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

Céline Gottin, Anne Dievart, Marilyne Summo, Gaëtan Droc, Christophe Périn, Vincent Ranwez, Nathalie Chantret

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 nonfunctionalization 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).

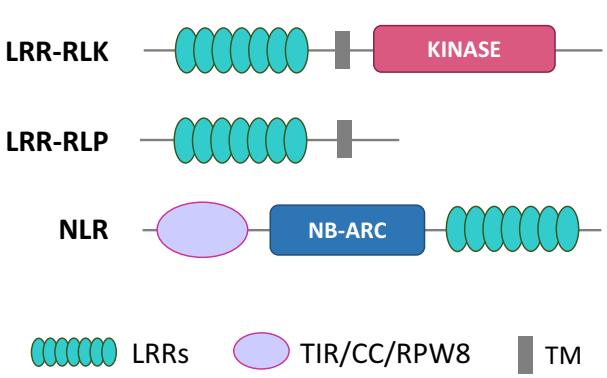

Figure 1. Schematic protein structure of the three LRR-CR subfamilies: LRR-RLK, LRR-RLP and NLR. TM, transmembrane domain; CC, coiled coil domain; TIR, toll-interleukin receptor; RPW8, resistance to powdery mildew 8 domain.

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.

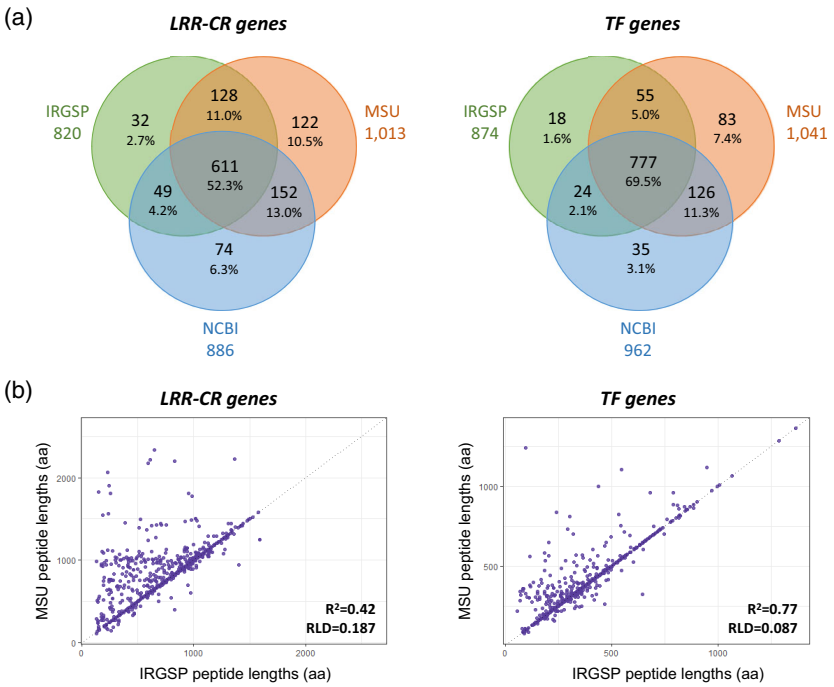

Figure 2. Comparison of publicly available MSU, IRGSP and NCBI annotations for the Nipponbare rice reference genome for two types of genes: LRR-containing receptors (LRR-CRs) and transcription factors (TFs).
(a) Venn diagrams representing the number of overlapping gene models for LRR-CRs and TFs among 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

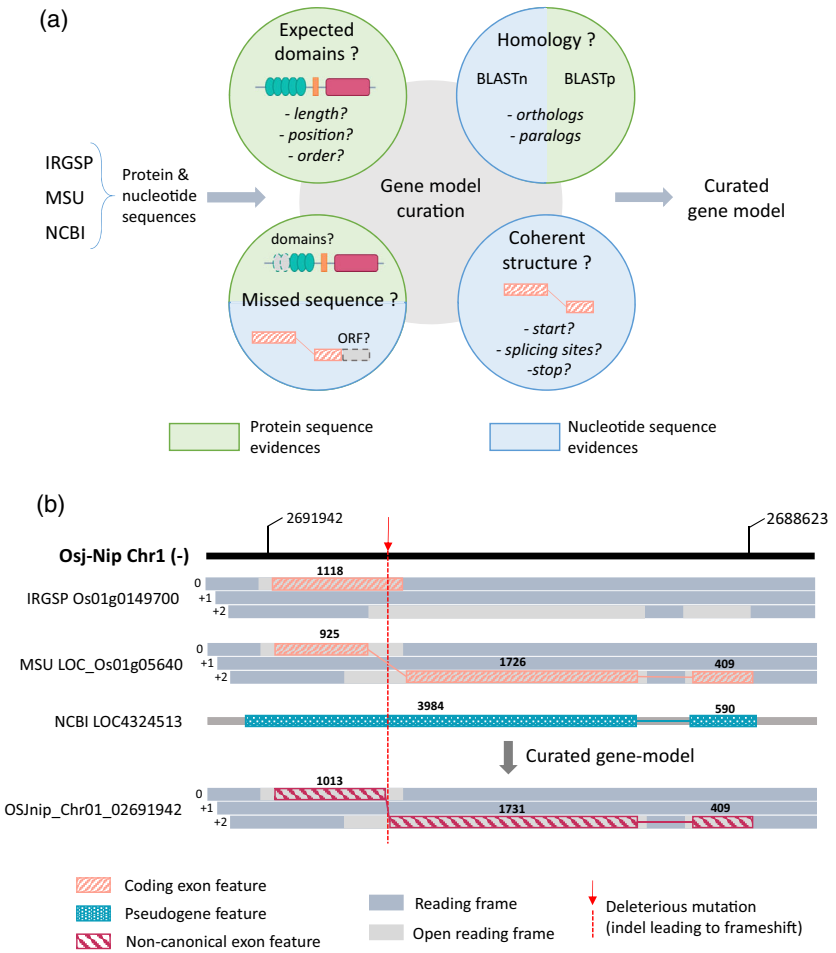

Figure 3. Manual curation of LRR-CR gene model strategy and example of annotation inconsistencies.
(a) Schematic representation of the strategy used to curate Nipponabre LRR-CR gene models. An initial gene model was selected from the three public annotations. This gene model was gradually modified based 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

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,

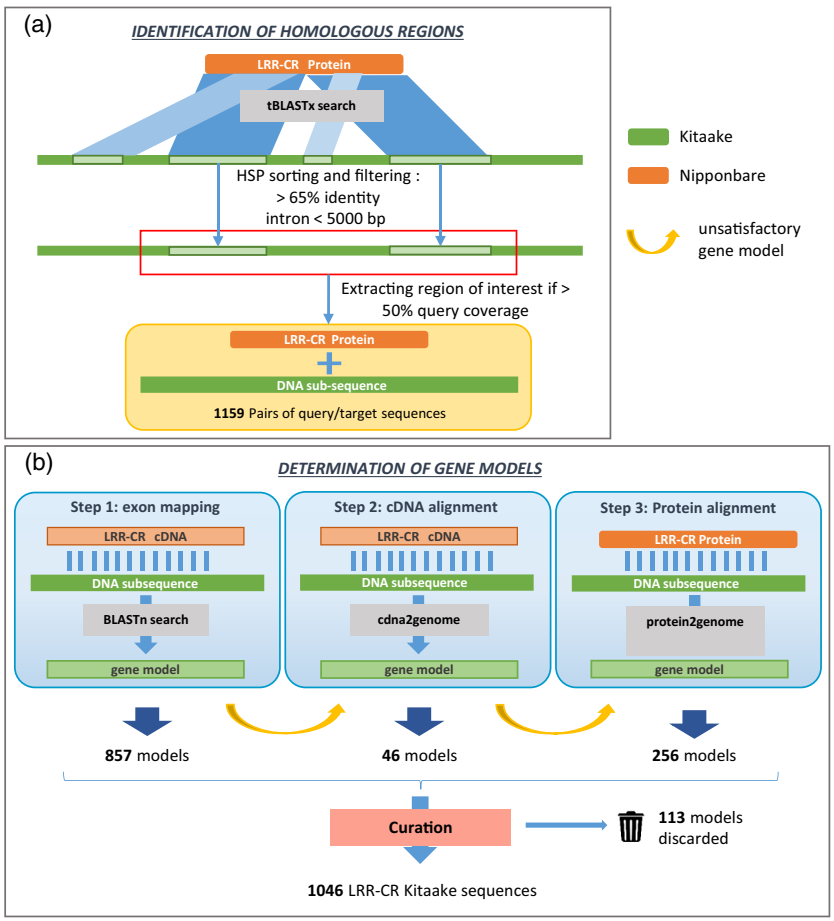

Figure 4. Schematic representation of the annotation transfer strategy between closely related genomes.
(a) Identification of LRR-CR homologous regions. Nipponbare LRR-CR proteins were used to search regions of interest in the Kitaake genome using tBLASTx. BLAST hits with over 65% identity were ranked 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

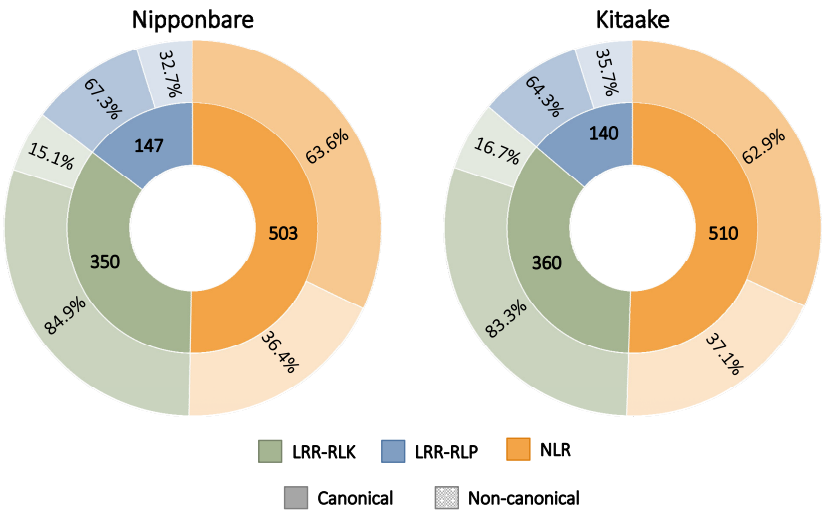

Figure 5. Proportion of canonical and non-canonical loci per gene subfamily in our Nipponbare and Kitaake expert annotations. Percentages were calculated per gene subfamily. The inner circle provides the number of loci per family, with a different color for each. The outer circle shows a lighter/darker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

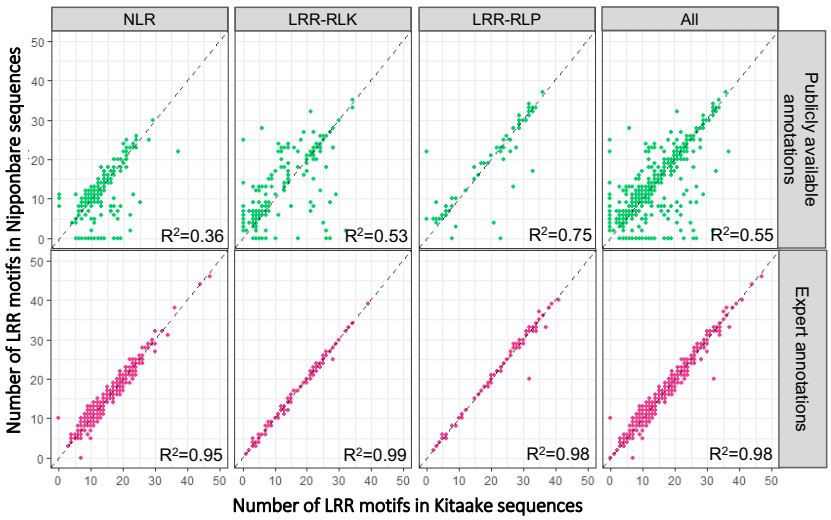

Figure 6. Comparison of LRR motif numbers between Nipponbare and Kitaake LRR-CR alleles, according to the annotations used. In green, comparison between two publicly available annotations of Nipponbare and Kitaake using IRGSP reference data for Nipponbare. In pink, comparison between our Nipponbare and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

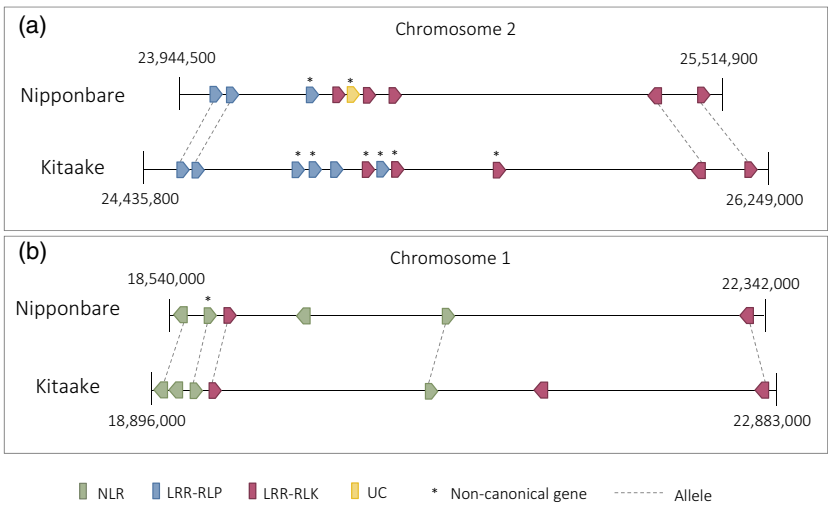

Figure 7. Schematic representation of two large loci on chromosomes 1 and 2 containing cultivar-specific LRR-CR genes.
(a) Representation of an unconserved cluster between Nipponbare and Kitaake on chromosome 2. Five and seven genes in Nipponbare and Kittake, respectively, were cultivar specific. The 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

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.