First '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]AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France.
[2]CIRAD, UMR AGAP, F-34398 Montpellier, France.

*Corresponding author: Nathalie Chantret nathalie.chantret@inrae.fr

Short title: **LRR-CR curated annotation in rice**

Material distribution: The author responsible for the distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Nathalie Chantret (nathalie.chantret@inrae.fr).

Abstract

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) subfamilies: the LRR-Receptor-Like Kinase (RLK), LRR-Receptor-Like Protein (RLP) and Nucleotide-binding LRR Receptor (NLR) subfamilies. Using LRRprofiler, a pipeline we developed to annotate and classify those proteins, we compared three publicly available annotations of the rice Nipponbare reference genome. The extended discrepancies we observed for LRR-CR gene models led us to perform in-depth manual curation of their annotations while paying special attention to nonsense mutations. We then transferred this manually curated annotation to Kitaake, a Nipponbare closely related cultivar, using an optimised strategy. Here we discuss the breakthrough achieved by manual curation when comparing genomes and, in addition to 'functional' and 'structural' annotations, we propose the community to adopt this new approach, which we call 'comprehensive' annotation. The resulting data are crucial for further studies on the natural variability and evolution of LRR-CR in order to promote their use in breeding future resilient varieties.

(198 words)

Introduction

Modern agriculture is at a critical juncture as the world's population continues to grow while there is a call for a shift away from chemical input applications 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) and declining produce quality. Despite awareness on the harmful environmental impacts, massive pesticide use is still a common means to prevent plant diseases today. Studying and understanding plant disease resistance and the underlying evolutionary mechanisms is 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 high and evolving genetic diversity of plant pathogens. The Leucine-Rich Repeat (LRR)-containing receptor (LRR-CR) is at the forefront among these genes. LRR-CRs share the common structural and functional LRR domain. This domain is composed of two to more than 30 repetitions of a ~24 amino acid motif characterised by a conserved skeleton composed mostly of leucine residues (Kajava, 1998; Bella et al., 2008; Kajava, 2012; Matsushima and Miyashita, 2012). LRR-CRs are classified in three main gene subfamilies: LRR Receptor-Like Kinase (LRR-RLK), LRR Receptor-Like Protein (LRR-RLP) and Nucleotide-Binding site LRR (NBS-LRR or NLR) (Sekhwal et al., 2015; Han, 2019) (**Figure 1**). LRR-RLKs and LRR-RLPs are referred to as pattern-recognition receptors (PRRs). These transmembrane receptors have an extracellular LRR domain and an intracellular domain. The latter is a kinase domain for LRR-RLK and a short cytoplasmic tail for LRR-RLP. Note that LRR-RLKs are known intercellular communication players, thus they are also involved in other biological processes such as development. NLRs are intracellular receptors composed of an NB-ARC domain followed by the LRR domain (Sekhwal et al., 2015; Burdett et al., 2019; Tamborski and Krasileva, 2020; Xiong et al., 2020).

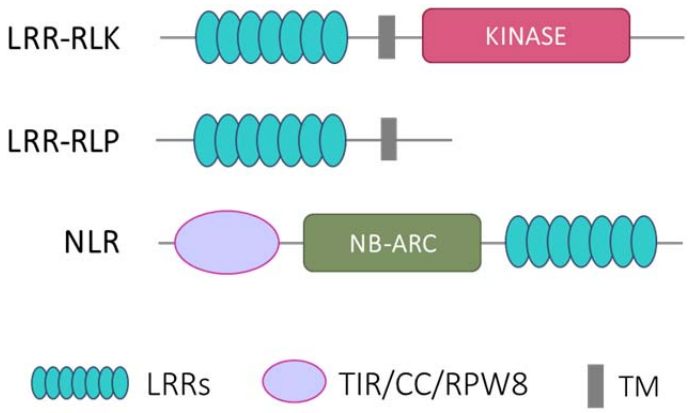


Figure 1. Schematic protein structure of the three LRR-CR subfamilies: LRR-RLK, LRR-RLP and NLR. TM, transmembrane domain; CC, coiled coil domain; TIR, toll-interleukin receptor; RPW8, resistance to powdery mildew 8 domain.

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 time scales. 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 (Wang et al., 2011; Sharma and Pandey, 2015; Fischer et al., 2016; Dufayard et al., 2017). A large proportion of LRR-CR genes are thought to evolve through a so called birth-and-death model (Michelmore and Meyers, 1998; Richter and Ronald, 2000; Nei and Rooney, 2005; McDowell and Simon, 2006). 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 one (neo-functionalization) or, more frequently, undergoing a non-functionalization process by accumulating nonsense mutations (Leister, 2004; Innan and Kondrashov, 2010). 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).

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 commonsense approach is to rely on automatic annotation. Gene and protein sequence annotation are thus crucial and the focus 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, UTRs) (Wilming and Harrow, 2009). The most widely used structural annotation pipelines (e.g. 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 (i) on *ab initio* gene structure determination according to rules learned on pre-existing annotations, and/or (ii) on 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 (Fawal et al., 2014; Bayer et al., 2018) and there are also genome construction issues (Torresen et al., 2019). The difficulty is twofold in the case of LRR-CRs: several similar genes are present in the genome due to gene duplication events, while each gene contains several similar domains due to the repetitive structure of the LRR domain. 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 Nipponbare rice variety, the number of genes per subfamily is variable, e.g. 374 to 498 NLR proteins (Zhou et al., 2004; Li et al., 2010; Li et al., 2016; Shao et al., 2016; Stein et al., 2018), 292 to 332 LRR-RLKs (Hwang et al., 2011; Sun and Wang, 2011; Dufayard et al., 2017) 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 manual curation of gene annotations to limit these uncertainties and achieve high quality comprehensive analyses, as for Arabidopsis and tomato NLR genes (Meyers et al., 2003; Jupe et al., 2013; Van de Weyer et al., 2019) or Nipponbare rice 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

125 (Kawahara et al., 2013), are currently available: one from the Michigan State University
126 Rice Genome Annotation Project (hereafter named MSU) (Yuan et al., 2003), one from
127 the National Center for Biotechnology Information (NCBI) and the current reference one
128 from the Rice Annotation Project of the International Rice Genome Sequencing Project
129 (IRGSP) (Sakai et al., 2013). We first implemented the LRRprofiler pipeline to compare
130 them with regard to the LRR-CR repertoire. This program builds subfamily- and genome-
131 specific LRR Hidden Markov Model (HMM) profiles, detects LRR-CR genes that contain
132 LRR motifs, and accurately locates all LRR motifs within these genes. We ran the
133 LRRprofiler pipeline in parallel on the three rice annotations and found that they greatly
134 differed in terms of the number of LRR-CR genes and their structural annotations. We
135 therefore performed a manual curation of the whole Nipponbare LRR-CR repertoire. We
136 decided not to force gene models to have to contain all expected features in functional
137 genes but rather, when appropriate, to supplement the gene models with sequence
138 fragments undoubtedly derived from LRR-CR encoding genes. In turn, we provided
139 objective information, i.e. whether the gene models were canonical or non-canonical. To
140 be qualified as canonical, a gene model had to fulfil all of the following conditions:
141 presence of a start codon, presence of a terminal stop codon, absence of an in-frame
142 stop codon, absence of frameshifts, and absence of unexpected intron splicing sites.
143 Conversely, any gene violating at least one of these constraints was qualified as non-
144 canonical. Finally, we also propose a strategy to transfer these manually curated LRR-
145 CR gene annotations to Kitaake, the closest related japonica genome whose sequence
146 is available (Jain et al., 2019). We then studied observed variations in gene numbers
147 and LRR motifs between Nipponbare and Kitaake genotypes when using available
148 automatic annotations and our manually curated annotations (hereafter referred to as
149 'comprehensive'). This comparison demonstrated how erroneous conclusions can
150 readily be drawn when relying solely on automatic structural and functional annotations
151 for this complex gene family. The curated comprehensive LRR-CR annotation
152 introduced in this article is available online through a dedicated website

153 (https://rice-genome-hub.southgreen.fr/content/geloc).

154

155

156

157 **Results**

158 **Inconsistencies among three publicly available Nipponbare rice LRR-CR**
159 **annotations**

160 We used LRRprofiler, a newly developed pipeline (see Material and Methods and
161 **Supplemental Methods**), to identify, annotate and classify into gene subfamilies the
162 LRR-CR protein sequences of the three publicly available Nipponbare proteomes (MSU,
163 IRGSP, NCBI). The total number of LRR-CRs identified varied markedly according to the
164 annotation: we identified 1,186 LRR-CRs in the MSU predicted proteome, 981 in that of
165 IRGSP and 1,025 in that of NCBI (**Table 1**). The distribution patterns of these proteins in
166 the different subfamilies also varied according to the annotations. For instance, the
167 number of predicted genes fluctuated less for the RLP subfamily than for the NLR
168 subfamily for which 64% more NLRs were detected in the MSU proteome (405 proteins)
169 compared to the IRGSP proteome (260 proteins).
170

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

	Total	LRR-RLKs	LRR-RLPs	NLRs	Others*
IRGSP	981	237 (24.2%)	160 (16.3%)	260 (26.5%)	324 (33.0%)
MSU	1,186	329 (27.7%)	140 (11.8%)	405 (34.2%)	312 (26.3%)
NCBI	1,025	305 (29.8%)	122 (11.9%)	339 (33.1%)	259 (25.2%)

174 * Others refer here to F-box-LRR and unclassified (UC) sequences

175

176 We assessed whether the identified genes were at the same genomic location or not by
177 measuring the overlap of the three predicted gene sets. We conducted a similar analysis
178 on nine transcription factor (TF) subfamilies, for which we assumed that the annotation
179 process would be easier since they had a more conserved structure and, although
180 having undergone expansion events, were not evolving under a birth-and-death model
181 (Lai et al., 2020). The TF dataset contained between 874 and 1,041 genes, according to
182 the annotations, and this number was similar to that of LRR-CR. The percentage of loci

183 for which a gene model was present in all three annotations was 52.9% for LRR-CR
184 genes and 69.5% for TF (**Figure 2A**), indicating that the three annotations were more
185 congruent for TF genes. Moreover, the percentage of loci in which only one annotation
186 detected a gene was 19.5% for LRR-CR genes, compared to only 12% for TF genes.

187 Even when a gene was predicted by the different annotations, the predicted structure of
188 the gene sometimes varied between predictions. One way to address this issue is to
189 compare the length of predicted proteins for genes positioned at the same locus. Note
190 that this is a conservative approach. Indeed, although a predicted protein length
191 difference between two gene models indicated that the gene models differed, the
192 reverse was not true as identical predicted protein lengths did not guarantee that the
193 gene models were identical. A comparison of predicted protein lengths for all LRR-CR
194 gene pairs located at the same locus but predicted by two different annotations is
195 presented in **Figure 2B** and **Supplemental Figure 1**. Here again, the number of genes
196 with a difference in the predicted protein length highlighted a substantial annotation
197 discrepancy. This difference was greater for LRR-CR genes than for TF genes. As an
198 example, when IRGSP and MSU were compared, the average difference between the
199 predicted product sizes was 18.7% for LRR-CR loci (with a R^2 of only 0.42) whereas this
200 difference was only 8.6% for TF loci (with a much higher R^2 of 0.77) (**Figure 2B**). These
201 results highlight the extent to which annotations generally differ, but more particularly for
202 LRR-CR gene subfamilies. These comparisons also showed that LRR-CR genes
203 predicted by IRGSP were generally shorter than those predicted by MSU or NCBI at the
204 same locus (**Figure 2B** and **Supplemental Figure 1**).

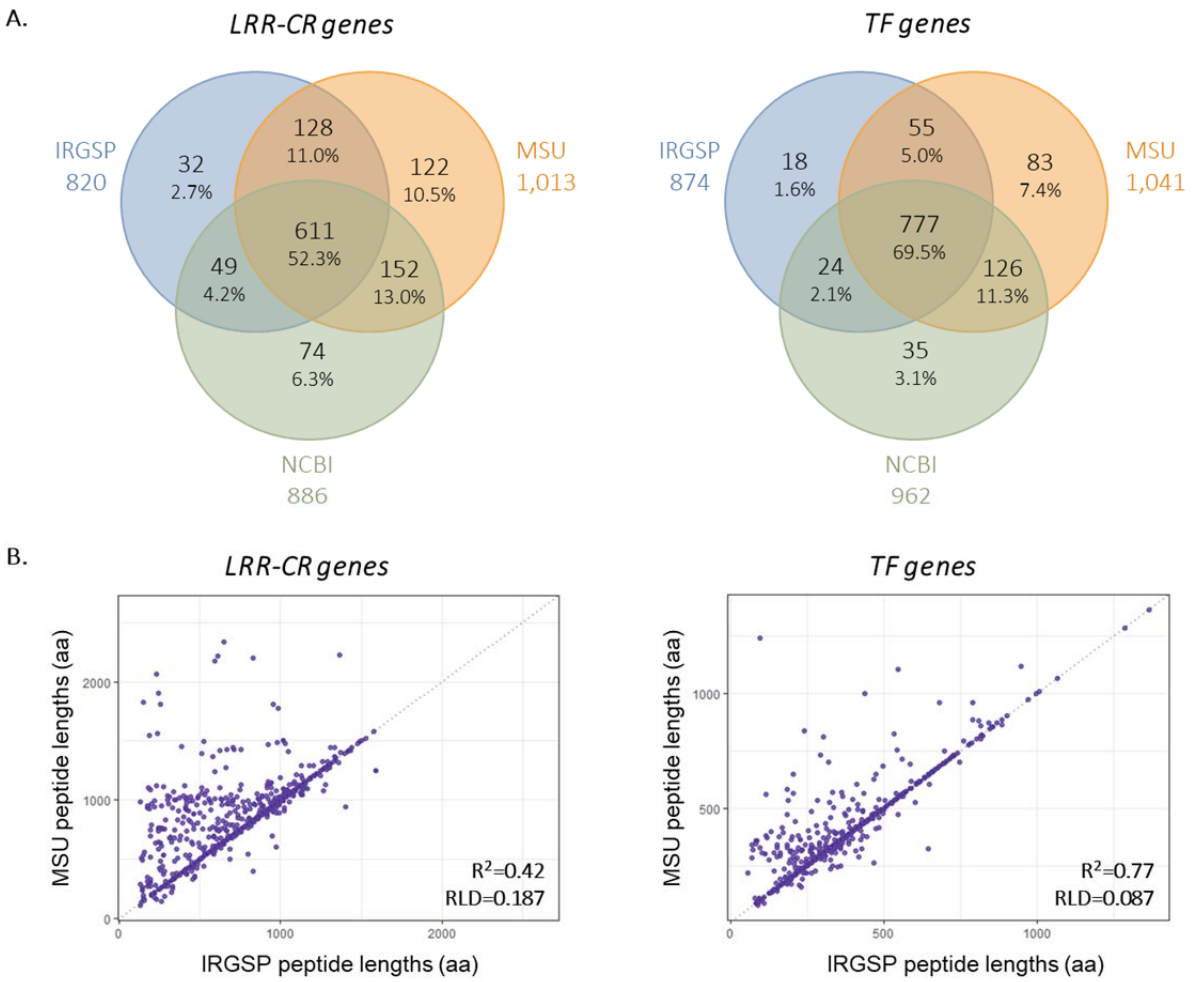


Figure 2. Comparison of MSU, IRGSP and NCBI publicly available annotations for the Nipponbare rice reference genome for two types of genes: LRR-containing receptors (LRR-CR) and transcription factors (TF).

(A) Venn diagrams representing the number of overlapping gene models for LRR-CR and TF between 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).

(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) are given at the bottom right for each gene family. See **Supplemental Figure 1** for all pairwise comparisons between IRGSP, MSU and NCBI.

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

Here we provide a brief description of the procedure we followed to manually curate LRR-CR annotations (**Figure 3A**). Note above all that, for traceability sake, the

219 procedure retained one of the three proposed gene models as much as possible. For a
220 given locus, we first selected one of the gene models among the available annotations
221 based on the completeness of the predicted protein. We then applied our expertise to
222 the selected gene model by combining protein and nucleic data. At the protein level, we
223 checked that all of the expected domains for each subfamily were present (e.g. LRRs
224 and TM for LRR-RLPs and LRR-RLKs, or NB-ARC for NLRs), in the right order, with the
225 expected length and inter-domain intervals. Protein domain information was particularly
226 useful for detecting potential gene fusion and fission. At the nucleic level, we examined:
227 (i) if the gene models had the expected structure (intron/exon), (ii) if nearby open
228 reading frames (ORF) belonging to LRR-CR encoding sequences were present, and (iii)
229 if the gene models included suspicious introns, such as short introns thereby enabling
230 the gene to sidestep stop codons or frameshifts, especially when they were never found
231 in homologs (**Figure 3B**). Any structural annotation containing an in-frame stop codon or
232 a frameshift (i.e. any gap in coding sequence that was not an intron but that changed the
233 translation phase), lacking a start codon or a terminal stop codon, or presenting an
234 unexpected splicing site (different from the GT-AG and GC-AG donor/acceptor canonical
235 splicing sites) was called 'non-canonical'. This careful inspection was facilitated by
236 viewing the sequence annotations with the Artemis editor (Carver et al., 2012).

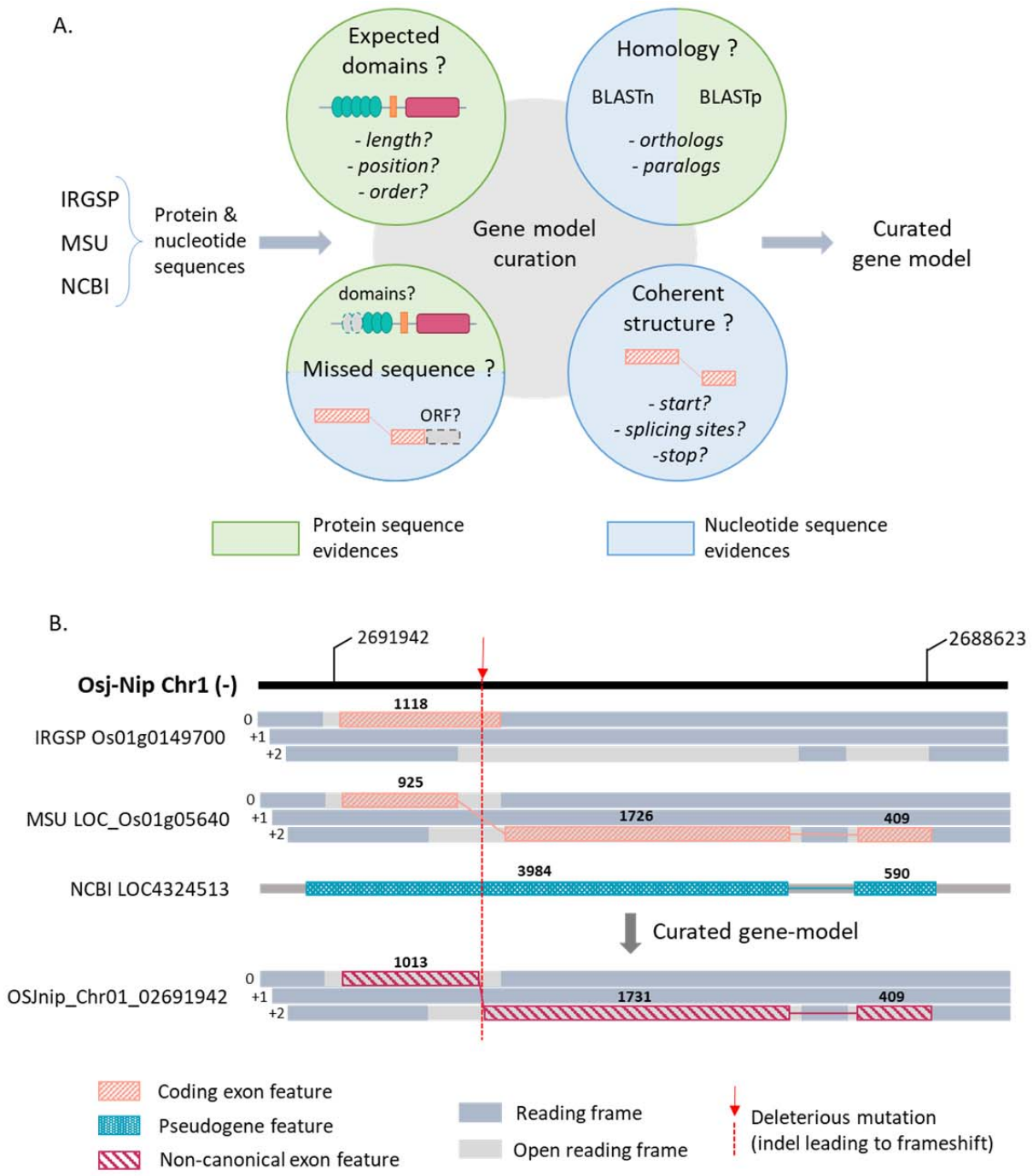


Figure 3. Manual curation of LRR-CR gene models strategy and example of annotation inconsistencies. **(A)** Schematic representation of the strategy used to curate Nipponabre LRR-CR gene models. An initial gene model was selected from the three public annotations. This gene model was gradually modified based on protein and nucleotide sequence evidence. The curated model was then classified as canonical or non-canonical.

243 **(B)** Schematic representation of an example of inconsistency between gene models from publicly
244 available annotations and how the curation was done. The gene is an LRR-RLK located on chromosome 1
245 of the Nipponbare genome. Numbers above boxes give the feature length. In this example, an indel
246 mutation caused a frameshift in the first exon of the gene. The IRGSP annotation retrieved the first part of
247 the coding sequence stopping at the first stop codon on frame 0. The MSU annotation retrieved a longer
248 coding sequence, but sidestep the indel mutation by introducing a 'dubious' intron in order to reach the
249 ORF on the +2 frame. This 'dubious' intron was abnormally short and contained a sequence highly
250 homologous to the coding sequence in other paralogous gene copies. The NCBI annotation gave a
251 pseudogene feature, i.e. a feature from which a protein sequence could not be deduced: the cDNA
252 sequence is available but would not allow protein translation as it would be in the wrong reading frame
253 after the mutation. The curation took advantage of the three annotations. It retried a cDNA sequence that
254 overlapped the complete former coding sequence in the two successive correct reading frames via
255 identification of the indel mutation. The identification of the indel mutation was clearcut since the gene was
256 tagged as 'non-canonical' with the presence of a frameshift, but it allowed a complete protein sequence to
257 be deduced and used for sequence comparison and alignment.

258

259 In a last step, we looked for LRR-containing sequences that would have been missed by
260 the three publicly available annotations. The genome was thus split into 1 kb segments
261 with overlapping 100 bp borders, translated into amino acid sequences in the six reading
262 frames, and domains (LRR, kinase, NB-ARC, etc.) were searched with hmmsearch. The
263 overall results were concatenated and filtered for redundancies. Domains falling inside
264 annotated exons or isolated from other LRR motifs (within a 5 kb range) were discarded.
265 Regions of interest (with at least three LRR motifs in tandem) were compared to other
266 plant genomes using BLAST to screen for the potential presence of a gene model in the
267 considered region.

268 The final set of manually validated LRR-CR loci on the Nipponbare genome consisted of
269 1,052 genes (351 LRR-RLKs, 147 LRR-RLPs, 495 NLRs and 59 UCs) (**Table 2** and
270 **Supplemental Data Set 1**). Among these 1,052 genes, eight (1 RLK, 3 RLP and 4 UC)
271 were located at loci for which none of the three publicly available annotations detected a
272 gene. The newly identified LRR-RLK was a canonical full-length sequence on the
273 forward strand of chromosome 2 (from 6,831,702 to 6,834,761). Note that this sequence
274 is actually present in GenBank under the accession number EAZ22278.1, but is located

275 on the reverse strand in a non-coding region of the Os02g0222500 gene. The seven
276 other are non-canonical truncated genes. In addition, for six of these 1,052 validated
277 genes, the LRRprofiler pipeline did not detect any further LRR motifs in the predicted
278 protein. LRR motifs were initially detected for these six genes, but at the threshold limit
279 when using HMM profiles built on the basis of the initial dataset (for details see
280 LRRprofiler pipeline section in the Material and Methods). When using the slightly
281 different HMM profiles obtained with the final dataset, the same LRR motifs were no
282 longer detected as they did not surpass the threshold. However, a careful manual
283 inspection showed that the LRR domain was present but contained divergent LRR
284 motifs, thereby complicating automatic detection. Consequently, these genes were kept
285 and classified according to the presence of the other domains (kinase or NB-ARC).
286 These eleven genes included two LRR-RLKs and nine NLRs.

287 Among these 1,052 LRR-CR genes, 327 (195 NLR, 57 LRR-RLK, 55 LRR-RLP and 20
288 UC) were manually modified because none of the three publicly available annotations
289 had a satisfactory gene model based on the previously defined criteria (**Figure 3**,
290 **Supplemental Data Set 1**). The overall proportion of modified loci was 31.3%
291 (327/1,052) and markedly varied according to the considered gene subfamily. Only
292 16.3% of LRR-RLK loci were modified, while 38.2% of the LRR-RLP loci and 39.4% of
293 the NLR loci were modified (**Table 2**).

294 Among these 1,052 LRR-CR genes, 305 (29.0%) were non-canonical (**Table 2**). More
295 precisely, 273 genes were both non-canonical and modified, representing 83.5% of the
296 total modified loci (273 over 327) and 89.5% of the non-canonical loci (273 over 305)
297 (**Supplemental Table 1**). The remaining 32 non-canonical genes were either unreported
298 by any of the annotations (7), or reported by the NCBI as pseudogene or gene models
299 having putative errors in the genomic sequence due to differences with respect to the
300 RNAseq data (25).

301 The different gene subfamilies did not contain the same proportion of non-canonical
302 gene models (**Table 2**). Very similar to what was observed regarding the proportion of
303 modified gene models according to the gene subfamily, non-canonical gene models

concerned only 15.1% of the LRR-RLK, compared to 32.7% and 36.8% of the LRR-RLP and NLR, respectively.

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]	1,052 (8)	351 (1)	147 (3)	495 (0)	59 (4)
Modified loci (%[b])	327 (31.3%)	57 (16.3%)	55 (38.2%)	195 (39.4%)	20 (36.4%)
Non-canonical loci (%[b])	305 (29.0%)	53 (15.1%)	48 (32.7%)	182 (36.8%)	22 (37.3%)

[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.

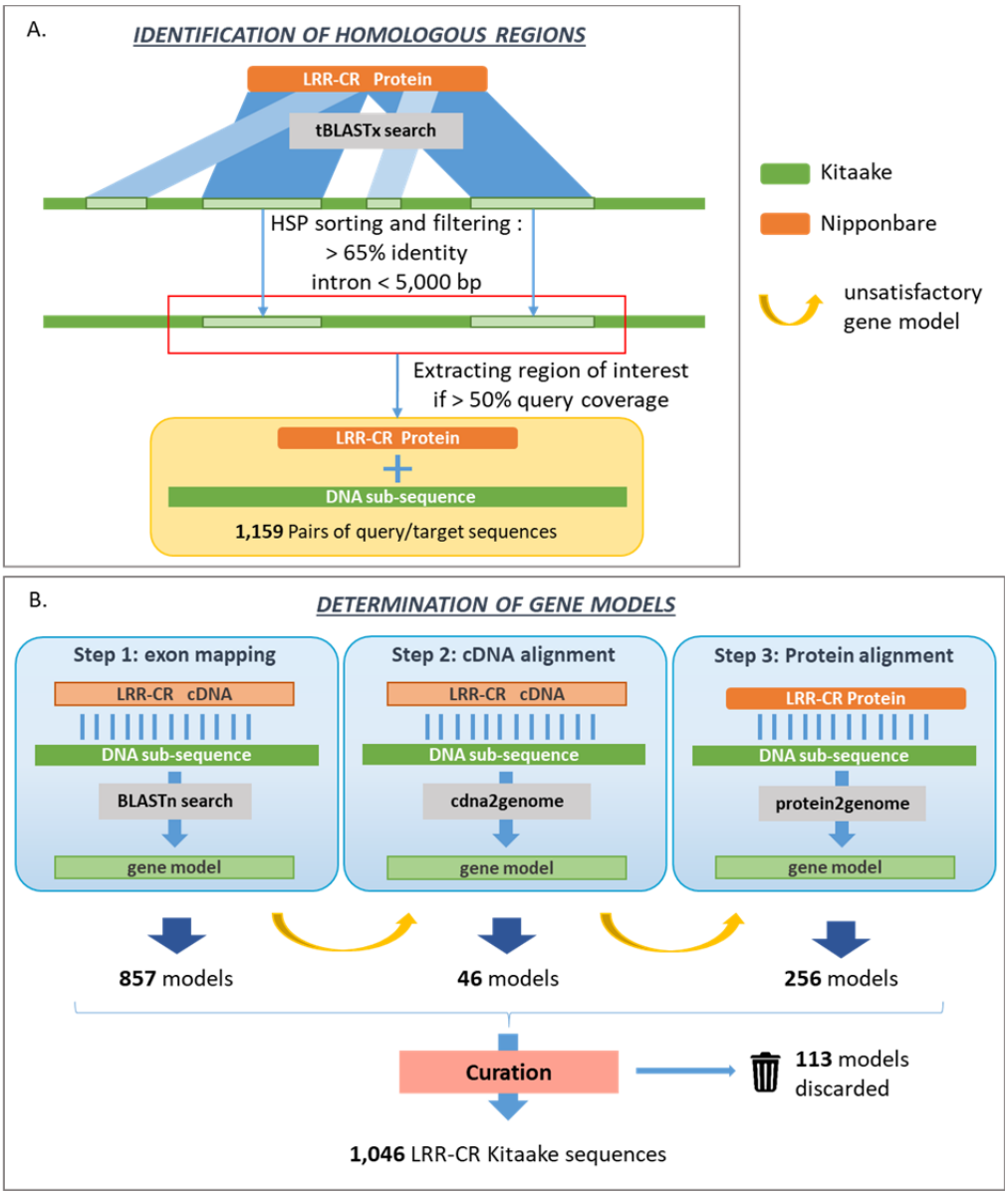
One way to assess the relevance of our expert LRR-CR annotations is to compare the number of functional domains (TM, NB-ARC, kinase, and LRRs) found in LRR-CR proteins derived from the reference annotations to the number of functional domains found in the proteins derived from our expert annotation (**Supplemental Figure 2**). These comparisons revealed that quite a few more LRR-CR domains were found in our manual annotation as compared to the publicly available annotations. For example, when compared to the IRGSP annotation, our expert annotation highlighted 29% more TM, 42% more NB-ARC, 33% more kinase and 20% more LRR motifs.

Comparison of LRR-CR loci between the Nipponbare and Kitaake rice cultivars

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 have 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

330 account three levels of annotation transfer, depending mostly on the level of sequence
331 identity of each considered region with the LRR-CR gene that identified it. At each locus,
332 our strategy strives to retrieve the most probable gene model with the idea that, if
333 possible, it should be canonical. At the end of the process, LRR-CR gene models that
334 are found to be non-canonical or having a dubious protein structure in Kitaake are
335 manually checked and corrected if needed. At this step, the transfer allowed us to
336 identify 1046 genes.



337

338

339 **Figure 4:** Schematic representation of the annotation transfer strategy between closely related genomes.
340 **(A)** Identification of LRR-CR homologous regions. Nipponbare LRR-CR proteins were used to search
341 regions of interest in the Kitaake genome using tBLASTx. BLAST hits with over 65% identity were ranked
342 in the LRR-CR query protein sequence order and used to define the region boundaries. If the filtered
343 BLAST hits within a Kitaake region covered more than 50% of the query LRR-CR sequence, then the
344 Kitaake sequence of this region was extracted and linked to the Nipponbare LRR-CR query protein.
345 **(B)** Determination of gene models. The process strives to give a gene model for each region of interest
346 identified in the Kitaake genome. The annotation is attempted in three consecutive steps. If the model
347 from one step is unsatisfactory, i.e. gives an alignment of poor quality with the Nipponbare query protein,
348 the process goes to the next step for this region. At the end of the third step, gene models that remained
349 unsatisfactory were manually checked. This process allowed us to annotate 1,046 genes in the Kitaake
350 genome.
351

352 As carried out for Nipponbare, the Kitaake genome was finally scanned with LRR HMM
353 profiles for new LRR-CR identifications. This procedure allowed us to annotate 18
354 additional genes, thereby leading to a total of 1,064 LRR-CR genes in the Kitaake
355 genome.

356 The LRRprofiler pipeline was used on the 1,064 predicted Kitaake proteins and allowed
357 us to detect LRR in 1,054 of them, 998 were further classified in a LRR-CR subfamily
358 and 56 remained in the UC group. Automatic detection of LRR failed for 10 genes. As
359 carried out for Nipponbare, at this step, manual validation of the protein annotations
360 confirmed the presence of an LRR domain of the expected size and located at the
361 expected positions. These 10 genes were therefore kept in the final dataset. The gene
362 subfamilies for these 10 loci were determined based on other functional domains and a
363 homology search against other LRR-CR protein sequences. Finally, the LRR-CR gene
364 set from Kitaake was composed of 360 LRR-RLKs, 140 LRR-RLPs, 508 NLRs and 56
365 UCs (**Figure 5** and **Supplemental Data Set 2**).

366 We then tagged all of these Kitaake LRR-CR gene models as either canonical or non-
367 canonical. We obtained 744 (69.9%) canonical genes and 320 (30.1%) non-canonical
368 genes. The proportions of canonical and non-canonical genes per subfamily for Kitaake
369 were very similar to those obtained for Nipponbare (**Figure 5**).

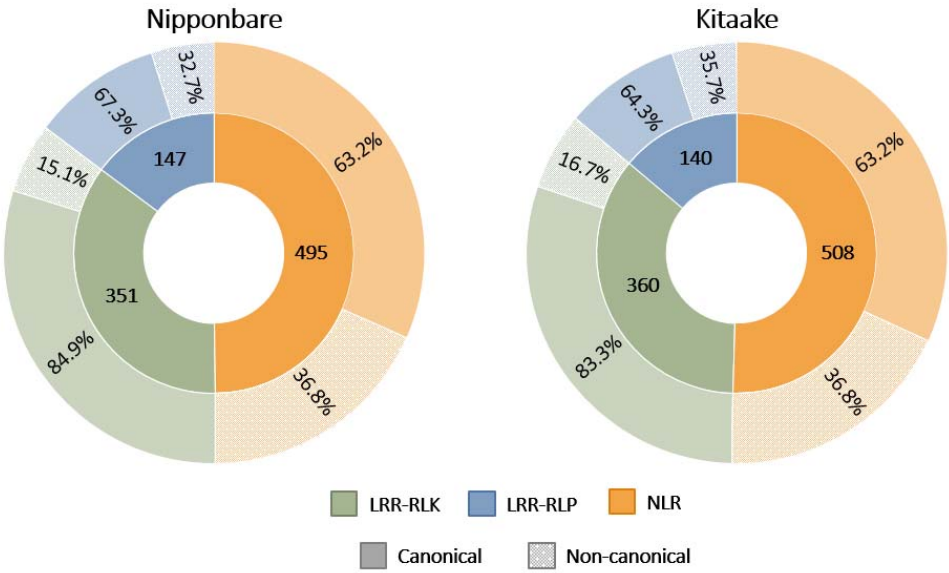


Figure 5. Proportion of canonical and non-canonical loci per gene subfamily in our Nipponbare and Kitaake expert annotations. Percentages were calculated per gene subfamily. The inner circle provides the number of loci per family, with a different colour for each. The outer circle shows a lighter (respectively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

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 alleles between Nipponbare and Kitaake

Nipponbare and Kitaake are two varieties of the same subspecies, i.e. *Oryza 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.

By using the Synmap program (Lyons and Freeling, 2008), we identified 999 allelic pairs (representing 90.0% of the total number of loci) between Nipponbare and Kitaake

388 (**Supplemental Data Set 3**). In addition, we noticed that for three NLR gene pairs
389 located close to each other on chromosome 9 in the Nipponbare genome, three
390 consecutive genes of the Kitaake chromosome 3 were found with 100% identity with
391 regard to their predicted coding sequence. The intergenic sequences of these two
392 regions also had a high level of identity (99% over 40.5 kb), suggesting that these three
393 genes are located in a translocated region of the genome. Among these 1,002 pairs
394 (999 allelic plus 3 translocated), 688 (68.7%) were pairs of canonical gene models, 267
395 (26.6%) were pairs of non-canonical gene models, and 47 (4.7%) were pairs of genes
396 found to be canonical only in one of the two cultivars. Interestingly, 83.3% of the LRR-
397 RLK pairs were canonical in both cultivars, compared to only 63.9% of the LRR-RLP
398 pairs and 60.5% of the NLR pairs (**Table 3**).

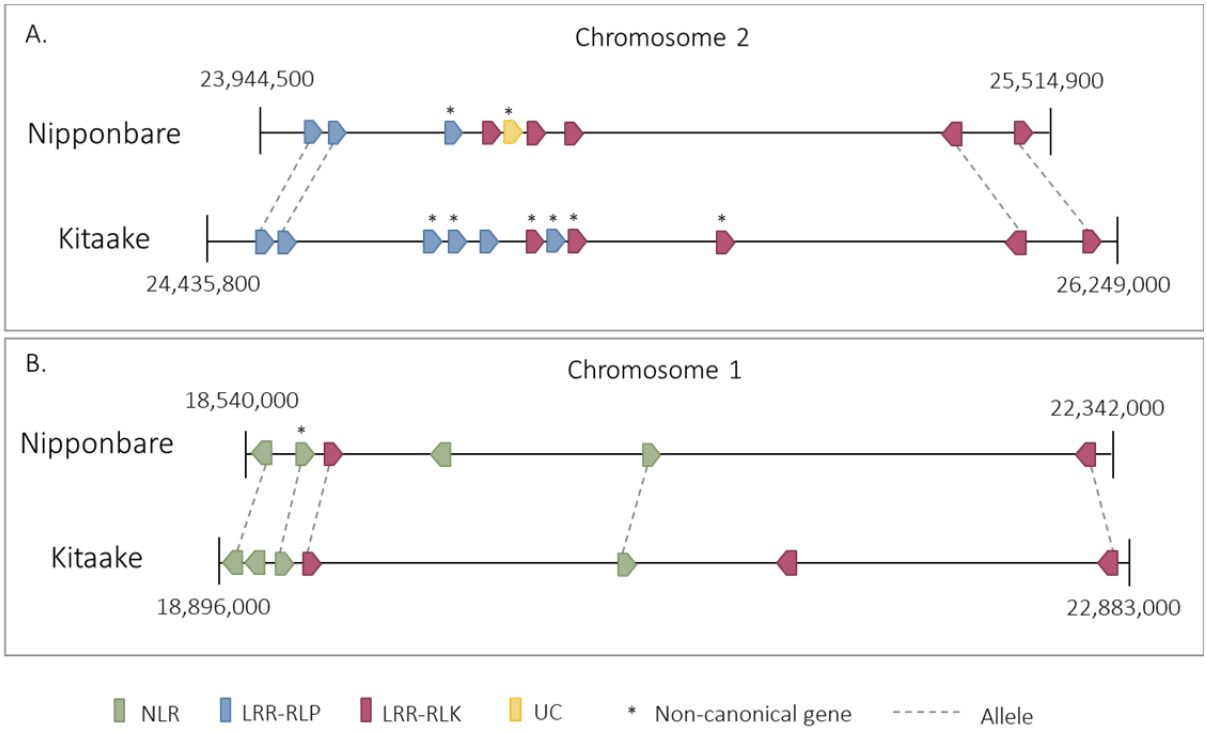
399 The remaining 111 genes were genotype-specific: 49 were present only in Nipponbare
400 and 62 only in Kitaake, of which 31 (63.3%) and 37 (59.7%), respectively, were
401 canonical. These genes were not evenly distributed on the genomes. Some of them
402 were located in entirely different regions in Nipponbare and Kitaake, in particular in
403 chromosomes 2, 11 and 12 where entire clusters were genotype-specific (**Figure 6A**).
404 Other genotype-specific genes were dispersed in regions containing conserved allelic
405 pairs in either Nipponbare or Kitaake (**Figure 6B**). Note that chromosome 11, which also
406 contained about a fifth of all LRR-CR genes, hosted 44 of the 111 (40.4%) cultivar-
407 specific loci.

408 **Table 3.** Number of allelic pairs between Nipponbare and Kitaake cultivars according to categories and
409 subfamilies.

Allele categories	Total	LRR-RLK	LRR-RLP	NLR	UC
C/C*	688 (68.7%)	285 (83.3%)	85 (63.9%)	288 (60.5%)	30 (58.8%)
NC/NC*	267 (26.6%)	46 (13.5%)	41 (30.8%)	162 (34.0%)	118 (35.5%)
C/NC*	28 (2.8%)	7 (2.0%)	4 (3.0%)	14 (2.9%)	3 (5.9%)
NC/C*	19 (1.9%)	4 (1.2%)	3 (2.3%)	12 (2.5%)	0 (0%)
Total	1002	342	133	476	51

410 * The four categories partitioned the loci according to whether they were canonical (C) or not (NC) in
411 Nipponbare/Kitaake. Numbers in parenthesis are the percentage per subfamily.

412



413

414 **Figure 6**. Schematic representation of two large loci on chromosomes 1 and 2 containing cultivar-specific
415 LRR-CR genes.

416 **(A)** Representation of an unconserved cluster between Nipponbare and Kitaake on chromosome 2. Five
417 and seven genes in Nipponbare and Kittake, respectively, were cultivar specific. The unconserved region
418 was framed by four conserved genes, i.e. two LRR-RLPs and two LRR-RLKs.

419 **(B)** Representation of a conserved region between Nipponbare and Kitaake on chromosome 1 hosting
420 cultivar-specific genes. The Nipponbare region hosted a cultivar-specific NLR while the Kitaake
421 corresponding region hosted two cultivar-specific genes, i.e. an NLR and an LRR-RLK.

422 We checked one specific feature of the sequenced Kitaake genome (KitaakeX), i.e. the
423 fact that it contains two *XA21* transgenes (Song et al., 1995; Jain et al., 2019). We
424 identified these two transgenes at positions 28,161,378 and 28,165,947 on chromosome
425 6 with no corresponding allele in Nipponbare, in accordance with published data.

426 For each of the 999 pairs of LRR-CR alleles (plus the three translocated pairs), the
427 fraction of exact matches along the cDNA pairwise alignment (i.e. their percentage of
428 identity) was computed. This cDNA identity was about 99.0% on average. The highest
429 identity rate (99.6%) was obtained for alleles belonging to the LRR-RLK subfamily,
430 followed by the NLR (98.7%) and LRR-RLP (98.0%). Non-canonical conserved gene
431 pairs (NC/NC category in **Table 3**) had, on average, a lower identity level (98.7%) than
432 the conserved canonical gene pairs (99.4%). The lowest level of sequence identity
433 (93.4%) was noted between gene pairs with one cultivar having a canonical form and
434 the other a non-canonical form (categories C/NC and NC/C in **Table 3**). Only 17 pairs of
435 alleles shared less than 80% cDNA identity due to both deletions (up to 1.7 kb) and high
436 sequence divergence (**Supplemental Table 2**).

437 **Impact of reannotations on LRR motif numbers in alleles**

438 To obtain a precise annotation of LRR motifs in each protein, we used the LRRprofiler
439 pipeline on proteomes predicted by the publicly available annotations on one hand, and
440 by our manually curated annotations on the other, for both cultivars. A total of 999 pairs
441 of alleles between Nipponbare and Kitaake were identified from our manual annotation
442 (plus three translocated pairs, see above). The same procedure was applied to retrieve
443 pairs of alleles of publicly available annotations. Then the number of LRR motifs
444 predicted in sequences was compared between Nipponbare and Kitaake alleles.

445 We observed a mean difference in LRR number per protein of 3.58 when comparing the
446 publicly available annotations (IRGSP for Nipponbare and the only one that exists for
447 Kitaake) (**Figure 7**, **Supplemental Figure 3**). This difference fell to 0.6 when our re-
448 annotated data were compared. Using our curated annotations hence led to LRR
449 number predictions that were much more consistent between Nipponbare and Kitaake
450 alleles and this trend was observed for all LRR-CR gene subfamilies. Moreover, the
451 mean difference in LRR number still varied between LRR-CR gene subfamilies, with
452 greater conservation of LRR motif numbers between LRR-RLK and LRR-RLP alleles
453 than between NLR alleles.

454

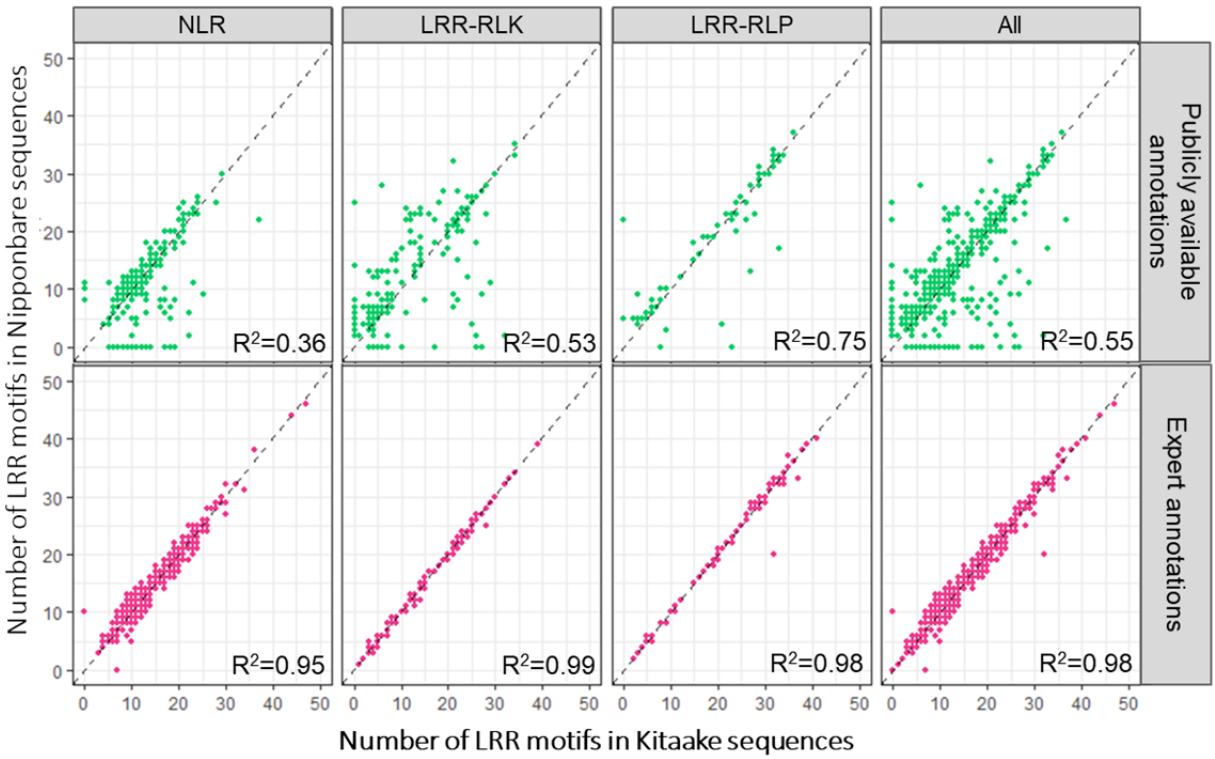


Figure 7. Comparison of LRR motif numbers between Nipponbare and Kitaake LRR-CR alleles, according to the annotations used. At the top, in green, comparison between two publicly available annotations of Nipponbare and Kitaake using IRGSP reference data for Nipponbare; at the bottom, in pink, comparison between our Nipponbare and Kitaake expert annotations.

462 **Discussion**

463 In recent years, in the wake of the gigantic amount of genome sequenced data
464 available, it was exciting to undertake evolutionary studies of gene families. We have
465 been part of this collective enthusiasm, and like many others, have based our research
466 conclusions on perfectible versions of automatic structural and functional gene
467 annotations (Fischer et al., 2016; Dufayard et al., 2017). Although previous genome-
468 wide phylogenetic approaches on LRR-CR gene families enhanced our knowledge on
469 their evolution, they almost never included a data curation step. Indeed, manual re-
470 annotation of gene families is a laborious time-consuming task, especially when dealing
471 with large complex gene families such as LRR-CR, and even more so when dealing with
472 many plant genomes. Despite the fact that automatic annotation tools are continuously
473 improving and remain indispensable at the genome level, human expertise is clearly still
474 needed to achieve an annotation accuracy level suitable for finer and deeper analyses.
475 Today we have finally undertaken this re-annotation work anyway because we are
476 convinced that these curated data are required to produce new reliable results on the
477 evolution of these gene families, especially in the current pangenomic era. Here we
478 describe a new so-called 'comprehensive' annotation strategy. We hope that this new
479 annotation process will get its place in addition to structural and functional annotations in
480 use thus far.

481 **Automatic annotations give inconsistent gene models on complex multigenic**
482 **families**

483 Our comparison of the three publicly available annotations for the Nipponbare rice
484 reference genome showed major discrepancies regarding the total number of LRR-CR
485 genes, the number of LRR-CR assigned to each subfamily, and the gene models (**Table**
486 **1**, **Figure 2** and **Supplemental Figure 1**). These differences were greater for LRR-CR
487 genes compared to other genes such as TF. Automatic annotation pipelines appeared to
488 be suitable overall for many gene families, but they led to a large proportion of
489 inconsistent gene models when annotating complex multigenic families like LRR-CR.
490 The annotation of fast-evolving multigene families is especially challenging for automatic
491 approaches because a high duplication rate is often accompanied by a loss of function

492 process (pseudogeneization) for many copies due, for instance, to mutations like
493 nucleotide substitutions, which may introduce premature stop codons or indels, in turn
494 generating frameshifts. This can lead to the presence of several gene copies sharing
495 high sequence similarity even though some of them may contain nonsense mutations,
496 frameshifts or be truncated. The annotation of gene copies harbouring nonsense
497 mutations is problematic compared to the initial unaltered copy. Some pipelines will be
498 able to detect the entire coding phase but will introduce false introns to sidestep stop
499 codons or frameshifts in order to retrieve a putatively translatable CDS (**Figure 3**). We
500 noticed that the MSU automatic annotation tended to sidestep nonsense and frameshift
501 mutations by introducing short introns. Such errors were observed previously by Meyers
502 when re-annotating the *A. thaliana* NLR gene family (Meyers et al., 2003). Two
503 arguments strengthen the assertion that such introns are false: (i) such introns are never
504 found in more than one copy whereas the intron positions are known to be well
505 preserved between closely related copies, and (ii) sequence comparisons performed
506 against recent paralogues (or orthologues from close relative species) have shown that
507 the sequences of these wrongly annotated introns are always clearly homologous to
508 coding sequences in other gene copies. Among the intron gain mechanisms,
509 intronization (i.e. the process by which an exonic sequence is changed into an intron by
510 mutation accumulation) is a complex process that is not yet very well documented or
511 understood (Yenerall and Zhou, 2012; Roy, 2016). If it really occurs in genomes, it
512 implies that a sufficiently long enough period of time must have passed for these
513 mutations to occur and generate novel splicing sites. It is thus very unlikely that so many
514 new introns arose in such a short period of time, as revealed by the low level of
515 divergence between these genes and their paralogue and/or orthologue counterparts.
516 Other annotation pipelines, such as that of the IRGSP consortium, are more
517 conservative in the sense that they give gene models with a more biologically
518 meaningful expected structure, e.g. truncated proteins, in accordance with the presence
519 of the first premature stop codons (either in-frame or caused by a frameshift, **Figure 3**).
520 This conservative choice could likely explain why we found more sequences classified
521 as LRR-RLP from the IRGSP annotation than from the two others (**Table 1**). Indeed, any
522 LRR-RLK with a premature stop codon somewhere before the kinase domain would be

523 considered as LRR-RLP (**Figure 1**, **Figure 3**). The annotation inconsistencies that we
524 pinpointed here were observed for LRR-CR genes and did not question the overall
525 quality of the three available rice genome annotations. They highlighted the limits of
526 automatic annotation pipelines to annotate such complex multigene families and the
527 consequences that given pipeline decision rules may have when drawing evolutionary
528 conclusions.

529 The comparisons of the different gene models proposed by the three Nipponbare
530 annotations led us to undertake a manual curation of the LRR-CR gene family. We are
531 not the first to get involved in this painstaking but necessary work. Several high-quality
532 studies have been based on re-annotated data, particularly in *A.thaliana* (Meyers et al.,
533 2003; Van de Weyer et al., 2019). Expert annotations could also of course contain
534 errors, but expert curation limits their number. Opting exclusively for automated
535 annotation should be avoided or otherwise operators should be aware that these
536 annotations may contain errors induced by gene family specificities. These biases must
537 be known and understood to avoid drawing misleading conclusions.

538 **The expert Nipponbare rice genome contains more than 1,000 LRR-CR loci of**
539 **which 30% have a non-canonical gene model**

540 We curated LRR-CR loci in the reference Nipponbare rice genome by first comparing
541 the three publicly available annotations at each locus: IRGSP, MSU and NCBI. Our aim
542 was to retrieve LRR-CR genes in their entirety, while taking the coding sequences as
543 they probably were before mutation accumulation into account. We obtained evidence
544 that the sequence portions we included in our new gene models were not random
545 genomic sequences but instead parts of the original gene CDS, as shown by the
546 recovery of protein domains belonging to LRR-CR genes (kinase, NB-ARC, TM.
547 **Supplemental Figure 2**). When a gene had a nonsense mutation (in-frame stop codon
548 or frameshift), an unexpected splicing site, or no terminal stop or start codon, we tagged
549 it as non-canonical. This canonical vs. non-canonical classification was based solely on
550 features observed in gene models and did not imply any judgements on gene
551 functionality. Genes tagged as non-canonical spanned a wide variety of cases, some of
552 which could very likely not be translated into a functional protein while others may have

553 had a function. As a first example, mutations inducing a premature stop codon could
554 lead to a shorter protein that might sometimes perform the same function. Yet in many
555 other cases shorter proteins might not perform the same function, if able to perform any
556 function at all. Another example concerns loss of the expected stop codon. When
557 screening a sequence, a stop codon will eventually be encountered, but determining the
558 functional consequences of this additional amino acid stretch would be impossible *in*
559 *silico*. The same holds when the start codon is lost. Determining the criteria according to
560 which an alternative start codon (if any) may become the new start codon, is a
561 hazardous task. These few examples highlight the extent to which sorting out different
562 functional scenarios is challenging. Moreover, mRNA molecules may play a regulating
563 role, even if they cannot be translated as such, thereby justifying the need to annotate
564 them. These reflexions led us to voluntarily disregard such interpretations in our re-
565 annotation process.

566 We observed that a third of the LRR-CR genes were non-canonical overall, but their
567 proportion varied according to the gene subfamily (**Table 2**). A lower proportion of LRR-
568 RLK genes were non-canonical (15%), compared to LRR-RLP (32%) and NLR (37%).
569 The LRR-RLK subfamily could be divided into 15-20 subgroups based on phylogenetic
570 study findings, and the duplication rate was shown to be quite variable according to the
571 subgroup considered (Tang et al., 2010; Fischer et al., 2016). Some subgroups, whose
572 genes have been described to be mostly involved in developmental processes, have
573 had a more stable copy number over the course of angiosperm evolution (Fischer et al.,
574 2016). These genes are less prone to duplication, and thus less likely to generate copies
575 accumulating nonsense mutations, thereby lowering the proportion of non-canonical
576 genes when the entire LRR-RLK subfamily is considered. The higher proportion of non-
577 canonical genes obtained for NLR and LRR-RLP suggests that these subfamilies
578 generally have higher birth and death rates. A quarter of the LRR-CR genes required
579 manual curation and were non-canonical (representing 83.5% of the curated loci and
580 89.5% of the non-canonical loci (**Supplemental table 1**)). In fact, manual curation was
581 conducted mainly when none of the three annotations gave satisfactory gene models
582 (such as the example presented in **Figure 3B**). The high correlation between non-

583 canonical and curated loci was thus likely due to the presence of mutations introducing
584 ambiguities, which are overcome to different extents by the three annotation pipelines.

585 We stress that this categorization of loci into a canonical or non-canonical models could
586 be impacted by the genomic sequence quality. Errors in the reference genome
587 sequence could introduce errors in the gene models. In our curated dataset, 27 non-
588 canonical genes were tagged in NCBI data as harbouring a difference when considering
589 the reference genome sequence and the RNAseq data. The mutations jeopardizing the
590 expected gene structure corresponded exactly to the positions where inconsistencies
591 had been highlighted between the genomic and RNAseq sequences. Without RNAseq
592 data, genome sequence errors cannot be differentiated from real mutations.
593 Redundancies in LRR-CR gene sequences can give rise to ambiguities during both
594 genome sequence assembly and expression data mapping, thus leading to errors
595 (Torresen et al., 2019). Only specific re-sequencing could resolve those potential
596 inconsistencies. In rice, this concerns only 2.6% of the genes in the current state of the
597 data presented in this study.

598 **LRR-CR repertoire in Kitaake and comparison with Nipponbare**

599 We propose a modular strategy to transfer our manually curated annotations to other
600 rice genomes. We applied this strategy to annotate LRR-CR genes from the genome of
601 the Kitaake cultivar, which also belongs to the *O. sativa* Japonica subspecies. A
602 comparison of the Nipponbare and Kitaake LRR-CR repertoires revealed an equivalent
603 number of loci. The distributions of LRR-CR loci per gene subfamily, chromosome and
604 category (canonical or not) were also consistent between these two cultivars (**Figure 5**).
605 In the Nipponbare genome, eight new LRR-CRs (1 LRR-RLK, 3 LRR-RLPs, and 4 UCs)
606 were identified. These genes had not been previously annotated in any of the three
607 publicly available annotations. In Kitaake, the same strategy enabled us to identify 114
608 new LRR-CR genes (48 of which were canonical): 17 LRR-RLKs, 24 LRR-RLPs, 50
609 NLRs and 23 UCs. The higher number of newly identified genes in Kitaake (114 vs 8)
610 suggested that annotation inaccuracies had a greater impact on recently sequenced
611 genomes that have not benefited from as much annotation investment as reference
612 genomes.

613 A comparison of the LRR motif number for all allelic pairs between Nipponbare and
614 Kitaake revealed a much greater difference in LRR number between alleles for publicly
615 available annotations (ranging from 2.68 to 3.58) in comparison to our manually curated
616 annotations (0.58) when the three subfamilies were all considered (**Figure 7** and
617 **Supplemental Figure 3**). When publicly available annotations are considered, some
618 rare allelic pairs harboring a very different number of LRR may be truly different
619 functional alleles. For instance, between two LRR-RLP alleles, one may contain a
620 premature stop-codon leading to the loss of a few motifs, but it may still have a biological
621 function. It is important to identify such a pair. In our expert annotation, alleles may
622 share an identical number of LRR but such allelic pairs would be clearly identifiable
623 because one of the alleles would be tagged as non-canonical while the other would be
624 canonical. Moreover, in non-canonical alleles, the causal mutation, its position and
625 impact on the gene (i.e. frameshift or premature stop-codon) could be identified.

626 The difference in LRR motif number observed between allele pairs was greater for NLR
627 than for the other subfamilies (**Figure 7**). This might be explained by the fact that NLR
628 motifs are more variable and hence harder to detect (Ng et al., 2011), which could lead
629 to apparent variations in the number of motifs in the two alleles. LRR motifs that have
630 been found in NLRs differed from the common “Plant Specific” LRR consensus
631 sequence, and were more irregular in terms of both length and residue conservation
632 (Kajava, 1998; Kuang et al., 2004; Matsushima and Miyashita, 2012; Sela et al., 2012).
633 Although we enhanced the LRR detection accuracy through the development of a new
634 LRR HMM profile for NLR, it is still not exhaustive. This also suggests that the number of
635 LRR motifs varies more in NLR than in other LRR-CR subfamilies.

636 High sequence similarity was observed between Nipponbare and Kitaake alleles (98.9%
637 identity for cDNA) (**Supplemental Table 2**), which was consistent with previous
638 comparisons (Jain et al., 2019). However, some allelic pairs showed a lower identity
639 level with a more ancient coalescent history between the two genomes. This
640 heterogeneity may have been the consequence of breeding programs from which these
641 varieties were derived. The breeding process involves crosses with more or less closely
642 related genotypes, sometimes from different subspecies, and may generate mosaic

643 genomes (Santos et al., 2019). No allelic pairs were found for 111 genes, i.e. 49 were
644 specific to Nipponbare and 62 to Kitaake. A majority of those genes were located in
645 clusters on chromosomes 2, 11 and 12, which have already been described as
646 containing a large number of LRR-CRs (Zhou et al., 2004; Mizuno et al., 2020)(**Figure**
647 **6**). Some clusters have also been shown to be less conserved (Mizuno et al., 2020).
648 More than half of these genes were classified as canonical, so they would generally be
649 more effectively functional.

650 The methods we developed allowed us to undertake an exhaustive comparison of the
651 LRR-CR repertoire between Nipponbare and Kitaake. Allelic pairs, including those
652 hosting nonsense mutations in either or both genotypes, were described (**Supplemental**
653 **Data Set 3**). Genotype-specific genes were also identified and localized, again along
654 with information related to the potential presence of nonsense mutations (**Figure 6**).
655 These results were achieved through a combination of an expert annotation and its
656 transfer to a second genotype for which a high quality *de novo* genome assembly was
657 available. Validation of the LRR-CR annotations of Kitaake was not very time consuming
658 compared to the initial work in Nipponbare where each gene was investigated
659 individually. Our study highlighted that investment in long-read sequence technologies
660 would be warranted in order to be able to work with high quality *de novo* assemblies,
661 especially when allelic diversity discovery is targeted.

662 The tools and curated datasets we generated in this study are available on these
663 websites: https://rice-genome-hub.southgreen.fr/content/geloc (data) and
664 https://github.com/cgottin/LRRprofiler (tools). We believe that evolutionary studies and
665 allele discovery initiatives on LRR-CRs would be more accurate and reliable when using
666 our new manually curated 'comprehensive' annotations for these genes. Moreover, we
667 feel that this 'comprehensive' annotation approach should be widely adopted by the
668 community in the light of its major potential benefits.

669 **Material and Methods**

670 **Genomes and annotation files**

671 Reference genomic sequences of Nipponbare (Kawahara et al., 2013) and Kitaake (Jain
672 et al., 2019) *Oryza sativa* ssp. *japonica* cultivars were downloaded from the Rice
673 Annotation Project Database (RAP-DB) website (https://rapdb.dna.affrc.go.jp/) and the
674 Phytozome website (https://phytozome.jgi.doe.gov/pz/portal.html). The general feature
675 format (GFF) and fasta files with coding DNA sequences (CDS) and protein sequences
676 for Nipponbare were downloaded for three different annotation projects: (i) the MSU
677 (v7.0) annotation was downloaded from the Rice Genome Annotation Project FTP
678 server (http://rice.plantbiology.msu.edu/), (ii) the IRGSP annotation files were
679 downloaded from the RAP-DB website (https://rapdb.dna.affrc.go.jp/), and (iii) the NCBI
680 annotation (release 102) annotated by the NCBI Eukaryotic Genome Annotation Pipeline
681 was downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov/). The IRGSP
682 annotation consists of two gene sets ('genes supported by FL-cDNAs, ESTs or proteins'
683 and 'computationally predicted genes') that were concatenated for the analyses (Sakai
684 et al., 2013).

685 **LRRprofiler implementation**

686 The LRRprofiler pipeline was implemented in two steps (**Supplemental Figure 4**). The
687 first step involved iterative refinement of LRR HMM profiles specific to a gene subfamily
688 (LRR-RLK or NLR) and proteome (inspired by (Ng et al., 2011), **Supplemental Figure**
689 **4A**). Only LRR-RLK and NLR were considered for profile refinement because they
690 contain a specific domain (i.e. kinase and NB-ARC domains, respectively), thereby
691 allowing clear identification of the subfamily to which they belong. A set of candidate
692 protein sequences was identified from a given proteome to refine the specific LRR
693 profiles. This set was composed of either LRR-RLKs identified with iTAK (Zheng et al.,
694 2016) or NLRs identified with the PF00931 Pfam NB-ARC profile. A first round of LRR
695 motif detection was performed in either of the candidate protein sets using hmmsearch
696 (HMMER, (Eddy, 2011)) with the SM00370 LRR profile from the SMART database.
697 Motifs of 20 to 26 amino acids in length were extracted, aligned with MAFFT with default
698 parameters and a new profile was built from the alignment using hmmbuilt (HMMER,
699 (Eddy, 2011)). This process was repeated using the HMM LRR profile built at the

700 previous iteration to search again for LRR motifs in the considered protein candidate set.
701 At each iteration, the sum of the amino acid lengths of the detected LRR motifs was
702 calculated. The process stopped when three iterations (not necessarily consecutive)
703 resulted in a decrease of this statistics. Finally, the process retrieved the HMM LRR
704 profile identifying the maximum number of LRR motifs in the candidate protein set.

705 The second step of the LRRprofiler pipeline consisted of identification of LRR-CR
706 proteins contained in a given proteome, annotation of their functional domains, as well
707 as their classification into a gene subfamily: LRR-RLK, LRR-RLP, NLR or UC
708 (**Supplemental Figure 4B**). Four publicly available LRR HMM profiles from the SMART
709 database, i.e. SM00367 (LRR_CC), SM00368 (LRR_RI), SM00369 (LRR_TYP) and
710 SM00370 (LRR), in addition to the newly built LRR profiles obtained in the first step,
711 were used to detect LRR motifs in the complete considered proteome using hmmsearch.
712 An annotation of LRR domains containing the start and end positions of each LRR motif
713 was output. The annotation of each protein was then supplemented using publicly
714 available profiles for other functional domains: TIR (PF01582), Malectin (PF11721),
715 Malectin-like (PF12819), RPW8 (PF05659), Cys-Pairs (Dievart and Clark, 2003;
716 Dufayard et al., 2017), F-box (PF00646) and FBD (PF08387). NB-ARC and kinase
717 domain annotations were retrieved from the first step, while transmembrane domains
718 (TM) were detected with TMHMM (v2.0c) with default parameters (Sonnhammer et al.,
719 1998). The subfamily assignment of each identified LRR-containing protein was
720 deduced from its domain structure. Proteins were classified into the LRR-RLK subfamily
721 if they contained LRRs and a kinase domain, and sometimes other domains such as the
722 malectin, malectin-like, cys-pair and TM domains. Proteins were classified in the NLR
723 subfamily if they included an NB-ARC domain and LRRs, sometimes with a TIR or an
724 RPW8 domain. The LRR-RLP subfamily included proteins with LRRs plus a TM,
725 malectin, malectin-like and/or cys-pair, or LRR-only structures when at least 10 plant-
726 specific LRR repeats were detected. Proteins containing an F-box or an FBD domain in
727 addition to LRRs were classified as F-box-LRR. All other LRR-containing proteins were
728 ranked in the UC group, and for these we performed a BLASTp search with default
729 parameters against the other gene sets (LRR-RLP, LRR-RLK, NLR and F-box) to
730 estimate their probable membership in one of these gene subfamilies. F-box proteins

731 were removed from our datasets and not further considered in the analyses. We ended
732 up with four gene sets: LRR-RLP, LRR-RLK, NLR and UC.

733 At the end of the construction phase, the complete LRRprofiler pipeline was tested on
734 the manually reviewed *Arabidopsis thaliana* protein dataset downloaded from the Swiss-
735 Prot section (https://www.uniprot.org/) (**Supplemental Methods, Supplemental Figure**
736 **5, Supplemental Table 3, Supplemental Data Set 4**) (Boutet et al., 2007) of the
737 UniProt databank (UniProt, 2019). This set was composed of 15,818 sequences.
738 Domain and repeat information was also extracted from the database, especially the
739 number of LRR motifs per sequence and the gene subfamily to which it belonged (LRR-
740 RLP, LRR-RLK, NLR, etc.).

741 **Rice transcription factor dataset**

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

746 **Annotation transfer from Nipponbare to Kitaake**

747 The first phase consisted of localization of regions of interest in the Kitaake genome, i.e.
748 region homologous to Nipponbare LRR-CR loci. Nipponbare LRR-CR protein sequences
749 from our expert annotations were aligned to the Kitaake genome with tBLASTx (Altschul
750 et al., 1990). Only high scoring pairs (HSPs) with more than 65% identity with
751 Nipponbare LRR-CR protein fragments and spanning at least 50% of the Nipponbare
752 query protein were retained. To define coherent candidate regions in Kitaake, HSPs
753 from the same Nipponbare query protein had to be located less than 5,000 bp apart,
754 except when the query Nipponbare homologous gene had a longer intron. In that case,
755 the Nipponbare intron length plus 500 bp was used as the upper bound for the distance
756 separating Kitaake HSPs. Multiple regions of interest could be found for a single
757 Nipponbare protein. This allowed us to annotate genes duplicated in the Kitaake
758 genome even if a single gene copy was present in the Nipponbare genome. In a second
759 phase, gene model determination was attempted in three consecutive steps for each

760 region of interest. Only regions that could not be successfully annotated at a given step
761 passed to the next one. In the first step, the Nipponbare query exons are mapped to the
762 target Kitaake region of interest with BLASTn. A gene model was then reconstructed
763 based on ordered HSPs. The gene model reconstruction quality was checked by
764 comparing the predicted protein with that of Nipponbare using BLASTp. The gene model
765 was retained if all expected exons were present and the Kitaake protein sequence had
766 more than 90% identity with the Nipponbare protein sequence. Otherwise annotation of
767 this region was delegated to the second step. In the second step, the Exonerate
768 cdna2genome model (Slater and Birney, 2005) was run independently for every
769 remaining query/target pair. The Exonerate output GFF file was parsed to construct the
770 target gene model and to document putative frameshift positions. Again, the Kitaake
771 predicted protein was compared to the Nipponbare query sequence with BLASTp and
772 retained if the coverage and identity were above 90% and 75%, respectively. Otherwise
773 annotation of this target region was delegated to the third step. In the third step, the
774 remaining loci were reconstructed with the Exonerate protein2genome model. This
775 model is better at finding the correct reading frames when the target and model loci are
776 more divergent, but it fails to correctly annotate type 1 and 2 splicing sites (intron/exon
777 junction falling inside a codon). This problem arises since it uses the same reading
778 frame to translate the whole genomic sequence (the six reading frames are tested, but
779 each resulting translation just uses one of them). To overcome this issue, intron
780 junctions are then corrected with a Python script that looks for canonical splicing sites in
781 a range of two nucleotides before and after the current junctions. Finally, gene models
782 highly divergent from the Nipponbare query sequence, with multiple premature stop
783 codons or without start or terminal stop codons, and overlapping frameshifts are tagged
784 to be manually checked.
785

786 **Identification of alleles between Nipponbare and Kitaake**

787 We used Synmap (Lyons and Freeling, 2008) to identify LRR-CR allelic pairs, i.e. genes
788 having the exact same chromosomic position in Nipponbare and Kitaake. Synmap was
789 developed to identify orthologous genes between different species based on
790 microcolinearity conservation, and it identifies blocks of genes of conserved order and

791 position. It retrieves a list of relationships between genic repertoires of two genomes. We
792 identified alleles by first selecting genes for which Synmap found a reciprocal
793 relationship, i.e. a relationship found in both Nipponbare to Kitaake and Kitaake to
794 Nipponbare comparisons. Genes for which allelic relationships could not be
795 unambiguously resolved by Synmap were manually resolved, when possible, using
796 VISTA (Mayor et al., 2000) and Artemis (Carver et al., 2012) software packages.
797

798 **Availability of new reannotated data**

799 A new identifier was allocated to each LRR-CR gene unravelled by this procedure.
800 These identifiers use the <OSJnip_ChrXX_00000000> or <OSJkit_ChrXX_00000000>
801 pattern for Nipponbare and Kitaake loci, respectively, with XX being the chromosome
802 number followed by the start codon position of the coding sequence (CDS) on the
803 chromosome (**Supplemental Data Sets 1 and 2**).

804 According to the MACSE convention (Ranwez et al., 2011; Ranwez et al., 2018), indels
805 causing frameshift mutations have been pinpointed by the presence of one or two '!'
806 characters in the nucleotide and protein sequences of non-canonical genes and are
807 available in an additional specific dataset posted on the website

808 (https://rice-genome-hub.southgreen.fr/content/geloc).

809

810 **Supplemental Data**

811 **Supplemental Methods**. Validation of the LRRprofiler pipeline.

812 **Supplemental Data Set 1**. LRR-CR loci from the Nipponbare rice reference genome.

813 Manually curated annotation of Nipponbare LRR-CR loci. For each LRR-CR gene, the
814 corresponding identifier (ID) from public annotations (IRGSP, MSU and NCBI) are given.
815 Known gene names (aliases) are also given and were retrieved from the NBRP
816 OryzaBase database. Substantial gene information is provided, including the gene sub-
817 family, gene category (canonical or non-canonical), the fact that the gene model was

818 modified (y) or not (n) during the curation process and the number of LRR motifs
819 annotated by the LRR profiler pipeline.

820 **Supplemental Data Set 2**. LRR-CR loci from the rice KitaakeX genome.

821 Manually curated annotation of Kitaake LRR-CR loci after transfer of Nipponbare loci.
822 For each LRR-CR gene, the corresponding identifier (ID) from public annotation are
823 given. Substantial gene information is provided, including the gene sub-family, gene
824 category (canonical or non-canonical) and the number of LRR motifs annotated by the
825 LRR profiler pipeline.

826 **Supplemental Data Set 3**. Allelic relationship and cDNA identity between Nipponbare
827 and Kitaake LRR-CR loci.

828 Each line provides the allelic pair identifiers (ID) from Nipponbare (NIP) and Kitaake
829 (KIT) cultivars. Allelic relationships were determined with Synmap, followed by manual
830 curation.

831 **Supplemental Data Set 4**. LRRprofiler results in the Swiss-Prot *A. thaliana* dataset.

832 LRR-containing proteins extracted from the Swiss-Prot database and results of the
833 LRRprofiler pipeline. The orange section presents information on gene sub-families and
834 number of annotated LRR motifs extracted from the Swiss-Prot database. The blue
835 section presents the LRRprofiler results for each gene.

836 **Supplemental Figure 1**. Comparison of predicted peptide lengths between Nipponbare
837 publicly available annotations (IRGSP, MSU, NCBI) for LRR-CR and TF loci.

838 The plots represent pairwise comparaisons of peptide lengths (in amino acids) predicted
839 by IRGSP and MSU (left), IRGSP and NCBI (centre) and MSU and NCBI (right) for LRR-
840 CR genes (upper) and TF genes (lower). Numbers at the bottom right of each graph are
841 the coefficient of determination (R^2) and gene family average relative length difference
842 (RLD). The grey dotted lines on each graph represent the diagonal y=x (dots are shown
843 along this line when the two considered annotations predicted peptides of the exact
844 same length for a given gene).

845 **Supplemental Figure 2**. Number of domains and motifs identified with LRRprofiler for
846 the Nipponbare proteomes predicted by publicly available and manualy curated
847 annotations.

848 **Supplemental Figure 3**. LRR motif number conservation between Nipponbare and
849 Kitaake LRR-CR loci depending on the compared annotation.

850 The public annotation for the Kitaake genome was compared to the Nipponbare IRGSP
851 (green), MSU (orange) and NCBI (blue) annotations. Our manually curated annotations
852 compared for Nipponbare and Kitaake are shown in pink. The 'Mean(diff)' statistic is the
853 average difference in the number of LRR motifs between the two cultivar alleles. The
854 dashed line on each graph represents the y=x diagonal.

855 **Supplemental Figure 4**. Schematic representation of the LRRprofiler pipeline.

856 **(A)** HMM profile refinement sub-process for LRR motifs. In a first step, candidate
857 proteins were identified for LRR-RLK or NLR subfamilies respectively using iTAK or
858 hmmsearch with the PF00931 NB-ARC profile. In a first round of LRR identification, one
859 of the two candidate protein sets was run with the SM00370 LRR profile as input. Then
860 identified LRRs were aligned with mafft and a new profile was built with hmmbuild. This
861 new profile was used to search LRR motifs in the chosen candidate protein set. The last
862 steps, i.e. LRR motif alignment, construction of a new profile and searchs for LRR motifs
863 with the new profile, were repeated until the stop condition was fulfilled. Finally, the best
864 profile from all iterations was retrieved as the new specific profile for the subfamily
865 depending on the candidate protein set (NLR or LRR-RLK).

866 **(B)** LRR-CR protein annotation and classification. Newly refined and SMART LRR
867 profiles (SM00367, SM00368, SM00369, and SM00370) were used to identify LRR-
868 containing proteins from a proteome and to annotate LRR motifs in these proteins, which
869 required handling of motif match overlaps. Other functional domains were annotated
870 with public HMM profiles. For each LRR-containing protein, the gene subfamily (LRR-
871 RLK, LRR-RLP, NLR or F-box-LRR) to which it belonged was deduced from its domain
872 structure. Proteins with at least an NB-ARC and LRRs were NLRs, proteins with at least
873 a kinase and LRRs were LRR-RLKs, while proteins with LRRs and a TM or a malectin(-

like) domain without kinase were LRR-RLPs. Structures with an F-box or FBD were F-box-LRRs. All proteins with alternative structures were pooled in the unclassified (UC) group. *F-box and FBD are domains specific to F-box-LRR proteins. F-box-LRRs are other intracellular receptors that were not studied here. Their identification allowed us to separate them from the other LRR-CR proteins.

Supplemental Figure 5. Comparison of expected and predicted LRR motifs in protein sequences from the Swiss-Prot *A. thaliana* dataset using publicly available and refined HMM profiles.

Number of expected (x axis) and predicted (y axis) LRR motifs in protein sequences from the Swiss-Prot *A. thaliana* dataset using public and new refined HMM profiles. The upper section compares motif numbers for LRR-RLP and LRR-RLK (concatenated in LRR-RL), and the lower section compares motif numbers for NLR.

Supplemental Table 1. Contingency table of canonical/non-canonical and modified/not modified LRR-CR loci from the Nipponbare manually curated annotation.

Supplemental Table 2. Percentage of cDNA identity between Nipponbare and Kitaake alleles according to gene sub-families and categories. Categories are given as Nipponbare/Kitaake. C, canonical; NC, non-canonical.

Supplemental Table 3. Performance of publicly available and refined LRR HMM profiles in the Swiss-Prot *A. thaliana* dataset.

The third line (Expected) presents the number of proteins belonging to each sub-family and the number of motifs annotated for these proteins in the Swiss-Prot database.

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 critical reading of the manuscript.

902

903 **Author contributions**

904 CG, NC, AD, CP and VR designed the research; CG, NC, AD performed the research;
905 CG, NC, AD, VR, MS, GD contributed to new analytic and computational tools; CG, NC,
906 AD analyzed the data; CG, NC, AD, VR wrote the paper.

907

908

Parsed Citations

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., Garcia Giron, C., Hourlier, T., Howe, K., Kahari, A., Kokocinski, F., Martin, F.J., Murphy, D.N., Nag, R., Ruffier, M., Schuster, M., Tang, Y.A., Vogel, J.H., White, S., Zadissa, A., Flicek, P., and Searle, S.M. (2016). The Ensembl gene annotation system. Database (Oxford) 2016.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Bailey-Serres, J., Parker, J.E., Ainsworth, E.A., Oldroyd, G.E.D., and Schroeder, J.I. (2019). Genetic strategies for improving crop yields. Nature 575, 109-118.
Google Scholar: Author Only Title Only Author and Title

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

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.
Google Scholar: Author Only Title Only Author and Title

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). UniProtKB/Swiss-Prot. Methods Mol Biol 406, 89-112.
Google Scholar: Author Only Title Only Author and Title

Burdett, H., Bentham, A.R., Williams, S.J., Dodds, P.N., Anderson, P.A., Banfield, M.J., and Kobe, B. (2019). The Plant "Resistosome": Structural Insights into Immune Signaling. Cell Host Microbe 26, 193-201.
Google Scholar: Author Only Title Only Author and Title

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

Dievart, A., and Clark, S.E. (2003). Using mutant alleles to determine the structure and function of leucine-rich repeat receptor-like kinases. Curr Opin Plant Biol 6, 507-516.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol 7, e1002195.
Google Scholar: Author Only Title Only Author and Title

FAO. (2018). The future of food and agriculture - Alternative pathways to 2050. (Rome).
Google Scholar: Author Only Title Only Author and Title

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

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Han, G.Z. (2019). Origin and evolution of the plant immune system. New Phytol 222, 70-83.
Google Scholar: Author Only Title Only Author and Title

Hwang, S.G., Kim, D.S., and 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.
Google Scholar: Author Only Title Only Author and Title

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11, 97-108.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Kajava, A.V. (1998). Structural diversity of leucine-rich repeat proteins. J Mol Biol 277, 519-527.
Google Scholar: Author Only Title Only Author and Title

Kajava, A.V. (2012). Tandem repeats in proteins: from sequence to structure. J Struct Biol 179, 279-288.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Lai, X., Chahtane, H., Martin-Arevalillo, R., Zubieta, C., and Parcy, F. (2020). Contrasted evolutionary trajectories of plant transcription factors. Curr Opin Plant Biol 54, 101-107.
Google Scholar: Author Only Title Only Author and Title

Leister, D. (2004). Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. Trends Genet 20, 116-122.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53, 661-673.
Google Scholar: Author Only Title Only Author and Title

Matsushima, N., and Miyashita, H. (2012). Leucine-Rich Repeat (LRR) Domains Containing Intervening Motifs in Plants. Biomolecules 2, 288-311.
Google Scholar: Author Only Title Only Author and Title

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. (2000). VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16, 1046-1047.
Google Scholar: Author Only Title Only Author and Title

McDowell, J.M., and Simon, S.A. (2006). Recent insights into R gene evolution. Mol Plant Pathol 7, 437-448.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Nei, M., and Rooney, A.P. (2005). Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39, 121-152.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Richter, T.E., and Ronald, P.C. (2000). The evolution of disease resistance genes. Plant Mol Biol 42, 195-204.
Google Scholar: Author Only Title Only Author and Title

Roy, S.W. (2016). How Common Is Parallel Intron Gain? Rapid Evolution Versus Independent Creation in Recently Created Introns in Daphnia. Mol Biol Evol 33, 1902-1906.
Google Scholar: Author Only Title Only Author and Title

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.C., Iwamoto, M., Abe, T., Yamada, Y., Muto, A., Inokuchi, H., Ikemura, T., Matsumoto, T., Sasaki, T., and Itoh, T. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol 54, e6.
Google Scholar: Author Only Title Only Author and Title

Santos, J.D., Chebotarov, D., McNally, K.L., Bartholome, J., Droc, G., Billot, C., and Glaszmann, J.C. (2019). Fine Scale Genomic Signals of Admixture and Alien Introgression among Asian Rice Landraces. Genome Biol Evol 11, 1358-1373.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Sela, H., Spiridon, L.N., Petrescu, A.J., Akerman, M., Mandel-Gutfreund, Y., Nevo, E., Loutre, C., Keller, B., Schulman, A.H., and Fahima, T. (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. Mol Plant Pathol 13, 276-287.
Google Scholar: Author Only Title Only Author and Title

Shao, Z.Q., Wang, B., and Chen, J.Q. (2016). Tracking ancestral lineages and recent expansions of NBS-LRR genes in angiosperms. Plant Signal Behav 11, e1197470.
Google Scholar: Author Only Title Only Author and Title

Sharma, M., and Pandey, G.K. (2015). Expansion and Function of Repeat Domain Proteins During Stress and Development in Plants. Front Plant Sci 6, 1218.
Google Scholar: Author Only Title Only Author and Title

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2, ii215-225.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Sun, X., and Wang, G.L. (2011). Genome-wide identification, characterization and phylogenetic analysis of the rice LRR-kinases. PLoS One 6, e16079.
Google Scholar: Author Only Title Only Author and Title

Tamborski, J., and Krasileva, K.V. (2020). Evolution of Plant NLRs: From Natural History to Precise Modifications. Annu Rev Plant Biol 71, 355-378.

Google Scholar: Author Only Title Only Author and Title

Tang, P., Zhang, Y., Sun, X., Tian, D., Yang, S., and Ding, J. (2010). Disease resistance signature of the leucine-rich repeat receptor-like kinase genes in four plant species. Plant Science 179, 399-406.
Google Scholar: Author Only Title Only Author and Title

Torresen, O.K., Star, B., Mier, P., Andrade-Navarro, M.A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A.V., Promponas, V.J., Anisimova, M., Jakobsen, K.S., and Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res 47, 10994-11006.
Google Scholar: Author Only Title Only Author and Title

UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47, D506-D515.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Wilming, L., and Harrow, J. (2009). Gene Annotation Methods. In Bioinformatics, D. Edwards, J. Stajich, and D. Hansen, eds (Springer, New York, NY.
Google Scholar: Author Only Title Only Author and Title

Xiong, Y., Han, Z., and Chai, J. (2020). Resistosome and inflammasome: platforms mediating innate immunity. Curr Opin Plant Biol 56, 47-55.
Google Scholar: Author Only Title Only Author and Title

Yenerall, P., and Zhou, L. (2012). Identifying the mechanisms of intron gain: progress and trends. Biol Direct 7, 29.
Google Scholar: Author Only Title Only Author and Title

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R., Lee, D., Quackenbush, J., and Buell, C.R. (2003). The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res 31, 229-233.
Google Scholar: Author Only Title Only Author and Title

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.
Google Scholar: Author Only Title Only Author and Title

Zhou, T., Wang, Y., Chen, J.Q., Araki, H., Jing, Z., Jiang, K., Shen, J., and Tian, D. (2004). Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Genet Genomics 271, 402-415.
Google Scholar: Author Only Title Only Author and Title