often been reused throughout evolution, e.g., the association of RLPs with kinase domains. Of course, many of these configurations are probably annotation errors (e.g., gene boundary delimitation leading to protein fusion prediction errors or a lack of domain recognition by the software) since these architectures are based on automatic genome annotations. Nevertheless, the 12 types of domain described in Section 2.1 are the most abundant predicted structures observed in RLKs in this data set (**Table 1**). The largest subfamilies are LRR and putative carbohydrate-binding [containing the S-domain, L-LEC, C-LEC, malectin, malectin-like (CrRLK1L), LysM (8) and CRR (147) kinases], with over 200 LRR and 150 putative carbohydrate-binding receptors per genome on average. Note that only one C-LEC RLK was present per genome, as mentioned above. Could the uniqueness of the C-LEC receptor in plant genomes be an indication of its functional importance? Among these ECD kinase associations, only one was probably not present in the angiosperm ancestors, i.e., the LRR malectin receptor kinase that is not found in the *Amborella* genome and whose expansion seems to have been monocot- and dicot-specific (38, 42) (**Figure 3**). Note, however, that the evolutionary history of this subfamily is perhaps more complex since some LRR malectin receptors—despite also being absent in gymnosperms—are found in basal vascular plants and in the *Physcomitrella* genome (one or two copies).

For the structures described in Section 2.2 (**Table 1**) and for many others (**Supplemental Table 1**), a sporadic distribution in monocot and dicot species has been observed without any obvious lineage-specific expansion. This rate of association is assumed to have been very low (78). If the events we noticed were not due to annotation errors, these domain swapping events could represent new randomly occurring domain associations. The putative ones that caught our attention are (*a*) the integration of a WAK domain in the S-domain organization, (*b*) the gain of a kinase domain in NBS-LRR proteins—but unlike Rpg5 (17), the kinase domain here is located at the N-terminal end with respect to the NB-ARC and LRR domains—and (*c*) receptor kinases containing only a malectin domain but no detected LRR repeats (**Supplemental Figure 4**). However, as these proteins have only been characterized by sequence analysis, further functional studies of these genes should provide more insight into the functions of these putative receptors.

4.2. Land Plant Clade

Prior to the publication of complete genomes of species belonging to the land plant group but outside of angiosperms, 29 cDNAs of RLKs (belonging to the LRR, C-LEC, Pro-rich/extensin, LysM, malectin-like (CrRLK1L), malectin-like LRR, and L-LEC subfamilies) were identified in *Marchantia polymorpha* (116). In the recent publication of the complete *M. polymorpha* genome (15), a whole section is devoted to the study of RLK genes, revealing the presence of many of them in this early diverging lineage of land plants. Although some subclasses were not represented (e.g., the BRI1 clade in the LRR subfamily), all domain RLK subfamilies were present except the CRR RLKs that appeared in vascular plants (*Selaginella*) (147) (see the caption for **Figure 3** for the discrepancy observed for the CRINKLY4 and S-domain subfamilies).

4.3. Charophyta Lineages

Before the complete genomes of charophyte algae were sequenced, a few studies explored the presence of RLKs in cDNA libraries, and some were found in two Charophycean algae, *Closterium ehrenbergii* (Zygnematophyceae, 14 RLKs) and *Nitella axillaris* (Charophyceae, 13 RLKs) (116). In another study, the presence of symbiotic gene homologs within the entire green lineage revealed that homologous sequences of LysM, malectin-like (CrRLK1L) and malectin-like LRR-domain RLKs were present in charophyte lineages, so these associations appeared before the emergence of land plants (34, 152) (**Figure 3**). The availability of the complete genome of *Chara braunii* revealed that the LysM RLK gene family was not only present but also expanded in this species (seven copies) (106). Since L-LEC, CRINKLY4, WAK, and putatively S-domain subfamilies are found in basal land plants but not in the *C. braunii* genome, the absence of these receptor configurations in Zygnematophyceae (23) and Coleochaetophyceae genomes would be highly indicative of their land plant specificity.

The earliest Streptophyta sequenced genome available today is *Klebsormidium flaccidum* (60). Although a detailed list of RLKs was not specifically established, RLKs were clearly identified as a gene family in *Klebsormidium*, for which the numbers of genes were significantly increased in land plant genomes (60). We found 94 RLKs in this genome, and the first canonical receptor configuration with an acquisition of a TM domain was observed in it. The only kinase-associated ECDs found in this genome are the LRR, C-LEC, and Pro-rich/extensin domains, thus suggesting that these fusions were the first that occurred and arose before the Klebsormidiophyceae split.

Mesostigmatophyceae (represented by *Mesostigma viride*) and Chlorokybophyceae (represented by *Chlorokybus atmophyticus*) are the earliest known diverging streptophyte lineages (81). So far, no RLKs have been identified from EST/cDNA data (127), but complete genomes in

these very early diverging lineages would be essential to delve deeper into the origin of the associations between LRR, C-LEC, and Pro-rich/extensin domains, which may have emerged early in the streptophyte lineage.

4.4. Chlorophyta Clade

In the Chlorophyta clade, there have been several attempts to screen for the presence of RLKs. A search in the *Ostreococcus tauri* (35) and *Chlamydomonas reinhardtii* genomes revealed the presence of two RLCK genes in *C. reinhardtii* but none in *O. tauri* (78). When using the standalone iTAK program (166) to assess several chlorophyte genomes, we found a few kinase domains classified as RLK/Pelle, some with various associated domains [e.g., ankyrin- or WD40-repeats annotated by the InterProScan program (102) but no TM domain detected by the TMHMM program (67)]. One domain that was found repeatedly in several species from the Chlorophyceae (three copies), Ulvophyceae (one copy), or Trebouxiophyceae (one copy) was a C-terminal U-box domain associated or not with other N-terminal domains resembling the UspA kinase U-box RLCKs described above (**Figure 3**). The presence of this particular assembly in some RLCK genes, also detected throughout the streptophyte lineage, could be a strong clue that this subfamily may have been present before the streptophyte-chlorophyte split, but after the Rhodophyceae divergence. Moreover, they could represent the ancestors from which RLKs arose.

Some LRR kinase domain associations have been observed in *Chlorella variabilis* (Trebouxiophyceae, Chlorellales) but not in any other chlorophytes analyzed previously [*C. reinhardtii* and *Volvox carteri* (Chlorophyceae), *Micromonas pusilla*, *Ostreococcus lucimarinus*, and *O. tauri* (prasinophytes)] (36, 90, 130). In the 14 species analyzed in this study, an LRR domain kinase association was again only detected in *C. variabilis*, and it was not found even in its close relative *Chlorella sorokiniana*. Previous phylogenetic analysis showed that this kinase domain did not cluster with the monophyletic RLK/Pelle subfamily (36). Overall, these results suggest that the LRR kinase association observed in this sole *Chlorella* genome could be a unique event that occurred independently of the successful association that spread in the streptophyte genomes.

5. CONCLUDING REMARKS

In conclusion, the RLK diversity observed in various plant genomes, in terms of domain combinations, did not occur in a single step over the course of plant evolution but instead took place gradually. Clearly, each new sequenced genome, particularly at the basal branches of streptophyte species, has made it possible to describe an increasingly precise history of the origin and evolution of these receptors. Although the picture is still only partial—especially with a lack of data on charophytes—the emergence of certain subfamilies in *C. braunii*, for example, led to the hypothesis that the high morphological complexity of *Chara* could result from the advent and/or expansion of certain gene subfamilies, including LysM, malectin-like (CrRLK1L), and malectin-like LRR RLKs (**Figure 4**) (106). Similarly, the appearance of RLK subfamilies that are found throughout land plants but missing in charophytes (WAK, L-LEC, CRINKLY4, and S-domain) is correlated with the diversification of both the developmental and defense-signaling mechanisms necessary for out-of-water adaptation (~450 Mya). The appearance of the LRR, C-LEC, and Pro-rich/extensin subfamilies in basal charophytes is still a mystery. How did they emerge? What was their function? Some tracks are beginning to appear, but the way is still long (158). Gaining in-depth understanding of the evolutionary history of these large gene families is a gargantuan task because many events, such as ancestral and recent lineage-specific duplications or domain rearrangements (e.g., fusions or exchanges), must be considered, and the

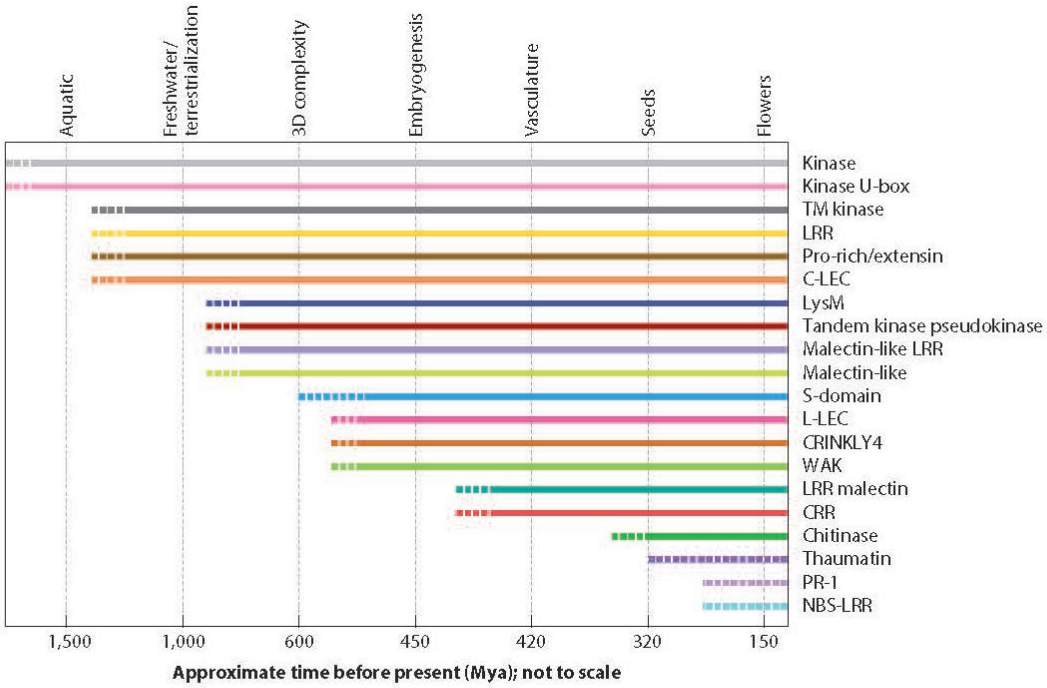


Figure 4

Relationships between the major stages of plant evolution and appearances of new structural organizations of domains. Dotted lines indicate that the exact starting point is unknown. Dates are from References 80 and 120. Abbreviations: C-LEC, calcium-dependent lectin; CRR, cysteine-rich repeat; L-LEC, L-type lectin; LRR, leucine-rich repeat; LysM, lysin motif; NBS, nucleotide-binding site; PR-1, pathogenesis-related protein 1; TM, transmembrane; WAK, wall-associated kinase.

evolutionary dynamics of ECDs and kinase domains must be studied in parallel. This huge work has been undertaken recently for the CRR kinase subfamily (147) and should open the path to new extensive analyses on other subfamilies.

SUMMARY POINTS

1. Plant receptor-like kinases (RLKs) are structurally related to animal receptor tyrosine kinase.
2. Some extracellular domains found in plants are unfamiliar in animals.
3. Domains contained in extracellular regions of plant RLKs are highly diversified.
4. RLKs have a monophyletic origin and are close relatives to animal Pelle and interleukin-1 receptor-associated kinases.
5. RLK diversity, in terms of domain combinations, did not occur in a single step over the course of plant evolution but took place gradually.

NOTE ADDED IN PROOF

While this review was in press, a new receptor configuration was described by Wu, Pei, and coauthors (159). This receptor is the first known cell-surface hydrogen peroxide (HP) sensor in plants.

The HP-INDUCED Ca^{2+} INCREASES 1 (HPCA1) receptor contains LRRs and a unique domain named the HP domain in its extracellular domain. This HP domain contains two cysteine pairs that are oxidized by HP to activate HPCA1.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was partly supported by the CGIAR Research Program on Rice (CRP RICE) (to A.D.) and by a PhD fellowship of SupAgro and CIRAD (to C.G.). Computational resources were provided by the South Green bioinformatics platform.

LITERATURE CITED

1. Ahlquist RP. 1948. A study of the adrenotropic receptors. *Am. J. Physiol.* 153:586–600
2. Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
3. Ardourel M, Demont N, Debelle F, Maillet F, de Billy F, et al. 1994. *Rhizobium meliloti* lipooligosaccharide nodulation factors: different structural requirements for bacterial entry into target root hair cells and induction of plant symbiotic developmental responses. *Plant Cell* 6:1357–74
4. Azevedo C, Santos-Rosa MJ, Shirasu K. 2001. The U-box protein family in plants. *Trends Plant Sci.* 6:354–58
5. Bacete L, Mélida H, Miedes E, Molina A. 2018. Plant cell wall-mediated immunity: cell wall changes trigger disease resistance responses. *Plant J.* 93(4):614–36
6. Baek D, Kim MC, Kumar D, Park B, Cheong MS, et al. 2019. AtPR5K2, a PR5-like receptor kinase, modulates plant responses to drought stress by phosphorylating protein phosphatase 2Cs. *Front. Plant Sci.* 10:1146
7. Becraft PW, Stinard PS, McCarty DR. 1996. CRINKLY4: a TNFR-like receptor kinase involved in maize epidermal differentiation. *Science* 273:1406–9
8. Bellande K, Bono JJ, Savelli B, Jamet E, Canut H. 2017. Plant lectins and lectin receptor-like kinases: How do they sense the outside? *Int. J. Mol. Sci.* 18:E1164
9. Bleecker AB. 1991. Genetic analysis of ethylene responses in *Arabidopsis thaliana*. *Symp. Soc. Exp. Biol.* 45:149–58
10. Boisson-Dernier A, Kessler SA, Grossniklaus U. 2011. The walls have ears: the role of plant CrRLK1Ls in sensing and transducing extracellular signals. *J. Exp. Bot.* 62:1581–91
11. Bonner TI, Oppermann H, Seeburg P, Kerby SB, Gunnell MA, et al. 1986. The complete coding sequence of the human raf oncogene and the corresponding structure of the c-raf-1 gene. *Nucleic Acids Res.* 14:1009–15
12. Borassi C, Sede AR, Mecchia MA, Salgado Salter JD, Marzol E, et al. 2016. An update on cell surface proteins containing extensin-motifs. *J. Exp. Bot.* 67(2):477–87
13. Bourdais G, Burdiak P, Gauthier A, Nitsch L, Salojärvi J, et al. 2015. Large-scale phenomics identifies primary and fine-tuning roles for CRKs in responses related to oxidative stress. *PLOS Genet.* 11(7):e1005373
14. Bourne Y, Abergel C, Cambillau C, Frey M, Rouge P, Fontecilla-Camps JC. 1990. X-ray crystal structure determination and refinement at 1.9 Å resolution of isolectin I from the seeds of *Lathyrus ochrus*. *J. Mol. Biol.* 214:571–84
15. Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, et al. 2017. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* 171:287–304.e15

16. Brown GD, Willment JA, Whitehead L. 2018. C-type lectins in immunity and homeostasis. *Nat. Rev. Immunol.* 18(6):374–89
17. Brueggeman R, Druka A, Nirmala J, Cavileer T, Drader T, et al. 2008. The stem rust resistance gene *Rpg5* encodes a protein with nucleotide-binding-site, leucine-rich, and protein kinase domains. *PNAS* 105:14970–75
18. Brueggeman R, Rostoks N, Kudrna D, Kilian A, Han F, et al. 2002. The barley stem rust-resistance gene *Rpg1* is a novel disease-resistance gene with homology to receptor kinases. *PNAS* 99:9328–33
19. Buendia L, Girardin A, Wang T, Cottret L, Lefebvre B. 2018. LysM receptor-like kinase and LysM receptor-like protein families: an update on phylogeny and functional characterization. *Front. Plant Sci.* 9:1531
20. Buist G, Steen A, Kok J, Kuipers OP. 2008. LysM, a widely distributed protein motif for binding to (peptido)glycans. *Mol. Microbiol.* 68:838–47
21. Chang C, Schaller GE, Patterson SE, Kwok SF, Meyerowitz EM, Bleecker AB. 1992. The *TMK1* gene from *Arabidopsis* codes for a protein with structural and biochemical characteristics of a receptor protein kinase. *Plant Cell* 4:1263–71
22. Chen Z. 2001. A superfamily of proteins with novel cysteine-rich repeats. *Plant Physiol.* 126:473–76
23. Cheng S, Xian W, Fu Y, Marin B, Keller J, et al. 2019. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* 179(5):1057–67
24. Choi J, Tanaka K, Cao Y, Qi Y, Qiu J, et al. 2014. Identification of a plant receptor for extracellular ATP. *Science* 343(6168):290–94
25. Clark SE, Williams RW, Meyerowitz EM. 1997. The *CLAVATA1* gene encodes a putative receptor kinase that controls shoot and floral meristem size in Arabidopsis. *Cell* 89:575–85
26. Cook DE, Mesarich CH, Thomma BPHJ. 2015. Understanding plant immunity as a surveillance system to detect invasion. *Annu. Rev. Phytopathol.* 53:541–63
27. Couto D, Zipfel C. 2016. Regulation of pattern recognition receptor signalling in plants. *Nat. Rev. Immunol.* 16(9):537–52
28. Czyzewicz N, Nikonorova N, Meyer MR, Sandal P, Shah S, et al. 2016. The growing story of (ARABIDOPSIS) CRINKLY 4. *J. Exp. Bot.* 67(16):4835–47
29. Dangl JL, Preuss D, Schroeder JI. 1995. Talking through walls: signaling in plant development. *Cell* 83:1071–77
30. Dardick C, Schwessinger B, Ronald P. 2012. Non-arginine-aspartate (non-RD) kinases are associated with innate immune receptors that recognize conserved microbial signatures. *Curr. Opin. Plant Biol.* 15(4):358–66
31. Day EK, Sosale NG, Lazzara MJ. 2016. Cell signaling regulation by protein phosphorylation: a multivariate, heterogeneous, and context-dependent process. *Curr. Opin. Biotechnol.* 40:185–92
32. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, et al. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 34:W362–65
33. Deeken R, Kaldenhoff R. 1997. Light-repressible receptor protein kinase: a novel photo-regulated gene from *Arabidopsis thaliana*. *Planta* 202:479–86
34. Delaux PM, Radhakrishnan GV, Jayaraman D, Cheem J, Malbreil M, et al. 2015. Algal ancestor of land plants was preadapted for symbiosis. *PNAS* 112:13390–95
35. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, et al. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *PNAS* 103:11647–52
36. Dievart A, Gilbert N, Droc G, Attard A, Gourgues M, et al. 2011. Leucine-rich repeat receptor kinases are sporadically distributed in eukaryotic genomes. *BMC Evol. Biol.* 11:367
37. Du S, Qu LJ, Xiao J. 2018. Crystal structures of the extracellular domains of the CrRLK1L receptor-like kinases ANXUR1 and ANXUR2. *Protein Sci.* 27(4):886–92
38. Dufayard JF, Bettembourg M, Fischer I, Droc G, Guiderdoni E, et al. 2017. New insights on leucine-rich repeats receptor-like kinase orthologous relationships in angiosperms. *Front. Plant Sci.* 8:381
39. Ebina Y, Ellis L, Jarnagin K, Edery M, Graf L, et al. 1985. The human insulin receptor cDNA: the structural basis for hormone-activated transmembrane signalling. *Cell* 40:747–58

40. Fan J, Bai P, Ning Y, Wang J, Shi X, et al. 2018. The monocot-specific receptor-like kinase SDS2 controls cell death and immunity in rice. *Cell Host Microbe* 23(4):498–510.E5
41. Feuillet C, Reuzeau C, Kjellbom P, Keller B. 1998. Molecular characterization of a new type of receptor-like kinase (wlrk) gene family in wheat. *Plant Mol. Biol.* 37:943–53
42. Fischer I, Dievart A, Droc G, Dufayard JF, Chantret N. 2016. Evolutionary dynamics of the leucine-rich repeat receptor-like kinase (LRR-RLK) subfamily in angiosperms. *Plant Physiol.* 170:1595–610
43. Ge Z, Dresselhaus T, Qu LJ. 2019. How CrRLK1L receptor complexes perceive RALF signals. *Trends Plant Sci.* 24(11):978–81
44. Goldberg JM, Manning G, Liu A, Fey P, Pilcher KE, et al. 2006. The dictyostelium kinome–analysis of the protein kinases from a simple model organism. *PLOS Genet.* 2:e38
45. Guo Y, Peng D, Zhou J, Lin S, Wang C, et al. 2019. iEKPD 2.0: an update with rich annotations for eukaryotic protein kinases, protein phosphatases and proteins containing phosphoprotein-binding domains. *Nucleic Acids Res.* 47:D344–50
46. Gust AA, Felix G. 2014. Receptor like proteins associate with SOBIR1-type of adaptors to form bimolecular receptor kinases. *Curr. Opin. Plant Biol.* 21:104–11
47. Hanks SK, Hunter T. 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* 9:576–96
48. Hanks SK, Quinn AM, Hunter T. 1988. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* 241:42–52
49. Hardie DG. 1999. Plant protein serine/threonine kinases: classification and functions. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50:97–131
50. Hashimoto C, Hudson KL, Anderson KV. 1988. The *Toll* gene of Drosophila, required for dorsal-ventral embryonic polarity, appears to encode a transmembrane protein. *Cell* 52:269–79
51. He Y, Zhou J, Shan L, Meng X. 2018. Plant cell surface receptor-mediated signaling—a common theme amid diversity. *J. Cell Sci.* 131:jcs209353
52. He ZH, Cheeseman I, He D, Kohorn BD. 1999. A cluster of five cell wall-associated receptor kinase genes, Wak1–5, are expressed in specific organs of Arabidopsis. *Plant Mol. Biol.* 39:1189–96
53. He ZH, Fujiki M, Kohorn BD. 1996. A cell wall-associated, receptor-like protein kinase. *J. Biol. Chem.* 271:19789–93
54. Hedges SB. 2002. The origin and evolution of model organisms. *Nat. Rev. Genet.* 3:838–49
55. Herve C, Dabos P, Galaud JP, Rouge P, Lescure B. 1996. Characterization of an *Arabidopsis thaliana* gene that defines a new class of putative plant receptor kinases with an extracellular lectin-like domain. *J. Mol. Biol.* 258:778–88
56. Hirayama T, Oka A. 1992. Novel protein kinase of *Arabidopsis thaliana* (APK1) that phosphorylates tyrosine, serine and threonine. *Plant Mol. Biol.* 20:653–62
57. Hohmann U, Lau K, Hothorn M. 2017. The structural basis of ligand perception and signal activation by receptor kinases. *Annu. Rev. Plant Biol.* 68:109–37
58. Hohmann U, Santiago J, Nicolet J, Olsson V, Spiga FM, et al. 2018. Mechanistic basis for the activation of plant membrane receptor kinases by SERK-family coreceptors. *PNAS* 115(13):3488–93
59. Hok S, Danchin EG, Allasia V, Panabieres F, Attard A, Keller H. 2011. An *Arabidopsis* (malectin-like) leucine-rich repeat receptor-like kinase contributes to downy mildew disease. *Plant Cell Environ.* 34:1944–57
60. Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, et al. 2014. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* 5:3978
61. Kerk D, Bulgrien J, Smith DW, Gribskov M. 2003. Arabidopsis proteins containing similarity to the universal stress protein domain of bacteria. *Plant Physiol.* 131:1209–19
62. Kim YS, Lee JH, Yoon GM, Cho HS, Park SW, et al. 2000. CHRK1, a chitinase-related receptor-like kinase in tobacco. *Plant Physiol.* 123:905–15
63. Klymiuk V, Yaniv E, Huang L, Raats D, Fatiukha A, et al. 2018. Cloning of the wheat *Yr15* resistance gene sheds light on the plant tandem kinase-pseudokinase family. *Nat. Commun.* 9:3735
64. Kohorn BD. 2016. Cell wall-associated kinases and pectin perception. *J. Exp. Bot.* 67(2):489–94

65. Kohorn BD, Lane S, Smith TA. 1992. An *Arabidopsis* serine/threonine kinase homologue with an epidermal growth factor repeat selected in yeast for its specificity for a thylakoid membrane protein. *PNAS* 89:10989–92
66. Krattinger SG, Keller B. 2016. Molecular genetics and evolution of disease resistance in cereals. *New Phytol.* 212(2):320–32
67. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305:567–80
68. Kroj T, Chanclud E, Michel-Romiti C, Grand X, Morel JB. 2016. Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol.* 210(2):618–26
69. Krupa A, Abhinandan KR, Srinivasan N. 2004. KinG: a database of protein kinases in genomes. *Nucleic Acids Res.* 32:D153–55
70. Krupa A, Anamika, Srinivasan N. 2006. Genome-wide comparative analyses of domain organisation of repertoires of protein kinases of *Arabidopsis thaliana* and *Oryza sativa*. *Gene* 380:1–13
71. Kutschera A, Dawid C, Gisch N, Schmid C, Raasch L, et al. 2019. Bacterial medium-chain 3-hydroxy fatty acid metabolites trigger immunity in *Arabidopsis* plants. *Science* 364(6436):178–81
72. Kwon A, Scott S, Taujale R, Yeung W, Kochut KJ, et al. 2019. Tracing the origin and evolution of pseudokinases across the tree of life. *Sci. Signal.* 12(578):eaav3810
73. Labbé J, Muchero W, Czarnecki O, Wang J, Wang X, et al. 2019. Mediation of plant-mycorrhizal interaction by a lectin receptor-like kinase. *Nat. Plants* 5(7):676–80
74. Lang D, Ullrich KK, Murat F, Fuchs J, Jenkins J, et al. 2018. The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* 93:515–33
75. Lee DS, Kim YC, Kwon SJ, Ryu CM, Park OK. 2017. The Arabidopsis cysteine-rich receptor-like kinase CRK36 regulates immunity through interaction with the cytoplasmic kinase BIK1. *Front. Plant Sci.* 8:1856
76. Lee JH, Takei K, Sakakibara H, Sun Cho H, Kim DM, et al. 2003. CHRK1, a chitinase related receptor-like kinase, plays a role in plant development and cytokinin homeostasis in tobacco. *Plant Mol. Biol.* 53(6):877–90
77. Lehti-Shiu MD, Shiu SH. 2012. Diversity, classification and function of the plant protein kinase superfamily. *Philos. Trans. R. Soc. B* 367:2619–39
78. Lehti-Shiu MD, Zou C, Hanada K, Shiu SH. 2009. Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiol.* 150:12–26
79. Lehti-Shiu MD, Zou C, Shiu SH. 2012. Origin, diversity, expansion history, and functional evolution of the plant receptor-like kinase/pelle family. In *Receptor-Like Kinases in Plants: Signaling and Communication in Plants*, Vol. 13, ed. F Tax, B Kemmerling, pp. 1–22. Berlin: Springer
80. Leliaert F, Verbruggen H, Zechman FW. 2011. Into the deep: new discoveries at the base of the green plant phylogeny. *Bioessays* 33:683–92
81. Lemieux C, Otis C, Turmel M. 2007. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biol.* 5:2
82. Li B, Ferreira MA, Huang M, Camargos LF, Yu X, et al. 2019. The receptor-like kinase NIK1 targets FLS2/BAK1 immune complex and inversely modulates antiviral and antibacterial immunity. *Nat. Commun.* 10(1):4996
83. Li J, Chory J. 1997. A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. *Cell* 90:929–38
84. Li W, Liu Y, Wang J, He M, Zhou X, et al. 2016. The durably resistant rice cultivar Digu activates defence gene expression before the full maturation of *Magnaporthe oryzae* appressorium. *Mol. Plant Pathol.* 17(3):354–68
85. Liang X, Zhou JM. 2018. Receptor-like cytoplasmic kinases: central players in plant receptor kinase-mediated signaling. *Annu. Rev. Plant Biol.* 69:267–99
86. Liebrand TW, van den Burg HA, Joosten MH. 2014. Two for all: receptor-associated kinases SOBIR1 and BAK1. *Trends Plant. Sci.* 19(2):123–32

87. Liener IE. 1964. Seed hemagglutinins. *Econ. Bot.* 18:27–33
88. Limpens E, Franken C, Smit P, Willemse J, Bisseling T, Geurts R. 2003. LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. *Science* 302:630–33
89. Limpens E, van Zeijl A, Geurts R. 2015. Lipochitooligosaccharides modulate plant host immunity to enable endosymbioses. *Annu. Rev. Phytopathol.* 53:311–34
90. Liu PL, Du L, Huang Y, Gao SM, Yu M. 2017. Origin and diversification of leucine-rich repeat receptor-like protein kinase (*LRR-RLK*) genes in plants. *BMC Evol. Biol.* 17:47
91. Loh YT, Martin GB. 1995. The disease-resistance gene *PTO* and the fenthion-sensitivity gene *FEN* encode closely related functional protein kinases. *PNAS* 92:4181–84
92. Madsen EB, Madsen LH, Radutoiu S, Olbryt M, Rakwalska M, et al. 2003. A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals. *Nature* 425:637–40
93. Maehle A-H. 2004. "Receptive substances": John Newport Langley (1852–1925) and his path to a receptor theory of drug action. *Med. Hist.* 48:153–74
94. Maehle A-H. 2009. A binding question: the evolution of the receptor concept. *Endeavour* 33:135–40
95. Manning G, Plowman GD, Hunter T, Sudarsanam S. 2002. Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.* 27:514–20
96. Manning G, Reiner DS, Lauwaet T, Dacre M, Smith A, et al. 2011. The minimal kinome of *Giardia lamblia* illuminates early kinase evolution and unique parasite biology. *Genome Biol.* 12:R66
97. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. 2002. The protein kinase complement of the human genome. *Science* 298:1912–34
98. Martin DM, Miranda-Saavedra D, Barton GJ. 2009. Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic Acids Res.* 37:D244–50
99. Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganal MW, et al. 1993. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science* 262:1432–36
100. Martinez M. 2016. Computational tools for genomic studies in plants. *Curr. Genom.* 17:509–14
101. McCarty DR, Chory J. 2000. Conservation and innovation in plant signaling pathways. *Cell* 103:201–9
102. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, et al. 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 43:D213–21
103. Moussu S, Augustin S, Roman AO, Broyart C, Santiago J. 2018. Crystal structures of two tandem malectin-like receptor kinases involved in plant reproduction. *Acta Crystallogr. D Struct. Biol.* 74(Pt 7):671–80
104. Nasrallah JB, Kao TH, Chen CH, Goldberg ML, Nasrallah ME. 1987. Amino-acid sequence of glycoproteins encoded by three alleles of the *S* locus of *Brassica oleracea*. *Nature* 326:617–19
105. Nikonorova N, Vu LD, Czyzewicz N, Gevaert K, De Smet I. 2015. A phylogenetic approach to study the origin and evolution of the CRINKLY4 family. *Front. Plant Sci.* 6:880
106. Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, et al. 2018. The *Chara* genome: secondary complexity and implications for plant terrestrialization. *Cell* 174:448–64.E24
107. Nissen KS, Willats WGT, Malinovsky FG. 2016. Understanding CrRLK1L function: cell walls and growth control. *Trends Plant Sci.* 21(6):516–27
108. Pan H, Stonoha-Arther C, Wang D. 2018. Medicago plants control nodulation by regulating proteolysis of the receptor-like kinase DMI2. *Plant Physiol.* 177(2):792–802
109. Pan J, Li Z, Wang Q, Yang L, Yao F, Liu W. 2020. An S-domain receptor-like kinase, OsESG1, regulates early crown root development and drought resistance in rice. *Plant Sci.* 290:110318
110. Park J, Kim TH, Takahashi Y, Schwab R, Dressano K, et al. 2019. Chemical genetic identification of a lectin receptor kinase that transduces immune responses and interferes with abscisic acid signaling. *Plant J.* 98(3):492–510
111. Pelagio-Flores R, Muñoz-Parra E, Barrera-Ortiz S, Ortiz-Castro R, Saenz-Mata J, et al. 2019. The cysteine-rich receptor-like protein kinase CRK28 modulates *Arabidopsis* growth and development and influences abscisic acid responses. *Planta* 251(1):2
112. Pruitt RE, Hulskamp M, Kopczak SD, Ploense SE, Schneitz K. 1993. Molecular genetics of cell interactions in *Arabidopsis*. *Dev. Suppl.* 1993:77–84

113. Radutoiu S, Madsen LH, Madsen EB, Felle HH, Umehara Y, et al. 2003. Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* 425:585–92
114. Rapp UR, Goldsborough MD, Mark GE, Bonner TI, Groffen J, et al. 1983. Structure and biological activity of v-raf, a unique oncogene transduced by a retrovirus. *PNAS* 80:4218–22
115. Sarris PF, Cevik V, Dagdas G, Jones JD, Krasileva KV. 2016. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol.* 14:8
116. Sasaki G, Katoh K, Hirose N, Suga H, Kuma K, et al. 2007. Multiple receptor-like kinase cDNAs from liverwort *Marchantia polymorpha* and two charophycean green algae, *Closterium ehrenbergii* and *Nitella axillaris*: extensive gene duplications and gene shufflings in the early evolution of streptophytes. *Gene* 401:135–44
117. Schellenberger R, Touchard M, Clément C, Baillieul F, Cordelier S, et al. 2019. Apoplastic invasion patterns triggering plant immunity: plasma membrane sensing at the frontline. *Mol. Plant Pathol.* 20(11):1602–16
118. Schlessinger J. 2014. Receptor tyrosine kinases: legacy of the first two decades. *Cold Spring Harb. Perspect. Biol.* 6:a008912
119. Schulze-Muth P, Irmler S, Schröder G, Schröder J. 1996. Novel type of receptor-like protein kinase from a higher plant (*Catharanthus roseus*). cDNA, gene, intramolecular autophosphorylation, and identification of a threonine important for auto- and substrate phosphorylation. *J. Biol. Chem.* 271:26684–89
120. Sessa EB, Banks JA, Barker MS, Der JP, Duffy AM, et al. 2014. Between two fern genomes. *GigaScience* 3:15
121. Shiu SH, Bleecker AB. 2001. Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *PNAS* 98:10763–68
122. Shiu SH, Bleecker AB. 2003. Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in Arabidopsis. *Plant Physiol.* 132:530–43
123. Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH. 2004. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell* 16:1220–34
124. Shumayla SS, Sharma S, Pandey AK, Singh K, Upadhyay SK. 2016. Molecular characterization and global expression analysis of lectin receptor kinases in bread wheat (*Triticum aestivum*). *PLOS ONE* 11:e0153925
125. Silva NF, Goring DR. 2002. The proline-rich, extensin-like receptor kinase-1 (PERK1) gene is rapidly induced by wounding. *Plant Mol. Biol.* 50:667–85
126. Silverstein AM. 2002. *Paul Ehrlich's Receptor Immunology: The Magnificent Obsession*. San Diego, CA: Academic
127. Simon A, Glockner G, Felder M, Melkonian M, Becker B. 2006. EST analysis of the scaly green flagellate *Mesostigma viride* (Streptophyta): implications for the evolution of green plants (Viridiplantae). *BMC Plant Biol.* 6:2
128. Smakowska-Luzan E, Mott GA, Parys K, Stegmann M, Howton TC, et al. 2018. An extracellular network of *Arabidopsis* leucine-rich repeat receptor kinases. *Nature* 553(7688):342–46
129. Snyders S, Kohorn BD. 1999. TAKs, thylakoid membrane protein kinases associated with energy transduction. *J. Biol. Chem.* 274:9137–40
130. Soanes DM, Talbot NJ. 2010. Comparative genome analysis reveals an absence of leucine-rich repeat pattern-recognition receptor proteins in the kingdom Fungi. *PLOS ONE* 5:e12725
131. Song WY, Han Z, Wang J, Lin G, Chai J. 2017. Structural insights into ligand recognition and activation of plant receptor kinases. *Curr. Opin. Struct. Biol.* 43:18–27
132. Song WY, Wang GL, Chen LL, Kim HS, Pi LY, et al. 1995. A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science* 270:1804–6
133. Stegmann M, Monaghan J, Smakowska-Luzan E, Rovenich H, Lehner A, et al. 2017. The receptor kinase FER is a RALF-regulated scaffold controlling plant immune signaling. *Science* 355(6322):287–89
134. Stein JC, Howlett B, Boyes DC, Nasrallah ME, Nasrallah JB. 1991. Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of *Brassica oleracea*. *PNAS* 88:8816–20
135. Stillmark H. 1889. Uber Ricin. *Arch. Pharmakol. Inst. Dorpat.* 3:59

136. Stracke S, Kistner C, Yoshida S, Mulder L, Sato S, et al. 2002. A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature* 417:959–62
137. Takahashi T, Mu JH, Gasch A, Chua NH. 1998. Identification by PCR of receptor-like protein kinases from Arabidopsis flowers. *Plant Mol. Biol.* 37:587–96
138. Tamborski J, Krasileva KV. 2020. Evolution of plant NLRs: from natural history to precise modifications. *Annu. Rev. Plant Biol.* 71:355–78
139. Tang D, Wang G, Zhou JM. 2017. Receptor kinases in plant-pathogen interactions: more than pattern recognition. *Plant Cell* 29:618–37
140. Teixeira PJ, Costa GG, Fiorin GL, Pereira GA, Mondego JM. 2013. Novel receptor-like kinases in cacao contain PR-1 extracellular domains. *Mol. Plant Pathol.* 14:602–9
141. Tobias CM, Howlett B, Nasrallah JB. 1992. An *Arabidopsis thaliana* gene with sequence similarity to the *S*-locus receptor kinase of *Brassica oleracea*: sequence and expression. *Plant Physiol.* 99:284–90
142. Tordai H, Bányai L, Patthy L. 1999. The PAN module: the N-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins. *FEBS Lett.* 461:63–67
143. Torii KU, Mitsukawa N, Oosumi T, Matsuura Y, Yokoyama R, et al. 1996. The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *Plant Cell* 8:735–46
144. Uknes S, Mauch-Mani B, Moyer M, Potter S, Williams S, et al. 1992. Acquired resistance in Arabidopsis. *Plant Cell* 4:645–56
145. Ullrich A, Bell JR, Chen EY, Herrera R, Petruzzelli LM, et al. 1985. Human insulin receptor and its relationship to the tyrosine kinase family of oncogenes. *Nature* 313:756–61
146. Ullrich A, Coussens L, Hayflick JS, Dull TJ, Gray A, et al. 1984. Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells. *Nature* 309:418–25
147. Vaattovaara A, Brandt B, Rajaraman S, Safronov O, Veidenberg A, et al. 2019. Mechanistic insights into the evolution of DUF26-containing proteins in land plants. *Commun. Biol.* 2:56
148. Vaid N, Pandey PK, Tuteja N. 2012. Genome-wide analysis of lectin receptor-like kinase family from Arabidopsis and rice. *Plant Mol. Biol.* 80:365–88
149. Valon C, Smalle J, Goodman HM, Giraudat J. 1993. Characterization of an *Arabidopsis thaliana* gene (TMKL1) encoding a putative transmembrane protein with an unusual kinase-like domain. *Plant Mol. Biol.* 23:415–21
150. van der Wel H, Loeve K. 1972. Isolation and characterization of thaumatin I and II, the sweet-tasting proteins from *Thaumatococcus daniellii* Benth. *Eur. J. Biochem.* 31:221–25
151. Vernié T, Camut S, Camps C, Rembliere C, de Carvalho-Niebel F, et al. 2016. PUB1 interacts with the receptor kinase DMI2 and negatively regulates rhizobial and arbuscular mycorrhizal symbioses through its ubiquitination activity in *Medicago truncatula*. *Plant Physiol.* 170(4):2312–24
152. Vigneron N, Radhakrishnan GV, Delaux PM. 2018. What have we learnt from studying the evolution of the arbuscular mycorrhizal symbiosis? *Curr. Opin. Plant Biol.* 44:49–56
153. Walker JC. 1993. Receptor-like protein kinase genes of *Arabidopsis thaliana*. *Plant J.* 3:451–56
154. Walker JC, Zhang R. 1990. Relationship of a putative receptor protein kinase from maize to the S-locus glycoproteins of *Brassica*. *Nature* 345:743–46
155. Wang C, Huang X, Li Q, Zhang Y, Li JL, Mou Z. 2019. Extracellular pyridine nucleotides trigger plant systemic immunity through a lectin receptor kinase/BAK1 complex. *Nat. Commun.* 10(1):4810
156. Wang X, Zafian P, Choudhary M, Lawton M. 1996. The PR5K receptor protein kinase from *Arabidopsis thaliana* is structurally related to a family of plant defense proteins. *PNAS* 93:2598–602
157. Wang Y, Bouwmeester K. 2017. L-type lectin receptor kinases: new forces in plant immunity. *PLOS Pathog.* 13(8):e1006433
158. Whitewoods C, Cammarata J, Nemec Venza Z, Sang S, Crook A, et al. 2018. *CLAVATA* was a genetic novelty for the morphological innovation of 3D growth in land plants. *Curr. Biol.* 28:2365–76
159. Wu F, Chi Y, Jiang Z, Xu Y, Xie L, et al. 2020. Hydrogen peroxide sensor HPCA1 is an LRR receptor kinase in *Arabidopsis*. *Nature* 578:577–81

160. Xi L, Wu XN, Gilbert M, Schulze WX. 2019. Classification and interactions of LRR receptors and co-receptors within the Arabidopsis plasma membrane—an overview. *Front. Plant Sci.* 10:472
161. Xiao Y, Stegmann M, Han Z, DeFalco TA, Parys K, et al. 2019. Mechanisms of RALF peptide perception by a heterotypic receptor complex. *Nature* 572(7768):270–74
162. Xing S, Li M, Liu P. 2013. Evolution of S-domain receptor-like kinases in land plants and origination of S-locus receptor kinases in Brassicaceae. *BMC Evol. Biol.* 13:69
163. Yang Y, Labbe J, Muchero W, Yang X, Jawdy SS, et al. 2016. Genome-wide analysis of lectin receptor-like kinases in *Populus*. *BMC Genom.* 17:699
164. Yeh YH, Panzeri D, Kadota Y, Huang YC, Huang PY, et al. 2016. The Arabidopsis malectin-like/LRR-RLK IOS1 is critical for BAK1-dependent and BAK1-independent pattern-triggered immunity. *Plant Cell* 28(7):1701–21
165. Zebisch A, Troppmair J. 2006. Back to the roots: the remarkable RAF oncogene story. *Cell Mol. Life Sci.* 63:1314–30
166. Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, et al. 2016. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9:1667–70
167. Zulawski M, Schulze G, Braginets R, Hartmann S, Schulze WX. 2014. The Arabidopsis Kinome: phylogeny and evolutionary insights into functional diversification. *BMC Genom.* 15:548

Discussion

Discussion

Table des matières

Au cours de cette thèse, je me suis intéressée à l'annotation et la caractérisation des gènes codants des récepteurs à LRR dans les génomes de riz (*Oryza*). En particulier, mes travaux se sont concentrés sur trois familles de récepteurs à LRR − les *Leucine-rich repeat receptor like kinases* (LRR-RLK), les *Leucine-rich repeat receptor like proteins* (LRR-RLP) et les *nucleotide binding-site Leucine-rich repeat* (NLR) − présentant un intérêt agronomique majeur car ils sont associés à des fonctions de signalisation cellulaire et de réponse aux stress environnementaux biotiques et abiotiques. Par exemple, le gène OsGIRL1 (LRR-RLK) est surexprimé en condition de stress salin ou de températures élevées (Park et al., 2014), le gène ERECTA (LRR-RLK) fonctionne avec le gène TMM (LRR-RLP) pour réguler l'efficacité d'utilisation de l'eau chez *Arabidopsis* (Masle et al., 2005), le gène Xa21 (LRR-RLK) confère une résistance au pathogène *Xanthomonas oryzae* pv. *Oryzae* responsable d'une part importante des pertes de rendement des cultures de riz en Asie (Song et al., 1995; Savary et al., 2019), ou les gènes RGA4 et RGA5 (NLR) confèrent ensemble une résistance au champignon *Magnaporthe Oryzae* (Cesari et al., 2013), etc. Actuellement, le contexte de changement global du climat tend à augmenter l'incidence des évènements de sécheresse et d'inondation, et favorise l'émergence de nouvelles maladies et la propagation des pathogènes à de nouvelles régions (Bebber et al., 2013). Les gènes impliqués dans les fonctions de signalisation, comme les récepteurs à LRR, sont alors en première ligne des programmes de recherche pour l'adaptation des espèces cultivées face aux maladies, aux changements climatiques et aux nouvelles pratiques culturales.

Dans ce contexte, il est pertinent de s'intéresser à l'ensemble de la diversité des récepteurs à LRR et de comprendre comment et à quelle vitesse ces gènes évoluent. Ces gènes suivent un modèle évolutif dit '*birth and death*'. La grande majorité des récepteurs à LRR sont fortement dupliqués dans les génomes, formant de larges clusters, et présentent des similarités de séquences très fortes entre différentes copies de gènes, dont certaines portent des mutations non-sens. L'ensemble de ces caractéristiques a un impact négatif sur la qualité des données disponibles et sur notre capacité à les étudier par des approches globales et génériques comme la pangénomique (Bayer et al., 2019; Outten and Warren, 2021).

Je me suis concentrée, au cours de ma thèse, sur deux aspects de l'annotation des récepteurs à LRR, l'annotation structurale des gènes dans les génomes et l'annotation des domaines et motifs dans les protéines, qui constituent les principaux prérequis pour mener des analyses

robustes, et améliorer notre compréhension de ces familles complexes. J'ai d'abord développé le pipeline LRRprofiler, qui fournit une annotation fiable et reproductible des domaines LRR en s'appuyant sur un ensemble de profils HMM adaptés aux données et aux familles de gènes (Chapitre 1). Ensuite, j'ai analysé et comparé les annotations publiques des récepteurs à LRR fournis par des méthodes automatisées. Les nombreuses incohérences observées m'ont amené à expertiser manuellement les modèles de gène de ces récepteurs et à proposer un nouveau format d'annotation tenant compte des gènes portant des mutations non-sens. Cette nouvelle annotation est associée à un pipeline de transfert d'annotation, LRRtransfer, permettant de 'calquer' l'annotation expertisée d'un génome sur des génomes proches et d'obtenir ainsi des annotations de très bonnes qualités avec une intervention manuelle limitée (Chapitre 2). Enfin, un site web dédié à la visualisation des annotations expertisées dans les génomes de riz a été créé pour valoriser le travail de la thèse et faciliter la diffusion et le partage des données associées (Chapitre 3).

A. Comprendre l'évolution des récepteurs à LRR à l'ère de la pangénomique

La démocratisation et les avancées technologiques du séquençage permettent d'avoir accès à des données génomiques, dont la qualité ne cesse d'augmenter, pour un nombre croissant d'espèces et d'accessions. Ces avancées ont ouvert la voie à des analyses descriptives et comparatives globales comme la pangénomique. L'approche pangénomique consiste à comparer des dizaines, des centaines, voire des milliers de génomes (Cheng et al., 2018), à différentes échelles de divergences, pour établir une liste exhaustive des gènes, ou autres séquences, présents dans les génomes comparés. Ces analyses distinguent alors les séquences présentes dans tous les génomes étudiés, formant le génome cœur (ou *core-genome*, indispensable), de celles présentes dans une partie seulement des génomes comparés, formant le génome accessoires (ou *dispensable genome*, facultatif) (Golicz et al., 2020). Cette approche descriptive permet d'appréhender et d'observer la diversité génétique et allélique au sein d'un compartiment donné, car elle ne se limite pas à la diversité moléculaire référencée (ou 'mappée') observée sur la base d'un seul génome de référence représentant le compartiment étudié. Dans le cadre des espèces d'intérêt agronomique, cela permet d'identifier de nouveaux gènes ou allèles pouvant être la cible des programmes de sélection pour l'amélioration et l'adaptation des plantes (Tao et al., 2019).

Différentes études pangénomiques menées sur des plantes cultivées ont montré que les récepteurs à LRR, notamment NLR et LRR-RLK, étaient particulièrement surreprésentés dans les génomes accessoires (Zhang et al., 2016; Dolatabadian et al., 2017). Nous avons observé, au cours du chapitre 2, qu'environ 30% des gènes codants les récepteurs à LRR sont non-canoniques. En particulier, les comparaisons réalisées entre Nipponbare et KitaakeX ont montré qu'environ 40 % des gènes présents uniquement chez l'un des cultivars étaient non-canoniques. Cette proportion plus élevée de gènes non-canoniques dans les génomes accessoires par rapport à l'ensemble des gènes indique qu'ils sont enrichis en gènes non-canoniques. Néanmoins, ce résultat n'est issu que de la comparaison entre deux génotypes et n'est peut–être pas généralisable. Des résultats complémentaires permettront de déterminer si cette tendance est retrouvée plus généralement, au sein des riz cultivés notamment.

Cette question de la nature des gènes composant le génome accessoire est particulièrement intéressante pour les récepteurs à LRR, notamment parce que l'analyse des génomes accessoires a justement pour but d'identifier de nouveaux gènes pouvant avoir un rôle dans l'adaptation des plantes (Jacquemin et al., 2013). La question de la 'réalité biologique' des gènes identifiés dans le génome accessoire a déjà été évoquée. Par exemple, Zhao *et al.* signalent que près de 65 % des gènes identifiés dans le génome accessoire du riz ne présentent pas de hit pour des domaines définis dans la base de données InterProScan, et que cela pourrait signaler un enrichissement en séquences artéfactuelles ou pseudogénisées (Zhao et al., 2018).

B. La qualité des séquences génomiques est un facteur limitant

Les récepteurs à LRR forment des familles de gènes avec une organisation génomique complexe qui ont une forte probabilité d'induire des biais lors de l'assemblage des séquences. Les régions des génomes portant ces gènes sont donc plus susceptibles de présenter des erreurs et des ambiguïtés de séquence. La qualité globale des séquences génomiques utilisées pour étudier ces récepteurs à LRR constitue donc un facteur limitant.

La séquence du cultivar *O. sativa japonica* cv. Nipponbare (IRGSP-1.0) est une séquence de haute qualité utilisée comme référence pour le riz et pour les autres monocotylédones. Au cours de mes travaux, je n'ai observé que peu d'erreurs et d'ambiguïtés pour cette séquence. Seuls deux gènes codant des récepteurs à LRR dont la séquence génomique pourrait être erronée, d'après des données d'expression diverses, ont été identifiés (Chapitre 2). La

validation de potentielles erreurs de séquence reste néanmoins difficile, en particulier pour ces familles complexes. Tout d'abord, le séquençage génomique d'une cible donnée peut se heurter à la difficulté de définir des amorces spécifiques à cause de l'existence de nombreuses séquences fortement similaires. Ensuite, en ce qui concerne le séquençage de transcrits, les récepteurs à LRR peuvent présenter des niveaux d'expression basale faibles ou ne pas être exprimés du tout en dehors des phases de développement ou d'infection. C'est d'autant plus probable pour les gènes pseudogénisés (en cas d'absence de promoteur par exemple). Par exemple, les mutations non-sens de quatre gènes non-canoniques de Nipponbare n'ont pas pu être validées car les orthologues de ces gènes chez KitaakeX ne présentaient pas les mêmes mutations et aucune séquence à 100 % d'identité nucléotidique n'a été trouvée dans les bases de données d'expression que nous avons explorées. Enfin, l'importante redondance dans les séquences de ces gènes rend difficile la validation des mutations observées par alignement. En effet, lors de l'alignement de *reads* des données de RNAseq, l'observation d'un *mismatch* entre la séquence génomique et un *read* peut correspondre à une erreur dans le *read*, à une erreur dans la séquence de référence ou à un positionnement erroné du *read* le long de cette séquence.

Outre les possibles erreurs de séquençage, les ambiguïtés dans les séquences génomiques sont également un frein pour l'analyse des gènes codants des récepteurs à LRR. Ces ambiguïtés sont caractérisées par la présence de bases « N » dans la séquence. La séquence de Nipponbare de référence (IRGSP-1.0) porte 0,04 % de « N ». Nous n'avons observé que deux loci LRR portant des ambiguïtés pour ce génome. Mais la qualité des génomes disponibles pour les différentes espèces et accessions de riz est encore très hétérogène. En particulier, la fréquence des ambiguïtés dans les séquences est extrêmement variable (Annexe 4). Par exemple, je me suis intéressée au génome de l'espèce *O. Rufipogon* W1943, l'espèce sauvage identifiée comme ancestrale des cultivars *japonica*, dont j'ai également ré-annoté les gènes LRR. Le génome de W1943 contient 1,88 % de « N » qui se retrouvent dans 129 gènes LRR, soit 13,2% d'un total de 978 gènes LRR-RLK, LRR-RLP et NLR. Ces ambiguïtés au niveau des gènes d'intérêt peuvent impacter les annotations et empêchent les comparaisons entre gènes. En particulier, ces ambiguïtés constituent une perte d'information et il sera également impossible de dire si une structure portant des ambiguïtés est canonique ou non. De plus, la présence d'ambiguïtés dans une région peut être le signe que la séquence est localement de

moins bonne qualité, et peut contenir plus d'erreurs (diminution de la profondeur de séquençage par exemple). Les ambiguïtés dans les génomes, même limitées à une faible proportion, peuvent alors fortement impacter les comparaisons entre différentes espèces, en particulier pour les gènes à LRR.

Cette importante hétérogénéité des séquences génomiques est actuellement limitante pour pouvoir comparer finement les récepteurs à LRR entre différentes accessions et étudier leur évolution. Récemment, les technologies de 3ème génération, dites « *long-read* », permettent le séquençage de fragments d'ADN de plusieurs kilobases (kb, 10^3 bases). Elles ont amélioré les capacités d'assemblage des génomes en permettant de lever des ambiguïtés non résolues avec les technologies précédentes, dites « *short-read »* (1er et 2ème générations), qui produisent des fragments de 70 à 250 bases (Heather and Chain, 2016). Ces limitations sont particulièrement fortes pour la reconstruction de régions portant des répétitions comme les clusters de gènes, les séquences satellites ou les zones accumulant des éléments transposables. Le développement de ces nouvelles technologies *long-read* et la baisse des coûts permettant d'augmenter la couverture et la profondeur (jusqu'à plus de 150x parfois) de séquençage ouvrent ainsi l'accès à des assemblages génomiques de très haute qualité. Ces progrès ont conduit à définir un nouveau standard de qualité pour décrire les données, appelé '*platinum standard reference sequences*' (PSRefSeqs). Récemment, 12 nouveaux génomes de riz couvrant une partie de la diversité de l'espèce domestiquée *Oryza sativa* ont été séquencés en répondant à ces normes de haute qualité (Mussurova et al., 2020; Zhou et al., 2020). Par ailleurs, le projet I-*O*MAP est également en train de produire des séquences PSRefSeqs pour 25 représentants sauvages du genre *Oryza*.

Ces 37 nouveaux génomes PSRefSeqs, associées à la nouvelle annotation des récepteurs à LRR, devraient fournir une base solide pour mener des analyses comparatives et évolutives plus exhaustives. Mais le travail manuel réalisé pour parvenir à une telle annotation chez Nipponbare pour les récepteurs à LRR a été long, en particulier car le riz présente beaucoup de gènes non-canoniques parmi ces récepteurs (> 30 %). Il serait tout simplement impossible de réaliser un tel travail sur de nombreux génomes. Dans ce contexte, le pipeline de transfert des annotations, développé au cours du chapitre 2 et implémenté dans le pipeline LRRtransfer, nous semble fournir un appui essentiel. En revanche, la stratégie développée est semi-automatique. Les annotations obtenues après le transfert nécessitent encore une

expertise manuelle pour corriger d'éventuelles erreurs ou gènes non identifiés. Même si l'investissement en temps est fortement réduit grâce à cette approche, cela reste une solution intermédiaire, viable pour quelques génomes. Il serait certainement possible d'augmenter davantage le niveau d'automatisation de l'annotation en s'inspirant de méthodes basées sur des approches de *machine learning* (Mahood et al., 2020), ou dédiées à l'annotation familles spécifique, comme Augustus-PPX (Stanke and Waack, 2003), BITACORA (Vizueta et al., 2020), TGFam-Finder (Kim et al., 2020), qui reposent souvent sur l'apprentissage de profils HMM protéiques. Ces outils doivent être testés dans ce contexte et devront probablement être modifiés pour prendre en considération les structures non-canoniques. Les méthodes de *machine learning* sont d'autant plus efficaces qu'elles sont entrainées sur un large jeu de données. De fait, la constitution d'une base d'apprentissage de taille suffisante est souvent un frein important empêchant l'utilisation de ces méthodes. Les données expertisées durant cette thèse, fournissent une base d'apprentissage conséquente pour l'annotation des récepteurs LRR du riz.

C. Le challenge de l'annotation structurale des familles de gènes complexes

C.1. La place des pseudogènes dans l'annotation

Les pseudogènes décrivent les séquences génomiques similaires à des séquences géniques, mais présentant des caractéristiques les rendant défectueuses (Vanin, 1985; Cheetham et al., 2020), c'est-à-dire non fonctionnelle. Lors des processus d'annotation automatique, beaucoup de pseudogènes sont annotés comme des gènes fonctionnels codant des protéines. Au cours des années 2000, une résolution a été prise de supprimer ces séquences qui créent du bruit dans la détermination des répertoires de gènes des espèces (van Baren and Brent, 2006; Zhang et al., 2006). Mais l'idée qu'un pseudogène soit strictement non-fonctionnel a été fréquemment remise en question, arguant que certaines séquences, par le biais des ARNm ou de protéines tronquées, présentaient des fonctions, notamment de régulation (Pink and Carter, 2013; Kovalenko and Patrushev, 2018; Xie et al., 2019). Il est donc particulièrement important de documenter les séquences présentant des mutations non-sens et de les intégrer dans les annotations, mais il est essentiel que leur annotation n'apparaisse pas comme une structure canonique au même titre que n'importe quel autre gène (Amid et al., 2009).

Au cours de cette thèse, nous avons comparé trois sources de données d'annotation publiques pour le riz Nipponbare. Cette comparaison a conduit à l'identification de nombreuses

divergences dans l'annotation structurale des gènes codant les récepteurs à LRR. Ces divergences semblent être inhérentes aux méthodes choisies mais accentuées par la complexité de la famille d'intérêt (Fawal et al., 2014). Il a été en particulier noté, chez l'humain et la souris, que l'annotation de gènes fortement dupliqués et organisés en cluster était particulièrement mal réalisée par les stratégies automatiques (gènes defensin et gènes homeobox) (Amid et al., 2009; Wilming et al., 2015). Les auteurs de ces deux papiers concluent par ailleurs que l'annotation manuelle est essentielle pour obtenir une annotation fiable dans ce contexte.

De nombreux pseudogènes ont été identifiés chez le riz, en particulier issus de duplications en tandem (Thibaud-Nissen et al., 2009; Xiao et al., 2016). Pour le génome de référence du riz, *O. sativa japonica* cv. Nipponbare, seule l'annotation fournie par le NCBI identifie certains gènes annotés comme étant des pseudogènes. L'annotation manuelle que nous avons réalisée nous a permis d'observer de nombreuses séquences de récepteurs à LRR présentant des mutations non-sens qui sont annotées comme des copies de gènes complètes et fonctionnelles, en intégrant parfois des erreurs pour y parvenir (comme de faux introns), ou qui ne sont pas annotées du tout en fonction de la stratégie d'annotation (Chapitre 2). D'autres études avant la poursuite de cette thèse ont pu faire le même constat pour ces gènes (Meyers et al., 2003; Jupe et al., 2013; Lv et al., 2014; Bayer et al., 2019).

C.2. L'identification et l'annotation des séquences non-canoniques pour une meilleure caractérisation des familles de gènes complexes

Les changements de stratégies sur la gestion des pseudogènes dans l'annotation au cours du temps révèle un problème du cadre conceptuel de la définition de ce qu'est un pseudogène. Cette définition décrit des copies de gènes non-fonctionnelles, alors que la réalité est très hétérogène. Les gènes dupliqués forment un gradient continu entre fonctionnel et non-fonctionnel, pendant qu'ils accumulent des mutations, dont des mutations non-sens. Par exemple, peut-on considérer une copie de gène présentant une mutation non-sens provoquant un codon stop précoce au ¾ de la séquence comme un pseudogène ? ou une copie de gène complète mais dont le codon *start* est muté et code pour un résidu autre que la méthionine ? Il est essentiel de fournir une annotation complète et descriptive, facilement réutilisable, caractérisant les séquences mutées, et permettant de les distinguer des autres gènes. Dans le cadre de cette thèse, nous n'avons pas eu un regard 'fonctionnaliste' sur les

annotations des gènes d'intérêt car il est impossible de déterminer de façon systématique sur la base uniquement de données bio-informatiques, le caractère fonctionnel ou non d'un gène. Nous avons donc fait le choix de ne pas utiliser le terme « pseudogène » mais de plutôt différencier les séquences « non-canoniques » des copies de gènes « canoniques » (Chapitre 2).

La stratégie que nous avons suivie est d'identifier la totalité des séquences codantes à chaque locus d'intérêt dans une approche évolutive. Ainsi, pour les gènes affectés par des mutations non-sens, nous avons annoté la totalité du locus « ancestral », c'est-à-dire en prenant en compte l'ensemble de la séquence codante qui présente de l'homologie avec des gènes de la même famille même après la ou les mutation(s). Cette stratégie permet d'intégrer les traces de l'histoire évolutive du locus directement dans l'annotation du gène. Cette annotation complète des gènes non-canoniques, et leur identification comme tel, constitue un changement de paradigme de l'annotation, par rapport au regard fonctionnaliste actuel. A notre sens, ce changement présente plusieurs avantages, notamment pour l'étude des familles de gènes complexes. Cela permet, par exemple, de faire moins d'erreur dans la classification des gènes et d'améliorer la détection des copies d'intérêt dans les jeux de données. En effet, des domaines et motifs protéiques d'intérêt peuvent être identifiables sur les fragments de séquences situés en 3' de la mutation. Par exemple, un récepteur LRR-RLK tronqué peut être classé LRR-RLP, perdant ainsi l'information de la présence d'un domaine kinase identifiable après le premier stop. Cette annotation permet également de faire des analyses entre génome (*Presence Absence Variation* notamment) en conservant un regard sur des orthologues maintenus mais présentant des mutations pouvant impacter la fonction initiale du gène (Amid et al., 2009). C'est particulièrement vrai pour les récepteurs à LRR qui suivent un modèle évolutif '*birth and death*'. Je l'ai notamment observé au cours du chapitre 3, en explorant le cluster de gènes de résistance étudié par Mizuno *et al.* (Mizuno et al., 2020). Alors que les auteurs concluent que la domestication a augmenté le nombre de gènes de résistance dans les génomes, la réalité pourrait être plus nuancée car les données fournies par les auteurs semblent sous-estimer le nombre de loci non-canoniques d'après nos observations. L'augmentation du nombre de copies de gènes de résistance mise en avant par ces auteurs n'est peut-être pas une réalité si la majorité de ces copies sont en cours de pseudogénisation.

D. Vers des analyses comparatives et évolutives plus fiables

D.1. Phylogénomique des familles de gènes complexes

L'une des conséquences du mode d'évolution '*birth and death*' est de rendre la distinction entre les orthologues vrais et les paralogues, entre différents génomes, très difficile (Dalquen et al., 2013). Faire des erreurs dans cette détermination peut avoir des conséquences importantes sur les conclusions qui découlent de l'analyse de ces relations. En particulier, définir comme orthologues des gènes qui sont en réalité des paralogues, peut conduire à une mauvaise évaluation des pressions de sélection agissant sur l'évolution d'un locus.

Dans ce contexte, l'annotation complète et la distinction des structures de gène non-canoniques permet de faciliter les recherches d'orthologues entre espèces, en particulier pour l'analyse des clusters de gènes dupliqués (Amid et al., 2009). Si les séquences non-canoniques ne sont pas annotées, des relations d'orthologie peuvent être mal établies (Figure 30). Par exemple, si le 'vrai' orthologue est non-canonique et que la moitié de sa séquence initiale n'est pas identifiée car après un stop, une copie paralogue canonique de ce gène peut être identifiée comme étant l'orthologue vrai (Figure 30b), car cette copie présenterait une similarité de séquence sur la totalité du gène (Dalquen et al., 2013; Dalquen and Dessimoz, 2013)(Chapitre 2). La possibilité d'aligner la totalité des séquences, même pour les gènes non-canoniques, grâce aux différents formats de fichier fournis pour la nouvelle annotation, permet de construire des arbres phylogénétiques et d'établir les relations d'orthologie plus proches de la réalité.

Les comparaisons de familles de gènes complexes peuvent être grandement facilitées par la visualisation. C'est un appui qui nous semble essentiel à la compréhension des familles et de leur évolution en particulier pour des gènes dupliqués contenant des domaines répétés. Premièrement, la visualisation des gènes dans leur contexte génomique permet d'observer des conservation/variation du nombre de copies de gènes et d'appréhender la distribution de ces gènes dans les génomes. Ensuite, la visualisation des structures et annotations des gènes permet d'observer la conservation/variation du nombre de motifs LRR entre gènes orthologues et paralogues, d'identifier des caractéristiques conservées entre gènes d'une même famille ou d'un même cluster (position des introns, nombre d'intron etc). Forts de ce constat, nous avons souhaité offrir une possibilité de visualisation de ces caractéristiques pour nos gènes d'intérêt à travers un site web dédié : Geloc. Ce site permet de visualiser en

parallèle les structures intron/exon des gènes, les mutations non-sens (*stop* précoces et *frameshifts*) et les annotations de domaines protéiques des gènes LRR-RLK, LRR-RLP et NLR (Chapitre 3).

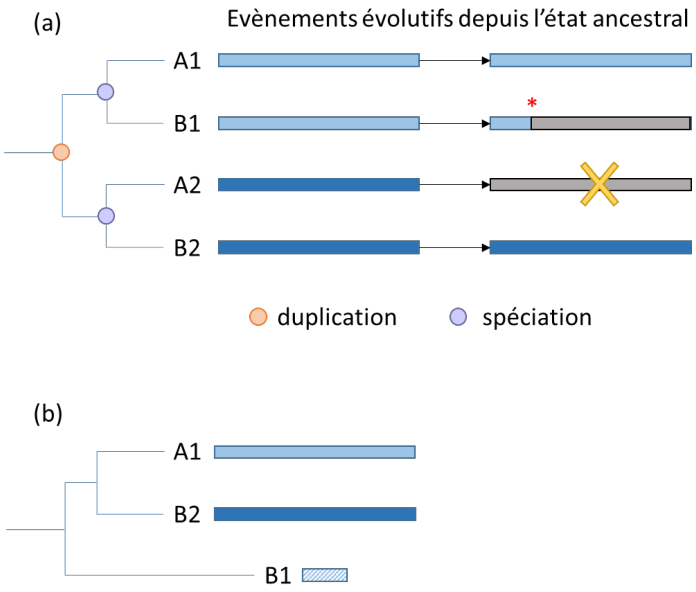


Figure 30 : Exemple de phylogénie de gènes affectés par des mutations non-sens.
(a) *Phylogénie réelle attendue entre les gènes 1 et 2 pour les espèces A et B. Les gènes 1 et 2 sont paralogues et issus d'une duplication ancestrale à la séparation des espèces A et B. Les gènes A1 et B1 sont orthologues. De même que les gènes A2 et B2. Au cours du temps, le gène 1 de l'espèce B a acquis un codon stop prématuré (étoile rouge) et le gène 2 de l'espèce A a été perdu. (b) En fonction de l'annotation du gène B1 (non-canonique), les gènes A1 et B2 peuvent être considérés comme orthologues car le gène B1 est incomplet.*

D.2. Analyses évolutives des récepteurs à LRR

La spécificité des récepteurs à LRR pour un ligand est majoritairement portée par le domaine LRR. Etudier l'évolution du domaine LRR est fondamental pour comprendre comment évolue cette spécificité, et comment émergent, et se maintiennent, les nouvelles fonctions de reconnaissance, c'est-à-dire comment un récepteur peut acquérir la capacité d'identifier un nouveau ligand. Un prérequis à ces analyses est de pouvoir annoter ce domaine, et l'ensemble des motifs qui le composent pour pouvoir faire des comparaisons fiables. Au cours de cette thèse, j'ai proposé un outil dédié à la détection et à l'annotation des motifs LRR dans les récepteurs à LRR : LRRprofiler (Chapitre 1). Cet outil fournit une annotation standardisée des motifs LRR qui est comparable entre protéines, entre familles et entre différents protéomes.

Il a été démontré que certains de ces gènes étaient sous sélection positive, notamment au niveau du domaine LRR (Mondragon-Palomino et al., 2002; Wang et al., 2011; Fischer et al., 2016). Des traces de sélection positives ont pu être détectées spécifiquement sur certains résidus. Les résidus concernés correspondent majoritairement à des résidus du segment conservé HCS exposés au solvant (Kobe and Deisenhofer, 1994; Jones and Jones, 1997; Chen, 2021), donc probablement impliqués dans les mécanismes d'interaction avec les ligands. Il a également été montré que le domaine LRR pouvait être affecté par des duplication de motifs et des évènements de conversion génique (Mondragon-Palomino and Gaut, 2005; Bjorklund et al., 2006). L'étude de l'évolution du domaine LRR est donc difficile car l'ensemble de ces mécanismes peuvent avoir cours en parallèle et peuvent alors constituer des signaux confondants impactant les résultats des analyses. Par ailleurs, nous avons observé que plus de 30% des gènes codants les récepteurs à LRR étaient non-canoniques, et que nombre d'entre eux étaient mal annotés ou absents des données d'annotation publiques. Les gènes présentant des mutations non-sens peuvent être non-fonctionnels et par conséquent être soumis à des forces évolutives différentes (relâchement de la sélection purifiante) de celles que subissent les gènes complets fonctionnels de la même famille (sélection purifiante ou diversifiante). Ne pas distinguer un gène partiel ou non-canonique des autres gènes peut alors induire des biais dans l'analyse des mécanismes évolutifs à l'échelle globale du génome, et dans la détermination de l'importance relative de ces différents mécanismes (mutations ponctuelles, indels de motifs, duplication de motifs ou conversion génique).

J'ai pu commencer à documenter différents paramètres évolutifs du domaine LRR à travers des analyses préliminaires s'appuyant sur la comparaison de ce domaine au sein des gènes orthologues identifiés entre Nipponbare (*O. sativa*) et l'accession W1943 de l'espèce *O. rufipogon*. Plusieurs études ont mis en évidence des différences de taux d'évolution (dN, dN/dS) entre les trois familles de récepteurs et entre les domaines au sein des gènes (Wang et al., 2011). En particulier, des traces de sélection ont été identifiées dans le domaine LRR (Mondragon-Palomino et al., 2002; Fischer et al., 2016). J'ai commencé par comparer les taux de substitutions observés dans le domaine LRR entre les trois familles d'intérêt (LRR-RLK, LRR-RLP et NLR). D'après ces analyses, le domaine LRR des gènes NLR semble accumuler plus de substitutions que le domaine LRR des deux autres familles de gènes. Néanmoins, des variations existent entre les catégories de gènes (canoniques ou non-canoniques). Si cette

observation est vraie pour les orthologues canoniques dans les deux espèces, elle n'est pas retrouvée pour les orthologues non-canoniques chez les deux espèces, ou canoniques chez l'une et non-canonique chez l'autre. Par ailleurs, aucune différence significative du taux de substitution n'est observée entre les autres domaines et le domaine LRR pour un même gène (kinase vs LRR pour les LRR-RLK et NB-ARC vs LRR pour les NLR). Donc les domaines LRR, dans leur globalité, ne semble pas accumuler pas plus de substitutions que le reste du gène. En revanche, des différences pourraient être observées dans la capacité d'un gène à purger ou fixer une mutation dans sa séquence. Des analyses complémentaires telles que celles faites sur les taux de substitution synonymes vs non-synonymes (dN/dS) permettraient de répondre à ces questions. Cependant, le temps de divergence entre les cultivars *japonica* et l'espèce apparentée sauvage *O. Rufipogon*, estimé à environ seulement 10 000 ans, et associé aux effets de la domestication, est court pour utiliser de telles méthodes. De nouvelles analyses devront être initiées en intégrant davantage d'espèces sauvages mieux réparties au sein du genre *Oryza*, comme par exemple *O. meridionalis* ou *O. longistaminata* pour les génomes AA (Figure 4).

De nombreuse questions restent à explorer pour mieux comprendre l'évolution du domaine LRR. Par exemple, est-ce que l'occurrence de mutations non-synonymes dans un motif LRR augmente la probabilité de mutations dans les motifs voisins ? Ceci pourrait être observé comme un effet de compensation des mutations pour stabiliser la structure 3D de la protéine. Ou encore, est-ce que les motifs LRR à proximité des autres domaines sont plus contraints que les autres ? En effet, la partie N-terminale du domaine LRR dans les NLR pourrait interagir avec le domaine NB-ARC, formant un repliement stable en absence de ligand pour inhiber le gène (Hwang and Williamson, 2003; Marone et al., 2013). De même, la partie C-terminale du domaine LRR chez les LRR-RLK et LRR-RLP serait impliquée dans la formation du complexe récepteur/corécepteur.

Parallèlement à ces perspectives d'analyses à court terme de l'évolution des domaines LRR sur le jeu de données expertisé au cours de ma thèse, l'ensemble de mes travaux a permis également l'élaboration de nouveaux projets de recherche. L'un d'entre eux a pour but d'exploiter les génomes de qualité PSRefSeq (Zhou et al., 2020) pour construire les pangénomes des récepteurs à LRR au sein des différents compartiments du riz asiatique et d'étudier leur évolution à différentes échelles de temps (structure de la diversité et recherche

de traces de sélection). Un autre axe concerne l'utilisation des méthodologies, outils et concepts que j'ai développé pour analyser d'autres familles de gènes d'intérêt, notamment certains facteurs de transcription impliqués dans la résistance à la submersion chez le riz.

Annexes

Annexe 1

Résultats d'identification des protéines LRR-RLK et motifs LRR lors de la procédure d'amélioration itérative des profils HMM.

itération	Nb Protéines	Nb motifs LRR	Longueur moyenne des motifs LRR	Nombre de résidus totaux dans les motifs LRR
0	171	1523	14.606	22245
1	172	1988	20.0201	39800
2	172	2009	19.8203	39819
3	172	2016	20.2376	40799
4	172	2019	20.1689	40721
5	172	2027	20.1707	40886
6	172	2021	20.0371	40495
7	172	2024	19.9921	40464

Annexe 2

Résultats d'identification des protéines NLR et motifs LRR lors de la procédure d'amélioration itérative des profils HMM.

Itération	Nb Protéines	Nb motifs LRR	Longueur moyenne des motifs LRR	Nombre de résidus totaux dans les motifs LRR
0	43	124	13.7097	1700
1	77	400	15.1725	6069
2	87	784	16.3878	12848
3	88	850	16.6012	14111
4	88	887	16.6573	14775
5	89	874	16.7735	14660
6	89	905	16.8	15204
7	89	716	16.8324	12052
8	88	871	16.8289	14658
9	88	849	16.861	14315

Annexe 3

Comparaison des 38 gènes LRR analysés par Mizuno *et al.* avec les données ré-annotées pour le cultivar Nipponbare. NC = non-canonique, C = canonique.

Identifiant des gènes mizuno *et al.*	pseudogènes	commentaire	Identifiants dans les données ré-annotées	NC/C
RG001			OSJnip_Chr11_27221855	C
RG002			OSJnip_Chr11_27234199	NC
RG003	pseudo		OSJnip_Chr11_27242070	NC
RG004			OSJnip_Chr11_27261093	C
RG005			OSJnip_Chr11_27269586	NC
RG006			OSJnip_Chr11_27282232	C
RG007			OSJnip_Chr11_27318942	NC
RG008			OSJnip_Chr11_27335214	C
RG009			OSJnip_Chr11_27342028	C
RG010			OSJnip_Chr11_27350120	C
RG011	pseudo		OSJnip_Chr11_27427776	NC
RG012	pseudo		OSJnip_Chr11_27604492	NC
RG013		putative Pi-ta	OSJnip_Chr11_27690035	C
RG014			OSJnip_Chr11_27691206	NC
RG015	pseudo		*Non identifiable*	
RG016			OSJnip_Chr11_27707213	C
RG017	pseudo		OSJnip_Chr11_27734371	NC
RG018			OSJnip_Chr11_27792722	NC
RG019			OSJnip_Chr11_27793837	C
RG020			OSJnip_Chr11_27818431	C
RG021			OSJnip_Chr11_27820309	C
RG022	pseudo		OSJnip_Chr11_27885376	NC
RG023	pseudo		OSJnip_Chr11_27894560	NC
RG024	pseudo		OSJnip_Chr11_27920064	NC
RG025		Pikm1	OSJnip_Chr11_27983597	C
RG026		Pikm2	OSJnip_Chr11_27985583	C
RLK001			OSJnip_Chr11_28224689	NC
RLK002			OSJnip_Chr11_28228913	NC
RLK003			OSJnip_Chr11_28236014	C
RLK004		Xa3/Xa26	OSJnip_Chr11_28246863	NC
RLK005			OSJnip_Chr11_28259210	C
RLK006	pseudo	inversion par rapport à la séquence de référence	OSJnip_Chr11_28454816	NC
RLK007			OSJnip_Chr11_28444371	C
RLK008			OSJnip_Chr11_28438467	NC
RLK009			OSJnip_Chr11_28411141	NC
RLK010			OSJnip_Chr11_28399360	NC
RLK011			OSJnip_Chr11_28376327	C
RLK012	pseudo		OSJnip_Chr11_28371257	NC

Annexe 4

Principaux génomes de riz disponibles et pourcentage d'ambiguïtés dans leurs séquences. Les séquences génomiques utilisées au cours de la thèse sont surlignées en orange.

Génome	Oryza	ssp/variant	version	nb scaffolds	Taille	%N
AA	barthii		1	12	313 410 183	0.8301
	glaberrima		AGI1.1.39	1951	321 694 182	4.0737
	glaberrima	G22	draft	49 662	297 116 702	0.9096
	glaberrima	CG14	draft	809 520	360 837 976	0.1765
	glaberrima	TOG5681	draft	745 779	338 389 125	1.5794
	glumeapatula		1.5	12	379 074 629	18.8863
	meridionalis		1.3	12	341 262 709	23.4544
	nivara		1.0	12	343 582 836	8.1365
	rufipogon	W1943		12	343 674 732	1.8836
	sativa	aus Kasalath		13	409 164 547	17.8140
	sativa	aus N22		12	368 317 088	0.0251
	sativa	aus DJ123	draft1.0	2 819	350 307 913	5.7030
	sativa	aus NATEL BORO		19	389 006 379	0.0001
	sativa	indica IR8		15	395 573 180	0.0016
	sativa	indica IR64		21	392 327 360	0.0150
	sativa	indica MH63	2	12	392 267 170	0.0006
	sativa	indica R498		14	390 983 864	0.0128
	sativa	indica KHAO YAI GUANG		46	404 239 987	0.0002
	sativa	indica ZS97		12	392 167 970	0.0004
	sativa	indica BALAM		575	496 281 742	0.0001
	sativa	indica PR 106		40	399 534 444	0.0001
	sativa	indica LIMA		93	410 466 678	0.0001
	sativa	japonica KitaakeX	3.1	33	387 930 332	1.0314
	sativa	japonica Nipponbare	7.0	14	380 712 436	0.0434
	sativa	japonica KETAN NANGKA		25	386 782 488	0.0002
	sativa	japonica CHAO MEO		81	388 033 956	0.0011
	sativa	japonica Azucena		37	386 339 773	0.0004
	sativa	aromatic ARC 10497		41	385 108 637	0.0007
	longistaminata		1.0	60 198	331 912 383	9.6366
BB	punctata			12	400 380 220	0.4157
CC	officinalis					
EE	australiensis					
FF	brachyanta		1.4b	7 485	265 188 840	6.6099
GG	granulata					

Références Bibliographiques

Aan den Toorn, M., Albrecht, C., and de Vries, S. (2015). On the Origin of SERKs: Bioinformatics Analysis of the Somatic Embryogenesis Receptor Kinases. Mol Plant **8,** 762-782.

Albert, I., Bohm, H., Albert, M., Feiler, C.E., Imkampe, J., Wallmeroth, N., Brancato, C., Raaymakers, T.M., Oome, S., Zhang, H., Krol, E., Grefen, C., Gust, A.A., Chai, J., Hedrich, R., Van den Ackerveken, G., and Nurnberger, T. (2015). An RLP23-SOBIR1-BAK1 complex mediates NLP-triggered immunity. Nat Plants **1,** 15140.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol **215,** 403-410.

Amid, C., Rehaume, L.M., Brown, K.L., Gilbert, J.G., Dougan, G., Hancock, R.E., and Harrow, J.L. (2009). Manual annotation and analysis of the defensin gene cluster in the C57BL/6J mouse reference genome. BMC Genomics **10,** 606.

Andolfo, G., Jupe, F., Witek, K., Etherington, G.J., Ercolano, M.R., and Jones, J.D. (2014). Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. BMC Plant Biol **14,** 120.

Andrade, M.A., Ponting, C.P., Gibson, T.J., and Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. J Mol Biol **298,** 521-537.

Anisimova, M., Pecerska, J., and Schaper, E. (2015). Statistical approaches to detecting and analyzing tandem repeats in genomic sequences. Front Bioeng Biotechnol **3,** 31.

Apic, G., Gough, J., and Teichmann, S.A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol **310,** 311-325.

Ashikawa, I., Hayashi, N., Yamane, H., Kanamori, H., Wu, J., Matsumoto, T., Ono, K., and Yano, M. (2008). Two adjacent nucleotide-binding site-leucine-rich repeat class genes are required to confer Pikm-specific rice blast resistance. Genetics **180,** 2267-2276.

Ausubel, F.M. (2005). Are innate immune signaling pathways in plants and animals conserved? Nat Immunol **6,** 973-979.

Bailey, P.C., Schudoma, C., Jackson, W., Baggs, E., Dagdas, G., Haerty, W., Moscou, M., and Krasileva, K.V. (2018). Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. Genome Biol **19,** 23.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res **34,** W369-373.

Baite, M.S., Raghu, S., Prabhukarthikeyan, S.R., Keerthana, U., Jambhulkar, N.N., and Rath, P.C. (2020). Disease incidence and yield loss in rice due to grain discolouration. Journal of Plant Diseases and Protection **127,** 9-13.

Baker, B., Zambryski, P., Staskawicz, B., and Dinesh-Kumar, S.P. (1997). Signaling in plant-microbe interactions. Science **276,** 726-733.

Baumgarten, A., Cannon, S., Spangler, R., and May, G. (2003). Genome-level evolution of resistance genes in Arabidopsis thaliana. Genetics **165,** 309-319.

Bayer, P.E., Edwards, D., and Batley, J. (2018). Bias in resistance gene prediction due to repeat masking. Nat Plants **4,** 762-765.

Bayer, P.E., Golicz, A.A., Tirnaz, S., Chan, C.K., Edwards, D., and Batley, J. (2019). Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. Plant Biotechnol J **17,** 789-800.

Bebber, D.P., Ramotowski, M.A.T., and Gurr, S.J. (2013). Crop pests and pathogens move polewards in a warming world. Nature Climate Change **3,** 985-988.

Bella, J., Hindle, K.L., McEwan, P.A., and Lovell, S.C. (2008). The leucine-rich repeat structure. Cell Mol Life Sci **65,** 2307-2333.

Biscotti, M.A., Olmo, E., and Heslop-Harrison, J.S. (2015). Repetitive DNA in eukaryotic genomes. Chromosome Res **23,** 415-420.

Bjorklund, A.K., Ekman, D., and Elofsson, A. (2006). Expansion of protein domain repeats. PLoS Comput Biol **2,** e114.

Bjorklund, A.K., Ekman, D., Light, S., Frey-Skott, J., and Elofsson, A. (2005). Domain rearrangements in protein evolution. J Mol Biol **353,** 911-923.

Bolger, M.E., Arsova, B., and Usadel, B. (2018). Plant genome and transcriptome annotations: from misconceptions to simple solutions. Brief Bioinform **19,** 437-449.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). UniProtKB/Swiss-Prot. Methods Mol Biol **406,** 89-112.

Boutrot, F., and Zipfel, C. (2017). Function, Discovery, and Exploitation of Plant Pattern Recognition Receptors for Broad-Spectrum Disease Resistance. Annu Rev Phytopathol **55,** 257-286.

Buchanan, S.G., and Gay, N.J. (1996). Structural and functional diversity in the leucine-rich repeat family of proteins. Prog Biophys Mol Biol **65,** 1-44.

Cannon, S.B., Zhu, H., Baumgarten, A.M., Spangler, R., May, G., Cook, D.R., and Young, N.D. (2002). Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. J Mol Evol **54,** 548-562.

Cesari, S. (2018). Multiple strategies for pathogen perception by plant immune receptors. New Phytol **219,** 17-24.

Cesari, S., Thilliez, G., Ribot, C., Chalvon, V., Michel, C., Jauneau, A., Rivas, S., Alaux, L., Kanzaki, H., Okuyama, Y., Morel, J.B., Fournier, E., Tharreau, D., Terauchi, R., and Kroj, T. (2013). The rice resistance protein pair RGA4/RGA5 recognizes the Magnaporthe oryzae effectors AVR-Pia and AVR1-CO39 by direct binding. Plant Cell **25,** 1463-1481.

CGIAR (Consultative Group on International Agricultural Research). (2018). RICE Annual Report. (Los Bãnos (Philippines): International Rice Research Institute (IRRI)).

Cheetham, S.W., Faulkner, G.J., and Dinger, M.E. (2020). Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet **21,** 191-201.

Chen, E., Huang, X., Tian, Z., Wing, R.A., and Han, B. (2019). The Genomics of Oryza Species Provides Insights into Rice Domestication and Heterosis. Annu Rev Plant Biol **70,** 639-665.

Chen, T. (2021). Identification and characterization of the LRR repeats in plant LRR-RLKs. BMC Mol Cell Biol **22,** 9.

Cheng, C., Motohashi, R., Tsuchimoto, S., Fukuta, Y., Ohtsubo, H., and Ohtsubo, E. (2003). Polyphyletic origin of cultivated rice: based on the interspersion pattern of SINEs. Mol Biol Evol **20,** 67-75.

Cheng, S., Melkonian, M., Smith, S.A., Brockington, S., Archibald, J.M., Delaux, P.M., Li, F.W., Melkonian, B., Mavrodiev, E.V., Sun, W., Fu, Y., Yang, H., Soltis, D.E., Graham, S.W., Soltis, P.S., Liu, X., Xu, X., and Wong, G.K. (2018). 10KP: A phylodiverse genome sequencing plan. Gigascience **7,** 1-9.

Choi, J.Y., Platts, A.E., Fuller, D.Q., Hsing, Y.I., Wing, R.A., and Purugganan, M.D. (2017). The Rice Paradox: Multiple Origins but Single Domestication in Asian Rice. Mol Biol Evol **34,** 969-979.

Civan, P., Craig, H., Cox, C.J., and Brown, T.A. (2015). Three geographically separate domestications of Asian rice. Nat Plants **1,** 15164.

Couto, D., and Zipfel, C. (2016). Regulation of pattern recognition receptor signalling in plants. Nat Rev Immunol **16,** 537-552.

Cubry, P., Tranchant-Dubreuil, C., Thuillet, A.C., Monat, C., Ndjiondjop, M.N., Labadie, K., Cruaud, C., Engelen, S., Scarcelli, N., Rhone, B., Burgarella, C., Dupuy, C., Larmande, P., Wincker, P., Francois, O., Sabot, F., and Vigouroux, Y. (2018). The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. Curr Biol **28,** 2274-2282 e2276.

Cui, H., Tsuda, K., and Parker, J.E. (2015). Effector-triggered immunity: from pathogen perception to robust defense. Annu Rev Plant Biol **66,** 487-511.

Dalquen, D.A., and Dessimoz, C. (2013). Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol Evol **5,** 1800-1806.

Dalquen, D.A., Altenhoff, A.M., Gonnet, G.H., and Dessimoz, C. (2013). The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PLoS One **8,** e56925.

Dawkins, R., and Krebs, J.R. (1979). Arms races between and within species. Proc R Soc Lond B Biol Sci **205,** 489-511.

DeFalco, T.A., and Zipfel, C. (2021). Molecular mechanisms of early plant pattern-triggered immune signaling. Mol Cell.

Deslandes, L., Pileur, F., Liaubet, L., Camut, S., Can, C., Williams, K., Holub, E., Beynon, J., Arlat, M., and Marco, Y. (1998). Genetic characterization of RRS1, a recessive locus in Arabidopsis thaliana that confers resistance to the bacterial soilborne pathogen Ralstonia solanacearum. Mol Plant Microbe Interact **11,** 659-667.

Die, J.V., Castro, P., Millan, T., and Gil, J. (2018). Segmental and Tandem Duplications Driving the Recent NBS-LRR Gene Expansion in the Asparagus Genome. Genes (Basel) **9**.

Dievart, A., Gottin, C., Perin, C., Ranwez, V., and Chantret, N. (2020). Origin and Diversity of Plant Receptor-Like Kinases. Annu Rev Plant Biol **71,** 131-156.

Dievart, A., Gilbert, N., Droc, G., Attard, A., Gourgues, M., Guiderdoni, E., and Perin, C. (2011). Leucine-rich repeat receptor kinases are sporadically distributed in eukaryotic genomes. BMC Evol Biol **11,** 367.

Dodds, P.N., and Rathjen, J.P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. Nat Rev Genet **11,** 539-548.

Dodds, P.N., Lawrence, G.J., Catanzariti, A.M., Teh, T., Wang, C.I., Ayliffe, M.A., Kobe, B., and Ellis, J.G. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. Proc Natl Acad Sci U S A **103,** 8888-8893.

Dolatabadian, A., Patel, D.A., Edwards, D., and Batley, J. (2017). Copy number variation and disease resistance in plants. Theor Appl Genet **130,** 2479-2490.

Dufayard, J.F., Bettembourg, M., Fischer, I., Droc, G., Guiderdoni, E., Perin, C., Chantret, N., and Dievart, A. (2017). New Insights on Leucine-Rich Repeats Receptor-Like Kinase Orthologous Relationships in Angiosperms. Front Plant Sci **8,** 381.

Eddy, S.R. (1996). Hidden Markov models. Curr Opin Struct Biol **6,** 361-365.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14,** 755-763.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol **7,** e1002195.

FAO. (2004). The State of Food and Agriculture 2003-2004. (Rome: FAO).

FAO. (2017). The future of food and agriculture - Trends and challenges. (Rome: FAO).

FAO. (2019). FAOSTAT Statistical Database (Rome: FAO).

FAO. (2020). FAOSTAT Statistical Database (Rome: FAO).

FAO. (2021). Food Outlook - Biannual Report on Global Food Markets. (Rome).

FAO, IFAD, UNICEFF, WFP, and WHO. (2020). The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets. (Rome FAO).

Fawal, N., Li, Q., Mathe, C., and Dunand, C. (2014). Automatic multigenic family annotation: risks and solutions. Trends Genet **30,** 323-325.

Fischer, I., Dievart, A., Droc, G., Dufayard, J.F., and Chantret, N. (2016). Evolutionary Dynamics of the Leucine-Rich Repeat Receptor-Like Kinase (LRR-RLK) Subfamily in Angiosperms. Plant Physiol **170,** 1595-1610.

Flick, J.S., and Johnston, M. (1991). GRR1 of Saccharomyces cerevisiae is required for glucose repression and encodes a protein with leucine-rich repeats. Mol Cell Biol **11,** 5101-5112.

Forsthoefel, N.R., Cutler, K., Port, M.D., Yamamoto, T., and Vernon, D.M. (2005). PIRLs: a novel class of plant intracellular leucine-rich repeat proteins. Plant Cell Physiol **46,** 913-922.

Friedman, A.R., and Baker, B.J. (2007). The evolution of resistance genes in multi-protein plant resistance systems. Curr Opin Genet Dev **17,** 493-499.

Fritz-Laylin, L.K., Krishnamurthy, N., Tor, M., Sjolander, K.V., and Jones, J.D. (2005). Phylogenomic analysis of the receptor-like proteins of rice and Arabidopsis. Plant Physiol **138,** 611-623.

Fuentes, R.R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J.F., Mohiyuddin, M., Wing, R.A., McNally, K.L., Tatarinova, T., Grigoriev, A., Mauleon, R., and Alexandrov, N. (2019). Structural variants in 3000 rice genomes. Genome Res **29,** 870-880.

Garris, A.J., Tai, T.H., Coburn, J., Kresovich, S., and McCouch, S. (2005). Genetic structure and diversity in Oryza sativa L. Genetics **169,** 1631-1638.

Glaszmann, J.C. (1987). Isozymes and classification of Asian rice varieties. Theor Appl Genet **74,** 21-30.

Golicz, A.A., Bayer, P.E., Bhalla, P.L., Batley, J., and Edwards, D. (2020). Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. Trends Genet **36,** 132-145.

Gomez-Gomez, L., and Boller, T. (2000). FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. Mol Cell **5,** 1003-1011.

Goodman, S.N., Fanelli, D., and Ioannidis, J.P. (2016). What does research reproducibility mean? Sci Transl Med **8,** 341ps312.

Gottin, C., Dievart, A., Summo, M., Droc, G., Perin, C., Ranwez, V., and Chantret, N. (2021). A new comprehensive annotation of leucine-rich repeat-containing receptors in rice. Plant J.

GRiSP (Global Rice Science Partnership). (2013). Rice almanac, 4th edition. (Los Baños (Philippines): International Rice Research Institute (IRRI)).

Guo, Y.L., Fitz, J., Schneeberger, K., Ossowski, S., Cao, J., and Weigel, D. (2011). Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis. Plant Physiol **157,** 757-769.

Ham, J.H., Melanson, R.A., and Rush, M.C. (2011). Burkholderia glumae: next major pathogen of rice? Mol Plant Pathol **12,** 329-339.

Hanks, S.K., and Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J **9,** 576-596.

Hanks, S.K., Quinn, A.M., and Hunter, T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science **241,** 42-52.

Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics **107,** 1-8.

Hecht, V., Vielle-Calzada, J.P., Hartog, M.V., Schmidt, E.D., Boutilier, K., Grossniklaus, U., and de Vries, S.C. (2001). The Arabidopsis SOMATIC EMBRYOGENESIS RECEPTOR KINASE 1 gene is expressed in developing ovules and embryos and enhances embryogenic competence in culture. Plant Physiol **127,** 803-816.

Hickey, D.A., Bally-Cuif, L., Abukashawa, S., Payant, V., and Benkel, B.F. (1991). Concerted evolution of duplicated protein-coding genes in Drosophila. Proc Natl Acad Sci U S A **88,** 1611-1615.

Hofsteenge, J., Kieffer, B., Matthies, R., Hemmings, B.A., and Stone, S.R. (1988). Amino acid sequence of the ribonuclease inhibitor from porcine liver reveals the presence of leucine-rich repeats. Biochemistry **27,** 8537-8544.

Holub, E.B. (2001). The arms race is ancient history in Arabidopsis, the wildflower. Nat Rev Genet **2,** 516-527.

Hothorn, M., Belkhadir, Y., Dreux, M., Dabi, T., Noel, J.P., Wilson, I.A., and Chory, J. (2011). Structural basis of steroid hormone perception by the receptor kinase BRI1. Nature **474,** 467-471.

Huang, X., and Han, B. (2015). Rice domestication occurred through single origin and multiple introgressions. Nat Plants **2,** 15207.

Huang, X., Kurata, N., Wei, X., Wang, Z.X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., Guo, Y., Lu, Y., Zhou, C., Fan, D., Weng, Q., Zhu, C., Huang, T., Zhang, L., Wang, Y., Feng, L., Furuumi, H., Kubo, T., Miyabayashi, T., Yuan, X., Xu, Q., Dong, G., Zhan, Q., Li, C., Fujiyama, A., Toyoda, A., Lu, T., Feng, Q., Qian, Q., Li, J., and Han, B. (2012). A map of rice genome variation reveals the origin of cultivated rice. Nature **490,** 497-501.

Hwang, C.F., and Williamson, V.M. (2003). Leucine-rich repeat-mediated intramolecular interactions in nematode recognition and cell death signaling by the tomato resistance protein Mi. Plant J **34,** 585-593.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. Nature **431,** 931-945.

IRGSP (International Rice Genome Sequencing Project), and Sasaki, T. (2005). The map-based sequence of the rice genome. Nature **436,** 793-800.

Jacquemin, J., Bhatia, D., Singh, K., and Wing, R.A. (2013). The International Oryza Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. Curr Opin Plant Biol **16,** 147-156.

Jain, R., Jenkins, J., Shu, S., Chern, M., Martin, J.A., Copetti, D., Duong, P.Q., Pham, N.T., Kudrna, D.A., Talag, J., Schackwitz, W.S., Lipzen, A.M., Dilworth, D., Bauer, D., Grimwood, J., Nelson, C.R., Xing, F., Xie, W., Barry, K.W., Wing, R.A., Schmutz, J., Li, G., and Ronald, P.C. (2019). Genome sequence of the model rice variety KitaakeX. BMC Genomics **20,** 905.

Janeway, C.A., Jr., and Medzhitov, R. (2002). Innate immune recognition. Annu Rev Immunol **20,** 197-216.

Jones, D.A., and Jones, J.D.G. (1997). The Role of Leucine-Rich Repeat Proteins in Plant Defences. In Advances in Botanical Research, J.H. Andrews, I.C. Tommerup, and J.A. Callow, eds (Academic Press), pp. 89-167.

Jones, D.A., Thomas, C.M., Hammond-Kosack, K.E., Balint-Kurti, P.J., and Jones, J.D. (1994). Isolation of the tomato Cf-9 gene for resistance to Cladosporium fulvum by transposon tagging. Science **266,** 789-793.

Jones, J.D., Vance, R.E., and Dangl, J.L. (2016). Intracellular innate immune surveillance devices in plants and animals. Science **354**.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30,** 1236-1240.

Jorda, J., Xue, B., Uversky, V.N., and Kajava, A.V. (2010). Protein tandem repeats - the more perfect, the less structured. FEBS J **277,** 2673-2682.

Jupe, F., Pritchard, L., Etherington, G.J., Mackenzie, K., Cock, P.J., Wright, F., Sharma, S.K., Bolser, D., Bryan, G.J., Jones, J.D., and Hein, I. (2012). Identification and localisation of the NB-LRR gene family within the potato genome. BMC Genomics **13,** 75.

Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G.J., Maclean, D., Cock, P.J., Leggett, R.M., Bryan, G.J., Cardle, L., Hein, I., and Jones, J.D. (2013). Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. Plant J **76,** 530-544.

Kajava, A.V. (1998). Structural diversity of leucine-rich repeat proteins. J Mol Biol **277,** 519-527.

Kajava, A.V. (2012). Tandem repeats in proteins: from sequence to structure. J Struct Biol **179,** 279-288.

Kajava, A.V., Vassart, G., and Wodak, S.J. (1995). Modeling of the three-dimensional structure of proteins with the typical leucine-rich repeats. Structure **3,** 867-877.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30,** 772-780.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., Childs, K.L., Davidson, R.M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S.S., Kim, J., Numa, H., Itoh, T., Buell, C.R., and Matsumoto, T. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) **6,** 4.

Kim, S., Cheong, K., Park, J., Kim, M.S., Kim, J., Seo, M.K., Chae, G.Y., Jang, M.J., Mang, H., Kwon, S.H., Kim, Y.M., Koo, N., Min, C.W., Kim, K.S., Oh, N., Kim, K.T., Jeon, J., Kim, H., Lee, Y.Y., Sohn, K.H., McCann, H.C., Ye, S.K., Kim, S.T., Park, K.S., Lee, Y.H., and Choi, D. (2020). TGFam-Finder: a novel solution for target-gene family annotation in plants. New Phytol **227,** 1568-1581.

Kobe, B., and Deisenhofer, J. (1993). Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. Nature **366,** 751-756.

Kobe, B., and Deisenhofer, J. (1994). The leucine-rich repeat: a versatile binding motif. Trends Biochem Sci **19,** 415-421.

Kobe, B., and Kajava, A.V. (2001). The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol **11,** 725-732.

König, S., Romoth, L., and Stanke, M. (2018). Comparative Genome Annotation. In Comparative Genomics: Methods and Protocols, J.C. Setubal, J. Stoye, and P. Stadler, eds (Humana Press.

Kovalenko, T.F., and Patrushev, L.I. (2018). Pseudogenes as Functionally Significant Elements of the Genome. Biochemistry (Mosc) **83,** 1332-1349.

Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X., and Morel, J.B. (2016). Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. New Phytol **210,** 618-626.

Kuang, H., Woo, S.S., Meyers, B.C., Nevo, E., and Michelmore, R.W. (2004). Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. Plant Cell **16,** 2870-2894.

Leister, D. (2004). Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. Trends Genet **20,** 116-122.

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. Nucleic Acids Res **49,** D458-D460.

Li, J., Wen, J., Lease, K.A., Doke, J.T., Tax, F.E., and Walker, J.C. (2002). BAK1, an Arabidopsis LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. Cell **110,** 213-222.

Li, J., Ding, J., Zhang, W., Zhang, Y., Tang, P., Chen, J.Q., Tian, D., and Yang, S. (2010). Unique evolutionary pattern of numbers of gramineous NBS-LRR genes. Mol Genet Genomics **283,** 427-438.

Liu, P.L., Du, L., Huang, Y., Gao, S.M., and Yu, M. (2017). Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. BMC Evol Biol **17,** 47.

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science **252,** 1162-1164.

Lv, H., Fang, Z., Yang, L., Zhang, Y., Wang, Q., Liu, Y., Zhuang, M., Yang, Y., Xie, B., Liu, B., Liu, J., Kang, J., and Wang, X. (2014). Mapping and analysis of a novel candidate Fusarium wilt resistance gene FOC1 in Brassica oleracea. BMC Genomics **15,** 1094.

Ma, X., Xu, G., He, P., and Shan, L. (2016). SERKing Coreceptors for Receptors. Trends Plant Sci **21,** 1017-1033.

Magalhaes, D.M., Scholte, L.L., Silva, N.V., Oliveira, G.C., Zipfel, C., Takita, M.A., and De Souza, A.A. (2016). LRR-RLK family from two Citrus species: genome-wide identification and evolutionary aspects. BMC Genomics **17,** 623.

Mahood, E.H., Kruse, L.H., and Moghe, G.D. (2020). Machine learning: A powerful tool for gene function prediction in plants. Appl Plant Sci **8,** e11376.

Man, J., Gallagher, J.P., and Bartlett, M. (2020). Structural evolution drives diversification of the large LRR-RLK gene family. New Phytol **226,** 1492-1505.

Marone, D., Russo, M.A., Laido, G., De Leonardis, A.M., and Mastrangelo, A.M. (2013). Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. Int J Mol Sci **14,** 7302-7326.

Martin, E.C., Sukarta, O.C.A., Spiridon, L., Grigore, L.G., Constantinescu, V., Tacutu, R., Goverse, A., and Petrescu, A.J. (2020). LRRpredictor-A New LRR Motif Detection Method for Irregular Motifs of Plant NLR Proteins Using an Ensemble of Classifiers. Genes (Basel) **11**.

Masle, J., Gilmore, S.R., and Farquhar, G.D. (2005). The ERECTA gene regulates plant transpiration efficiency in Arabidopsis. Nature **436,** 866-870.

Matsushima, N., and Kamiya, M. (1999). Super-Motifs of Leucine-Rich Repeats (LRRs) Proteins. Genome Informatics **11,** 343-345.

Matsushima, N., and Kretsinger, R.H. (2016). Leucine Rich Repeats. Sequences, Structures, Ligand-Interactions and Evolution. (Germany: LAMBERT Academic Publishing).

Matsushima, N., Miyashita, H., Mikami, T., and Kuroki, Y. (2010). A nested leucine rich repeat (LRR) domain: the precursor of LRRs is a ten or eleven residue motif. BMC Microbiol **10,** 235.

McDowell, J.M., and Simon, S.A. (2006). Recent insights into R gene evolution. Mol Plant Pathol **7,** 437-448.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell **15,** 809-834.

Meyers, B.C., Dickerman, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sobral, B.W., and Young, N.D. (1999). Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J **20,** 317-332.

Michelmore, R.W., and Meyers, B.C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res **8,** 1113-1130.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., and Bateman, A. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res **49,** D412-D419.

Mizuno, H., Katagiri, S., Kanamori, H., Mukai, Y., Sasaki, T., Matsumoto, T., and Wu, J. (2020). Evolutionary dynamics and impacts of chromosome regions carrying R-gene clusters in rice. Sci Rep **10,** 872.

Monat, C., Pera, B., Ndjiondjop, M.N., Sow, M., Tranchant-Dubreuil, C., Bastianelli, L., Ghesquiere, A., and Sabot, F. (2017). De Novo Assemblies of Three Oryza glaberrima Accessions Provide First Insights about Pan-Genome of African Rices. Genome Biol Evol **9,** 1-6.

Mondragon-Palomino, M., and Gaut, B.S. (2005). Gene conversion and the evolution of three leucine-rich repeat gene families in Arabidopsis thaliana. Mol Biol Evol **22,** 2444-2456.

Mondragon-Palomino, M., Meyers, B.C., Michelmore, R.W., and Gaut, B.S. (2002). Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Res **12,** 1305-1315.

Monosi, B., Wisser, R.J., Pennill, L., and Hulbert, S.H. (2004). Full-genome analysis of resistance gene homologues in rice. Theor Appl Genet **109,** 1434-1447.

Monteiro, F., and Nishimura, M.T. (2018). Structural, Functional, and Genomic Diversity of Plant NLR Proteins: An Evolved Resource for Rational Engineering of Plant Immunity. Annu Rev Phytopathol **56,** 243-267.

Mussurova, S., Al-Bader, N., Zuccolo, A., and Wing, R.A. (2020). Potential of Platinum Standard Reference Genomes to Exploit Natural Variation in the Wild Relatives of Rice. Front Plant Sci **11,** 579980.

Nandakumar, R., Rush, M.C., and Correa, F. (2007). Association of Burkholderia glumae and B. gladioli with Panicle Blight Symptoms on Rice in Panama. Plant Dis **91,** 767.

Nandakumar, R., Shahjahan, A.K.M., Yuan, X.L., Dickstein, E.R., Groth, D.E., Clark, C.A., Cartwright, R.D., and Rush, M.C. (2009). Burkholderia glumae and B. gladioli Cause Bacterial Panicle Blight in Rice in the Southern United States. Plant Dis **93,** 896-905.

Nei, M., and Rooney, A.P. (2005). Concerted and birth-and-death evolution of multigene families. Annu Rev Genet **39,** 121-152.

Ng, A.C., Eisenberg, J.M., Heath, R.J., Huett, A., Robinson, C.M., Nau, G.J., and Xavier, R.J. (2011). Human leucine-rich repeat proteins: a genome-wide bioinformatic categorization and functional analysis in innate immunity. Proc Natl Acad Sci U S A **108 Suppl 1,** 4631-4638.

Noel, L., Moores, T.L., van Der Biezen, E.A., Parniske, M., Daniels, M.J., Parker, J.E., and Jones, J.D. (1999). Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. Plant Cell **11,** 2099-2112.

Oerke, E.C. (2006). Crop losses to pests. The Journal of Agricultural Science **144,** 31-43.

Oh-e, I., Saitoh, K., and Kuroda, T. (2007). Effects of High Temperature on Growth, Yield and Dry-Matter Production of Rice Grown in the Paddy Field. Plant Production Science **10,** 412-422.

Ohkura, H., and Yanagida, M. (1991). S. pombe gene sds22+ essential for a midmitotic transition encodes a leucine-rich repeat protein that positively modulates protein phosphatase-1. Cell **64,** 149-157.

Oka, H.-I. (1988). Origin of Cultivated Rice. (Japan Scientific Societies Press, Tokyo, Elsevier Science Publishers, Amsterdam).

Outten, J., and Warren, A. (2021). Methods and Developments in Graphical Pangenomics. J Indian Inst Sci**,** 1-14.

Park, S., Moon, J.C., Park, Y.C., Kim, J.H., Kim, D.S., and Jang, C.S. (2014). Molecular dissection of the response of a rice leucine-rich repeat receptor-like kinase (LRR-RLK) gene to abiotic stresses. J Plant Physiol **171,** 1645-1653.

Patthy, L. (2003). Modular assembly of genes and the evolution of new functions. Genetica **118,** 217-231.

Pingali, P.L. (2012). Green revolution: impacts, limits, and the path ahead. Proc Natl Acad Sci U S A **109,** 12302-12308.

Pink, R.C., and Carter, D.R. (2013). Pseudogenes as regulators of biological function. Essays Biochem **54,** 103-112.

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E.J. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One **6,** e22594.

Ranwez, V., Douzery, E.J.P., Cambon, C., Chantret, N., and Delsuc, F. (2018). MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol Biol Evol **35,** 2582-2584.

Rice Annotation Project, Tanaka, T., Antonio, B.A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H., Wu, J., Itoh, T., Sasaki, T., Aono, R., Fujii, Y., Habara, T., Harada, E., Kanno, M., Kawahara, Y., Kawashima, H., Kubooka, H., Matsuya, A., Nakaoka, H., Saichi, N., Sanbonmatsu, R., Sato, Y., Shinso, Y., Suzuki, M., Takeda, J., Tanino, M., Todokoro, F., Yamaguchi, K., Yamamoto, N., Yamasaki, C., Imanishi, T., Okido, T., Tada, M., Ikeo, K., Tateno, Y., Gojobori, T., Lin, Y.C., Wei, F.J., Hsing, Y.I., Zhao, Q., Han, B., Kramer, M.R., McCombie, R.W., Lonsdale, D., O'Donovan, C.C., Whitfield, E.J., Apweiler, R., Koyanagi, K.O., Khurana, J.P., Raghuvanshi, S., Singh, N.K., Tyagi, A.K., Haberer, G., Fujisawa, M., Hosokawa, S., Ito, Y., Ikawa, H., Shibata, M., Yamamoto, M., Bruskiewich, R.M., Hoen, D.R., Bureau, T.E., Namiki, N., Ohyanagi, H., Sakai, Y., Nobushima, S., Sakata, K., Barrero, R.A., Sato, Y., Souvorov, A., Smith-White, B., Tatusova, T., An, S., An, G., S, O.O., Fuks, G., Fuks, G., Messing, J., Christie, K.R., Lieberherr, D., Kim, H., Zuccolo, A., Wing, R.A., Nobuta, K., Green, P.J., Lu, C., Meyers, B.C., Chaparro, C., Piegu, B., Panaud, O., and Echeverria, M. (2008). The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res **36,** D1028-1033.

Rice Genome Project (2014). The 3,000 rice genomes project. Gigascience **3,** 7.

Salamov, A.A., and Solovyev, V.V. (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Res **10,** 516-522.

Sallet, E., Gouzy, J., and Schiex, T. (2019). EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. Methods Mol Biol **1962,** 97-120.

Sang, T., and Ge, S. (2007). Genetics and phylogenetics of rice domestication. Curr Opin Genet Dev **17,** 533-538.

Santana Silva, R.J., and Micheli, F. (2020). RRGPredictor, a set-theory-based tool for predicting pathogen-associated molecular pattern receptors (PRRs) and resistance (R) proteins from plants. Genomics **112,** 2666-2676.

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. Nat Ecol Evol **3,** 430-439.

Schaper, E., and Anisimova, M. (2015). The evolution and function of protein tandem repeats in plants. New Phytol **206,** 397-410.

Schulze, B., Buhmann, M.T., Rio Bartulos, C., and Kroth, P.G. (2015). Comprehensive computational analysis of leucine-rich repeat (LRR) proteins encoded in the genome of the diatom Phaeodactylum tricornutum. Mar Genomics **21,** 43-51.

Sekhwal, M.K., Li, P., Lam, I., Wang, X., Cloutier, S., and You, F.M. (2015). Disease Resistance Gene Analogs (RGAs) in Plants. Int J Mol Sci **16,** 19248-19290.

Shao, Z.Q., Wang, B., and Chen, J.Q. (2016a). Tracking ancestral lineages and recent expansions of NBS-LRR genes in angiosperms. Plant Signal Behav **11,** e1197470.

Shao, Z.Q., Xue, J.Y., Wu, P., Zhang, Y.M., Wu, Y., Hang, Y.Y., Wang, B., and Chen, J.Q. (2016b). Large-Scale Analyses of Angiosperm Nucleotide-Binding Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns. Plant Physiol **170,** 2095-2109.

She, J., Han, Z., Kim, T.W., Wang, J., Cheng, W., Chang, J., Shi, S., Wang, J., Yang, M., Wang, Z.Y., and Chai, J. (2011). Structural insight into brassinosteroid perception by BRI1. Nature **474,** 472-476.

Shevchenko, D.V., Akins, D.R., Robinson, E., Li, M., Popova, T.G., Cox, D.L., and Radolf, J.D. (1997). Molecular characterization and cellular localization of TpLRR, a processed leucine-rich repeat protein of Treponema pallidum, the syphilis spirochete. J Bacteriol **179,** 3188-3195.

Shiu, S.H., and Bleecker, A.B. (2001a). Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. Proc Natl Acad Sci U S A **98,** 10763-10768.

Shiu, S.H., and Bleecker, A.B. (2001b). Plant receptor-like kinase gene family: diversity, function, and signaling. Sci STKE **2001,** re22.

Shiu, S.H., and Bleecker, A.B. (2003). Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in Arabidopsis. Plant Physiol **132,** 530-543.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31,** 3210-3212.

Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol **7 Suppl 1,** S10 11-12.

Song, W.Y., Wang, G.L., Chen, L.L., Kim, H.S., Pi, L.Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.X., Zhu, L.H., Fauquet, C., and Ronald, P. (1995). A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. Science **270,** 1804-1806.

Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol **6,** 175-182.

Souvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, V., Tatusova, T., and Lipman, D. (2010). Gnomon - NCBI eukaryotic gene prediction tool.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19 Suppl 2,** ii215-225.

Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C., Chougule, K., Gao, D., Iwata, A., Goicoechea, J.L., Wei, S., Wang, J., Liao, Y., Wang, M., Jacquemin, J., Becker, C., Kudrna, D., Zhang, J., Londono, C.E.M., Song, X., Lee, S., Sanchez, P., Zuccolo, A., Ammiraju, J.S.S., Talag, J., Danowitz, A., Rivera, L.F., Gschwend, A.R., Noutsos, C., Wu, C.C., Kao, S.M., Zeng, J.W., Wei, F.J., Zhao, Q., Feng, Q., El Baidouri, M., Carpentier, M.C., Lasserre, E., Cooke, R., Rosa Farias, D.D., da Maia, L.C., Dos Santos, R.S., Nyberg, K.G., McNally, K.L., Mauleon, R., Alexandrov, N., Schmutz, J., Flowers, D., Fan, C., Weigel, D., Jena, K.K., Wicker, T., Chen, M., Han, B., Henry, R., Hsing, Y.C., Kurata, N., de Oliveira, A.C., Panaud, O., Jackson, S.A., Machado, C.A., Sanderson, M.J., Long, M., Ware, D., and Wing, R.A. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet **50,** 285-296.

Steinegger, M., and Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol **35,** 1026-1028.

Sun, J., Li, L., Wang, P., Zhang, S., and Wu, J. (2017). Genome-wide characterization, evolution, and expression analysis of the leucine-rich repeat receptor-like protein kinase (LRR-RLK) gene family in Rosaceae genomes. BMC Genomics **18,** 763.

Sun, R., Wang, S., Ma, D., and Liu, C. (2018). Genome-Wide Analysis of LRR-RLK Gene Family in Four Gossypium Species and Expression Analysis during Cotton Development and Stress Responses. Genes (Basel) **9**.

Sun, X., and Wang, G.L. (2011). Genome-wide identification, characterization and phylogenetic analysis of the rice LRR-kinases. PLoS One **6,** e16079.

Sun, X., Cao, Y., Yang, Z., Xu, C., Li, X., Wang, S., and Zhang, Q. (2004). Xa26, a gene conferring resistance to Xanthomonas oryzae pv. oryzae in rice, encodes an LRR receptor kinase-like protein. Plant J **37,** 517-527.

Sun, Y., Li, L., Macho, A.P., Han, Z., Hu, Z., Zipfel, C., Zhou, J.M., and Chai, J. (2013). Structural basis for flg22-induced activation of the Arabidopsis FLS2-BAK1 immune complex. Science **342,** 624-628.

Szalkowski, A.M., and Anisimova, M. (2013). Graph-based modeling of tandem repeats improves global multiple sequence alignment. Nucleic Acids Res **41,** e162.

Takahashi, N., Takahashi, Y., and Putnam, F.W. (1985). Periodicity of leucine and tandem repetition of a 24-amino acid segment in the primary structure of leucine-rich alpha 2-glycoprotein of human serum. Proc Natl Acad Sci U S A **82,** 1906-1910.

Tamborski, J., and Krasileva, K.V. (2020). Evolution of Plant NLRs: From Natural History to Precise Modifications. Annu Rev Plant Biol **71,** 355-378.

Tao, Y., Zhao, X., Mace, E., Henry, R., and Jordan, D. (2019). Exploring and Exploiting Pan-genomics for Crop Improvement. Mol Plant **12,** 156-169.

Terrapon, N., Gascuel, O., Marechal, E., and Brehelin, L. (2012). Fitting hidden Markov models of protein domains to a target species: application to Plasmodium falciparum. BMC Bioinformatics **13,** 67.

Thibaud-Nissen, F., Ouyang, S., and Buell, C.R. (2009). Identification and characterization of pseudogenes in the rice gene complement. BMC Genomics **10,** 317.

Tommaso, P.D., Floden, E.W., Magis, C., Palumbo, E., and Notredame, C. (2017). [Nextflow, an efficient tool to improve computation numerical stability in genomic analysis]. Biol Aujourdhui **211,** 233-237.

Urbach, J.M., and Ausubel, F.M. (2017). The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. Proc Natl Acad Sci U S A **114,** 1063-1068.

van Baren, M.J., and Brent, M.R. (2006). Iterative gene prediction and pseudogene removal improves genome annotation. Genome Res **16,** 678-685.

Van de Weyer, A.L., Monteiro, F., Furzer, O.J., Nishimura, M.T., Cevik, V., Witek, K., Jones, J.D.G., Dangl, J.L., Weigel, D., and Bemm, F. (2019). A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. Cell **178,** 1260-1272 e1214.

Vanin, E.F. (1985). Processed pseudogenes: characteristics and evolution. Annu Rev Genet **19,** 253-272.

Vaughan, D.A., Lu, B.-R., and Tomooka, N. (2008). Was Asian Rice (Oryza sativa) Domesticated More Than Once? Rice **1,** 16-24.

Vicat, T. (2021). Amélioration, mise en forme et distribution d'un pipeline de transfert d'annotation (rapport de stage).

Vizueta, J., Sanchez-Gracia, A., and Rozas, J. (2020). bitacora: A comprehensive tool for the identification and annotation of gene families in genome assemblies. Mol Ecol Resour **20,** 1445-1452.

Vollset, S.E., Goren, E., Yuan, C.-W., Cao, J., Smith, A.E., Hsiao, T., Bisignano, C., Azhar, G.S., Castro, E., Chalek, J., Dolgert, A.J., Frank, T., Fukutaki, K., Hay, S.I., Lozano, R., Mokdad, A.H., Nandakumar, V., Pierce, M., Pletcher, M., Robalik, T., Steuben, K.M., Wunrow, H.Y., Zlavog, B.S., and Murray, C.J.L. (2020). Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study. The Lancet **396,** 1285-1306.

Wambugu, P.W., Brozynska, M., Furtado, A., Waters, D.L., and Henry, R.J. (2015). Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. Sci Rep **5,** 13957.

Wang, G., Ellendorff, U., Kemp, B., Mansfield, J.W., Forsyth, A., Mitchell, K., Bastas, K., Liu, C.M., Woods-Tor, A., Zipfel, C., de Wit, P.J., Jones, J.D., Tor, M., and Thomma, B.P. (2008). A

genome-wide functional investigation into the roles of receptor-like proteins in Arabidopsis. Plant Physiol **147,** 503-517.

Wang, G.L., Ruan, D.L., Song, W.Y., Sideris, S., Chen, L., Pi, L.Y., Zhang, S., Zhang, Z., Fauquet, C., Gaut, B.S., Whalen, M.C., and Ronald, P.C. (1998). Xa21D encodes a receptor-like molecule with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. Plant Cell **10,** 765-779.

Wang, J., Tan, S., Zhang, L., Li, P., and Tian, D. (2011). Co-variation among major classes of LRR-encoding genes in two pairs of plant species. J Mol Evol **72,** 498-509.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K.C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., Zhao, X., Shen, F., Cui, X., Yu, H., Li, Z., Chen, M., Detras, J., Zhou, Y., Zhang, X., Zhao, Y., Kudrna, D., Wang, C., Li, R., Jia, B., Lu, J., He, X., Dong, Z., Xu, J., Li, Y., Wang, M., Shi, J., Li, J., Zhang, D., Lee, S., Hu, W., Poliakov, A., Dubchak, I., Ulat, V.J., Borja, F.N., Mendoza, J.R., Ali, J., Li, J., Gao, Q., Niu, Y., Yue, Z., Naredo, M.E.B., Talag, J., Wang, X., Li, J., Fang, X., Yin, Y., Glaszmann, J.C., Zhang, J., Li, J., Hamilton, R.S., Wing, R.A., Ruan, J., Zhang, G., Wei, C., Alexandrov, N., McNally, K.L., Li, Z., and Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature **557,** 43-49.

Wang, Y., Zhang, H., Zhong, H., and Xue, Z. (2021). Protein domain identification methods and online resources. Comput Struct Biotechnol J **19,** 1145-1153.

Way, M.J., and Heong, K.L. (1994). The role of biodiversity in the dynamics and management of insect pests of tropical irrigated rice—a review. Bulletin of Entomological Research **84,** 567-587.

Whitham, S., Dinesh-Kumar, S.P., Choi, D., Hehl, R., Corr, C., and Baker, B. (1994). The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor. Cell **78,** 1101-1115.

Wilming, L.G., Boychenko, V., and Harrow, J.L. (2015). Comprehensive comparative homeobox gene annotation in human and mouse. Database (Oxford) **2015**.

Xiao, J., Sekhwal, M.K., Li, P., Ragupathy, R., Cloutier, S., Wang, X., and You, F.M. (2016). Pseudogenes and Their Genome-Wide Prediction in Plants. Int J Mol Sci **17**.

Xie, J., Li, Y., Liu, X., Zhao, Y., Li, B., Ingvarsson, P.K., and Zhang, D. (2019). Evolutionary Origins of Pseudogenes and Their Association with Regulatory Sequences in Plants. Plant Cell **31,** 563-578.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat Rev Genet **13,** 329-342.

Yeh, R.F., Lim, L.P., and Burge, C.B. (2001). Computational inference of homologous gene structures in the human genome. Genome Res **11,** 803-816.

Yue, J.X., Meyers, B.C., Chen, J.Q., Tian, D., and Yang, S. (2012). Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes. New Phytol **193,** 1049-1063.

Zhang, J., Chen, L.L., Xing, F., Kudrna, D.A., Yao, W., Copetti, D., Mu, T., Li, W., Song, J.M., Xie, W., Lee, S., Talag, J., Shao, L., An, Y., Zhang, C.L., Ouyang, Y., Sun, S., Jiao, W.B., Lv, F., Du, B., Luo, M., Maldonado, C.E., Goicoechea, J.L., Xiong, L., Wu, C., Xing, Y., Zhou, D.X., Yu, S., Zhao, Y., Wang, G., Yu, Y., Luo, Y., Zhou, Z.W., Hurtado, B.E., Danowitz, A., Wing, R.A., and Zhang, Q. (2016). Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. Proc Natl Acad Sci U S A **113,** E5163-5171.

Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P.M., and Gerstein, M. (2006). PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics **22,** 1437-1439.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., Xu, Q., Wang, Z.X., Wei, X., Han, B., and Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet **50,** 278-284.

Zheng, Y., Jiao, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P., Banf, M., Dai, X., Martin, G.B., Giovannoni, J.J., Zhao, P.X., Rhee, S.Y., and Fei, Z. (2016). iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. Mol Plant **9,** 1667-1670.

Zhong, Y., and Cheng, Z.M. (2016). A unique RPW8-encoding class of genes that originated in early land plants and evolved through domain fission, fusion, and duplication. Sci Rep **6,** 32923.

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., Parakkal, P., Arbelaez, L.J., Franco, N., Alexandrov, N., Hamilton, N.R.S., Leung, H., Mauleon, R., Lorieux, M., Zuccolo, A., McNally, K., Zhang, J., and Wing, R.A. (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. Sci Data **7,** 113.

Zipfel, C., and Felix, G. (2005). Plants and animals: a different taste for microbes? Curr Opin Plant Biol **8,** 353-360.