agronomy

MDPI

Article

SNPs, InDels, and Microsatellites within and Near to Rice NBS-LRR Resistance Gene Candidates

Mark J. Quinton-Tulloch and Katherine A. Steele *

School of Natural Sciences, Bangor University, Gwynedd LL57 2UW, UK; m.quinton-tulloch@bangor.ac.uk
* Correspondence: k.a.steele@bangor.ac.uk

Abstract: Plant resistance genes (R-genes) drive the immune responses of crops against specific pathotypes of disease-causing organisms. Over time, genetic diversity in R-genes and R-pseudogenes has arisen among different rice varieties. This bioinformatics study was carried out to (i) predict the full sets of candidate nucleotide-binding site leucine-rich repeat (NLR) R-genes present in six rice genomes; (ii) detect variation within candidate R-genes; (iii) identify potential selectable markers within and near to LRR genes among 75 diverse *indica* rice genomes. Four high quality *indica* genomes, plus the standard *japonica* and *indica* reference genomes, were analysed with widely available bioinformatic tools to identify candidate R-genes and R-pseudogenes. They were detected in clusters, consistent with previous studies. BLAST analysis of cloned protein sequences of 31 R-gene loci gave confidence in this approach for detection of cloned NLR R-genes. Approximately 10% of candidate R-genes were located within 1 kb of a microsatellite (SSR) marker. Sequence comparisons among *indica* rice genomes detected SNPs or InDels in 334 candidate rice R-genes. There were significantly more SNPs and InDels within the identified NLR R-gene candidates than in other types of gene. The genome-wide locations of candidate R-genes and their associated markers are presented here for the potential future development of improved disease-resistant varieties. Limitations of in silico approaches used for R-gene discovery are discussed.

Keywords: pathogen; resistance; NBS-LRR; NLR; *Oryza sativa*; rice blast; bacterial blight

Citation: Quinton-Tulloch, M.J.; Steele, K.A. SNPs, InDels, and Microsatellites within and Near to Rice NBS-LRR Resistance Gene Candidates. *Agronomy* **2021**, *11*, 2297. https://doi.org/10.3390/agronomy11112297

Academic Editor: Yong-Bao Pan

Received: 27 September 2021
Accepted: 10 November 2021
Published: 13 November 2021

1. Introduction

Plant pathogens contribute to yield losses of up to 30% of global rice (*Oryza sativa* L.) production, leading to as much as 40.9% of crop losses in some major food security hotspots [1]. Crop varieties that are genetically immune to pathogens are considered to be more environmentally sustainable than the use of chemicals for pathogen control. The immune response of plants to pathogens is characterised by two interconnected pathways, primarily defined by the pathogen molecules recognised by the plant [2]. PAMP-triggered immunity (PTI) involves the recognition of conserved pathogen-associated molecular patterns (PAMPs), essential for pathogen survival, by host cell surface transmembrane pattern recognition receptors (PRRs). This triggers intracellular signalling events and changes in host gene expression that restrict microbial movement within the host. Non-host-specific resistance is generally considered to be controlled by PTI. The second pathway of the immune system is known as the effector-triggered immunity (ETI) pathway. ETI controls host-specific resistance through recognition of pathogen effectors, also known as avirulence genes, that encode species-, race-, or even strain-specific proteins. Plants have coevolved specific resistance (R) genes, which encode cytoplasmic protein receptors that bind to the complementary avirulence proteins and activate ETI, often leading to localised cell death that prevents pathogen spread [2]. R-genes are considered to have a large qualitative effect and follow Mendeilan inheritance [2]. Some ETI and PTI genes can contribute to quantitative disease resistance, although several other mechanisms can be involved in quantitative resistance [2].

Broad-spectrum resistance to multiple races of a pathogen can be provided through stacking different R-genes with complementary resistance spectra, and this strategy can potentially delay the breakdown of resistance. In some cases, the combination of R-genes can result in quantitative complementation with stronger resistance being seen in the offspring than in parental lines [3,4]. Pyramiding of functional R-genes by combining multiple R-genes into a single genome is a breeding technique that has been effectively used to improve the resistant phenotype of crop lines. Due to the masking of the action of individual R-genes by other R-genes [3], it is not always possible to select plants with multiple resistance genes based on phenotype. Therefore, the availability of polymorphic DNA markers, such as microsatellites or simple sequence repeats (SSRs), single nucleotide polymorphisms (SNPs) or insertions/deletions (InDels) are necessary for effective pyramiding of R-genes in breeding programs. Using closely linked markers flanking a target R-gene can greatly reduce the problem of linkage drag that occurs due to selection of an introgressed gene, a problem exemplified by the large introgression of wild rice genome in the region around *Pi33* of IR64 [5].

The largest group of R-genes in plants are nucleotide-binding site leucine-rich repeat (NBS-LRR; NLR) genes, characterised by nucleotide-binding sites (NBS) and LRR domains with variable amino- and carboxy-terminal domains. There are three groups of leucine-rich repeat (LRR)-containing receptor families: NLR and two other groups of pattern recognition receptor proteins which include a transmembrane (TM) domain [6]. The NBS domain is highly conserved, functioning as a molecular switch in disease signalling pathways, while the variable LRR domain is involved in the recognition of specific R proteins. NLR proteins are divided into two major subfamilies, according to the presence of either Toll/interleukin-1 receptors (TIR) or coiled-coil (CC) motifs in the amino-terminal domain. Both subfamilies are involved in pathogen recognition but have distinct sequence motifs and initiate different signalling pathways [7]. In contrast to *Arabidopsis* and other dicots, TIR-NLR proteins (TNLs) are not found in cereal species, suggesting that early angiosperms had few TNLs, which were subsequently lost in the cereal lineage. Genes coding for TIR domains have been found in the rice genome, but these have diverged from the NLR family of genes [8].

Previously, degenerate PCR primers, designed to target conserved sequence domains, have been used to isolate and identify novel resistance genes in a variety of plant species [9–11]. Bioinformatics approaches use similar conserved motifs in the NBS domain of plant R-genes to detect R-gene analogues in silico from genome sequences. P-loop, kinase-2, kinase-3a, RNBS-B, GLPL, and MHDL domains are common to both subfamilies of NLR genes, while TIR and non-TIR R-genes contain family-specific RNBS-A and RNBS-D motifs [12,13]. Recently, new motif searching strategies for the genome-wide identification of R-gene models have been developed. The increasing availability of whole genome sequences for different varieties is broadening the scope for using bioinfomatics for novel R-gene detection in breeding [14].

Bioinformatics studies of annotated genes in Nipponbare (*japonica*) and 93-11 (*indica*) predicted over 500 rice NLR genes, across all 12 chromosomes [15–18], and more were predicted in rice wild relatives [19] with frequent changes in copy number among genomes. In both *indica* and *japonica*, approximately one third of identified rice NLR genes were considered R-pseudogenes, having mutations that prevent functional resistance proteins from being encoded. Although similar numbers of R-genes were identified in 93-11 and Nipponbare, the 93-11 reference genome is incomplete and could contain undetected R-genes. However, the whole-genome shotgun sequencing strategy used for the 93-11 genome [20] could also artificially inflate the number of identified genes, due to single locus fragments erroneously being allocated to multiple loci during genome assembly [16].

This study aimed to catalogue the candidate R-genes or R-pseudogenes of *indica* rice, which accounts for more than 75% of global rice production. It complements other studies that have predicted rice R-gene candidates from the annotated genes of next generation studies by using high-quality genomes of four *indica,* alongside the older 93-11

and Nipponbare reference genomes. Diversity of nucleotide variation within the identified R-genes and pseudogenes was mined using genomes for nine *indica* rice lines [21] and 66 *indica* genomes selected from the 3000 Rice Genomes Project [22]. The SNPs, InDels, and SSR loci in the vicinity of R-genes and R-pseudogenes are provided in supplementary files as a community resource.

2. Materials and Methods

2.1. MEME/MAST/NLR-Parser R-Gene Search in Five Indica and One Japonica Reference Genomes

Jupe et al. [23] utilised the MEME suite [24] to identify 20 sequence motifs associated with NLR proteins, based on a positive training set of 53 NLR protein sequences and a negative training set of 16 non-NLR, nucleotide binding, or pattern recognition receptor proteins. These motifs were used here to conduct MAST [25] searches of the annotated amino acid sequences from the four high quality *indica* genomes—Shuhui498 [26], Zhenshan 97, Minghui 63 [27], and IR8 [28]—as well as 93-11 [20] and Nipponbare [29] genome assembly projects. Zhenshan 97 and Minghui 63 [27] are parents of the most widely cultivated hybrid rice variety in China (Shanyou 63). IR8 is the variety which helped drive the Green Revolution in Asia [28]. Shuhui498 (R498) is a restorer line used in a three-line hybrid system [26] and has platinum standard genome assemblies. The resulting outputs were parsed with NLR-parser [30] to identify candidate NLR proteins.

Annotated gene DNA sequences from each reference assembly were translated in all six reading frames and the above search strategy was repeated.

2.2. HMMER Motif Search in Five Indica and One Japonica Reference Genomes

HMMER version 3.1b2 (Available online: http://hmmer.org (accessed on 10 September 2019)) was used to search the same sets of amino acid sequences and 6-frame-translated DNA sequences described above for sequence homologs to PFAM domains related to TIR (PF01582), NBS (PF00931), and LRR (PF00560, PF01030, PF03382, PF05725, PF07723, PF07725, PF12799, PF13306, PF13516, PF13855, and PF14580) motifs.

2.3. Read Processing and Variant Calling in 75 Resequenced Indica Genomes

Details of the 75 *indica* lines used, including the BioSample IDs for the sequencing reads, are given in Supplementary File S1. Genome resequencing of nine *indica* rice lines (IR64, IR71033-4-1-127B, IR65482-4-136-2-2, IRBB60, Khumal-4, Loktantra (NR1487-2-1-2-2-1-1), Anamol Masuli, Sunaulo Sugandha, and Sugandha-1) was carried out as described in [21]. Sequencing reads are available from the NCBI Sequence Read Archive (BioProject accession PRJNA395505). In addition, sequencing reads were obtained from [22] for 66 *indica* rice lines. Reads were aligned against the assembled genome sequences of the four previously stated high-quality *indica* genomes, plus 93-11 [20] and Nipponbare, using Bowtie2 [31]. Neither discordant nor mixed alignment of paired reads were permitted and all other parameters were set as their default values.

SAMtools was used for variant calling with filtering carried out using the vcfutils package [32]. Variants with a read depth of greater than 200 were removed due to the likelihood of them resulting from variation between variable copy number repeats, and non-deletion variants with a read depth of less than 5 were also removed due to there being insufficient evidence of true variation at these sites.

2.4. Identification of Candidate Genes

The BLAST+ suite [33] was used to carry out a number of searches against annotated databases in order to annotate the identified candidate R-genes. BLASTN searches were carried out, querying the DNA sequences of the identified R-genes from the Shuhui498 assembly project [26] against the complete sets of gene DNA sequences from (a) the *Oryza sativa* ssp. *japonica* cv. Nipponbare genome assembly [34]—based on the IRGSP gene

models imported from RAP-DB [35]—and (b) the Beijing Genome Institute's assembly and annotation of the *O. sativa* ssp. *indica* cv. 93-11 genome [20,36].

A BLASTP search was also conducted, querying the Shuhui498 candidate R-gene amino acid sequences against the NCBI non-redundant protein database. In the case of genes with multiple annotated transcript isoforms, the first annotated isoform was used as the search query. Similarly, a BLASTN search was carried out, querying the annotated gene DNA sequences against the NCBI non-redundant nucleotide database.

The results of the motif searches of the annotated protein sequences of the six reference genomes were used to identify candidate R-genes. Genes were identified as candidate R-genes by two criteria: (i) if protein sequences corresponding to those genes were identified as complete NLR sequences by NLR-parser, or (ii) if there was evidence of both NBS and LRR domains from the HMMER motif searches.

The results of motif searches of the 6-frame-translated gene DNA were used to identify candidate resistance pseudogenes—these do not code for complete, functional, resistance proteins. Genes were considered to be candidate R-pseudogenes if the translated DNA sequences showed truncation or frameshift mutations and also met the criteria described above for candidate R-genes using protein sequences, which had not been identified as candidate R-genes by having both NBS and LRR domains identified in the HMMER motif searches.

2.5. Ortholog Identification

Ortholog identification was carried out with JustOrthologs [37]; therefore, we chose to use the term "ortholog" to describe the genes detected using this software, although, strictly speaking, they are allelic, or the result of recent duplication events with the same species, or both. Analyses were carried out for all pairwise combinations of the six reference genomes, using the complete sets of CDS DNA sequences as inputs. If JustOrthologs failed to detect an ortholog between an R-gene or R-pseudogene and any gene in one or more of the other reference genomes, further attempts to identify orthologs for that gene were undertaken by means of reciprocal BLAST analyses. Complete gene DNA sequences, including non-coding regions, were used for the reciprocal BLAST, with the gene sequences in question being queried against the full set of gene DNA sequences of the reference genome being searched. Genes identified as best hits were used as queries against the full set of gene DNA sequences from the original R-gene reference genome, and if the best hit matched the original gene, then the genes were considered to be orthologs.

2.6. Comparison with Cloned Resistance Genes

A total of 50 cloned R-gene protein sequences were downloaded from Genbank (Table 1): these were for 38 known rice blast resistance protein sequences (corresponding to 34 alleles at 20 loci) and 12 bacterial blight (BB) resistance protein sequences (corresponding to 11 alleles at 11 loci, of which only one was known to be an NLR). BLASTP searches against the annotated amino acid sequences for all six reference genomes were conducted.

Multiple sequence alignments were carried out for each resistance locus using Clustal Omega [38], aligning the cloned sequences against the reference amino acid sequence(s) with the best BLAST alignment scores across all six reference genomes searched, and any orthologs to these genes that had been identified in the other reference genomes.

Table 1. Cloned R-gene sequences used in BLASTP searches of the six reference genomes. Details of gene annotations are available via GenBank (www.ncbi.nim.nih.gov (accessed on 10 January 2019)). The allele in the source variety gives resistance to the pathogen indicated.

Locus	Chromosome	Gene Allele	GenBank ID	Source Variety
Resistance Conferred against Rice Blast Pathogen (*Magnaporthe grisea*)				
Pi2	6	*Pi2*	ABC94599.1	C101A51 (*indica*)
	6	*Pi9*	ABB88855.1	75-1-127 (*indica*)
	6	*Pi50*	AKS24975.1	NIL-e1 (*japonica*)
	6	*Piz-t*	ABC73398.1	Toride 1 (*japonica*)
Pi5-1	9	*Pi5-1*	ACJ54697.1	RIL260 (*japonica*)
	9	*Pii*	BAN59294.1	Hitomebore (*japonica*)
Pi5-2	9	*Pi5-2*	ACJ54698.1	RIL260 (*japonica*)
Pi21	4	*Pi21*	Q7XL73.1	(*japonica*)
Pi36	8	*Pi36*	ABI64281.1	Kasalath (*indica*)
Pi37	1	*Pish*	XP_015646397.1	Nipponbare (*japonica*)
	1	*Pi37*	ABI94578.1	St. No. 1 (*japonica*)
Pi54	11	*Pi54 (partial)*	AAY33493.1	Tetep (*indica*)
	11	*Pi54 (partial)*	CCD32366.1	Basmati 386 (*indica*)
Pi64	1	*Pi64*	BAS74649.1	Nipponbare (*japonica*)
Pi-b	2	*Pi-b*	BAA85975.1	BL-1 (*japonica*)
Pi-CO39	11	*Pi-CO39*	ARF20187.1	Yixiang1B (*indica*)
Pid-2	6	*Pid-2*	ALU57428.1	Nipponbare (*japonica*)
Pid3	6	*Pid3*	ACN62383.1	Digu (*indica*)
	6	*Pid3-A4*	ACN62387.1	A4 (*O. rufipogon*)
	6	*Pi25*	AFM35701.1	Zhongjian100 (*indica*)
Pik-1	11	*Pik1-KA*	BAL63005.1	Kanto51 (*japonica*)
	11	*Pi1-5*	AEB00617.1	C101LAC (*indica*)
	11	*Pikm-1*	BAG72135.1	Tsuyake (*japonica*)
	11	*Pikp-1*	ADV58352.1	(*japonica*)
Pik-2	11	*Pik2-KA*	BAL63006.1	Kanto51 (*japonica*)
	11	*Pikp-2*	ADV58351.1	(*japonica*)
	11	*Pi1-6*	AEB00618.1	C101LAC (*indica*)
	11	*Pikm-2*	BAG72136.1	Tsuyake (*japonica*)
Pit	12	*Pit*	BAH20864.1	Nipponbare (*japonica*)
Pi-ta	12	*Pi-ta*	AAO45178.1	Tsuyake (*japonica*)
Pb1	11	*Pb1*	BAJ25849.1	Yixiang1B (*indica*)
	11	*Pb1*	BAJ25848.1	St. No. 1 (*japonica*)
LABR_64-1	9	*LABR_64-1*	AIP86911.1	312007 (*japonica*)
	9	*LABR_64-1*	AIP86912.1	301279 (*japonica*)
LABR_64-2	9	*LABR_64-2*	AIP86925.1	312007 (*japonica*)
	9	*LABR_64-2*	AIP86926.1	301279 (*japonica*)
SasRGA4	11	*SasRGA4*	BAK39922.1	Sasanishiki, Aichi-asahi (*japonica*)
SasRGA5	11	*SasRGA*	BAK39930.1	Sasanishiki, Aichi-asahi (*japonica*)
Resistance Conferred against Bacterial Blight Pathogen (*Xanthomonas oryzae* pv. *oryzae*)				
Xa1	4	*Xa1*	BAA25068.1	IR-BB1 (*indica*)
Xa3/Xa26	11	*Xa3/Xa26*	AYH53004.1	Wase Aikoku 3 (*japonica*)
Xa4	11	*Xa4*	AQQ72929.1	Nipponbare (*japonica*)
	11	*Xa4*	AQQ72925.1	Minghui 63 (*indica*)
Xa5	5	*Xa5*	AHC94895.1	PB1 (*indica*)
Xa10	11	*Xa10*	AGE45112.1	IRBB10A (*indica*)
Xa13	8	*Xa13*	ABD78944.1	IR24 (*indica*)
Xa21	11	*Xa21*	AAC49123.1	IRBB21 (*indica*)
Xa23	11	*Xa23*	AIX09985.1	JG30 (*indica*)
Xa25	12	*Xa25*	AGS56391.1	Zhenshan 97 (*indica*)
Xa27	6	*Xa27*	AFO69279.1	Taichung Native 1 (*indica*)
Xa41	11	*Xa41(t)*	B8BKP4.1	93-11 (*indica*)

2.7. Sequence Variation Analysis

Custom Perl scripts were written to identify homozygous variations in 75 *indica* rice lines that were located within the candidate R-genes and R-pseudogenes of the reference genomes analysed. The type of mutation was predicted for each variation, based on its position within the annotated gene and the sequence changes.

Further analysis was carried out on variations compared against the Shuhui498 genome sequence [26]. Mutation type was predicted for all homozygous variations identified within all annotated genes. For each annotated transcript isoform, all variations occurring in that region were classified as one of the following gene positions: 5′ UTR; 5′ UTR intron; 3′ UTR; 3′ UTR intron; start codon; stop codon; CDS; intron; exon donor splice site; exon acceptor splice site; intron donor splice site; intron acceptor splice site. Exon donor and acceptor sites were defined as being 3 bases from the intron/exon boundary, and intron donor and acceptor sites were defined as being 8 and 30 bases from the boundary, respectively.

SNPs occurring within coding regions were categorised as non-synonymous or synonymous, while InDels within or overlapping coding regions were categorised as frameshift or non-frameshift. Synonymous SNPs and insertions within the stop codon that maintained stop codon position were considered as non-functional variations, all other variations within coding regions were considered to be functional variations that would likely impact gene expression or function.

In order to assess whether there are differences in the amount of variation found in R-genes in comparison with other genes, summary metrics (V^X) were calculated for each annotated gene, where X indicates the type of variation being analysed. Metrics were calculated for all variations within the gene (V^A) and for each of the following functional or positional groups of variations: functional (V^F); non-functional (V^{NF}); non-functional within coding regions (V^{NFC}); non-functional within non-coding regions (V^{NFNC}); exon donor splice site (V^{EDS}); exon acceptor splice site (V^{EAS}); intron donor splice site (V^{IDS}); intron acceptor splice site (V^{IAS}); intron between coding exons (V^{CI}); 5′ UTR exon (V^{5UE}); 5′ UTR intron (V^{5UI}); 3′ UTR exon (V^{3UE}); 3′ UTR intron (V^{3UI}). For each group of variations, the corresponding summary metric was calculated for each gene, according to Equation (1), where t is the number of annotated transcript isoforms, l is the number of sequenced lines, $v_{i,j}$ is the number of variations identified within transcript i in line j, and n_i is a normalisation factor for the number of possible mutation sites for the transcript isoform. In the case of exon–intron splice site variation metrics, n_i is defined as the number of introns between coding exons in transcript i. For all other variation groups, n_i is defined as the number of bases in transcript i that could contain a variation that would be counted by that metric (e.g., n_i is the number of bases in transcript i that are within the 5′ UTR exons for V^{5UE} and would be the total number of coding bases in transcript i for V^F).

$$V^X = \frac{\sum_{i=1}^{t} \sum_{j=1}^{l} \frac{v_{i,j}}{n_i}}{t.l} \quad (1)$$

Two-sample t-tests, carried out at the 0.01% significance level, were used to determine whether the means of each variation summary metric significantly differed between non-R-genes and the combined set of R-genes and R-pseudogenes, and between R-genes and R-pseudogenes. The F-test was used to test whether variation was equal in the two groups being tested. In the case of unequal variation, Welch's t-test was applied.

2.8. Genomic Locations of SSR Primers

The forward and reverse primer sequences for 19,311 rice simple sequence repeat (SSR) microsatellite markers were downloaded from Gramene (Available online: www.gramene.org (accessed on 10 January 2019)) and were aligned against the six reference genome sequences using BLAST (Available online: blast.ncbi.nlm.nih.gov (accessed on 10 January 2019)).

3. Results

3.1. Identification of R-Genes, R-Pseudogenes, and Partial R-Genes

The combination of two in silico identification tools detected more candidate R-genes in all five *indica* genomes than in Nipponbare (Table 2 and Supplementary File S3). The total number detected in *indicas* ranged from 278 in Shuhui498 to 394 in Minghui 63, compared with 232 in Nipponbare. Although there was a significant amount of overlap between the sets of genes identified, many were not identified as R-genes in some genomes by either of the motif-searching methods. Across the six reference genomes, 94 to 191 genes met both of the criteria for R-candidate genes. Of those considered R-pseudogenes in this study, 25–55 were identified as complete by NLR-parser but the other motif searches did not detect both types of domain, and 113 to 176 were identified as having both types of domain but not identified as complete NLR genes by NLR-parser. Locations of the candidate R-genes and R-pseudogenes are given in Supplementary File S2, as well as the names of any overlapping SSR markers.

Table 2. Total number of candidate R-genes and R-pseudogenes identified by using both NLR-Parser and HMMER searches of the annotated protein sequences and 6-frame-translated gene DNA sequences in five *indica* reference genomes (Shuhui498, Minguhi 63, Zhenshan 97, IR8, and 93-11 [2013 assembly]) and Nipponbare (*japoinica*).

Candidate Type	Motif Search				Gene Counts					
	Search Sequence Type	NLR-Parser	HMMER NB-ARC	HMMER LRR	Shuhui498	Minguhi 63	Zhenshan 97	IR8	93-11	Nipponbare
R-genes	Protein	Complete	+	+	126	191	179	156	140	94
		Complete	+	-	35	46	50	35	55	25
		Partial	+	+	115	155	140	152	166	106
		-	+	+	2	2	4	4	10	7
		R-genes totals			***278***	***394***	***373***	***347***	***371***	***232***
R-pseudogenes	6-frame translated gene DNA	Complete	+	+	21	12	10	30	13	46
		Complete	+	-	10	8	15	13	8	25
		Complete	-	+	0	0	0	2	0	0
		Complete	-	-	0	0	0	0	0	1
		Partial	+	+	23	12	12	15	10	36
		-	+	+	2	1	0	0	1	2
		R-pseudogenes totals			***56***	***33***	***37***	***60***	***32***	***110***

Of the R-pseudogenes, between 10 and 46 in each reference genome were identified as complete by NLR-parser and as having both NBS domains and LRR domains by the HMMER motif searches, and 19–64 met only one of the two motif searching selection criteria. The Nipponbare genome was found to have the highest number of R-pseudogenes, with 110, as compared with between 32 and 60 for the *indica* genomes tested.

Partial R-genes contain NBS or LRR domain but do not meet the criteria for selection as candidate genes between were identified in each reference genome. There were between 988 and 1189 partial R-genes per genome (Supplementary File S3) that are not considered here as R-gene candidates.

The two bioinformatic motif searching approaches were compared for R-gene identification with each other and with the BLAST searching. The NLR-parser searches of protein sequences from each of the six reference genome assemblies detected sequences corresponding to between 119 and 237 genes considered to be complete NLR genes, and 330–475 genes encoding partial NLR sequences. The search of 6-frame-translated gene DNA sequences detected a higher number of genes containing both complete and par-

tial NLR sequences in all reference genomes, with 205–279 complete and 560–653 partial NLR genes identified. HMMER motif searches of the protein sequences of each reference genome identified between 398 and 629 NBS domain-encoding genes, and 753–1,058 LRR domain-encoding genes. Of these, 207–348 genes were identified that encoded both NBS and LRR domains. In addition, across all reference genomes, a total of 12 genes were identified that encoded TIR domains, but only five of these also encoded an NBS domain and none encoded LRR domains. Unlike the NLR-parser pipeline, HMMER searches of the 6-frame-translated gene DNA sequences identified fewer genes containing motifs typical of NLR genes in each reference genome; although, a higher number of genes with NBS domains were detected in all references except for 93-11. In each of the reference genomes, the number of NBS or LRR domain-containing genes was between 513 and 623 for NBS domains, and between 827 and 926 for LRR domains, with between 132 and 175 of those genes containing both types of domains.

BLAST searches of the Shuhui498 R-genes and R-pseudogenes against the alternative rice reference annotations and NCBI non-redundant sequence databases identified significant hits against sequences annotated as either NLR genes or resistance genes for all 278 candidate R-genes and 51 of 56 candidate R-pseudogenes.

3.2. R-Gene Distribution

R-genes were identified on all 12 rice chromosomes of all the reference genomes. However, their distribution in the genomes is uneven and they tend to be found clustered together on chromosomes (Figure 1). The number of R-genes on each chromosome also varies greatly, ranging from 136 R-genes and R-pseudogenes on chromosome 11 of Minghui 63, down to 12 on chromosome 3 of Shuhui498 (Figure 2). Chromosome 11 had the highest number of R-genes for all reference genomes with between 18.3% (93-11) and 31.9% (Minghui 63) of the total complement of identified R-genes and R-pseudogenes.

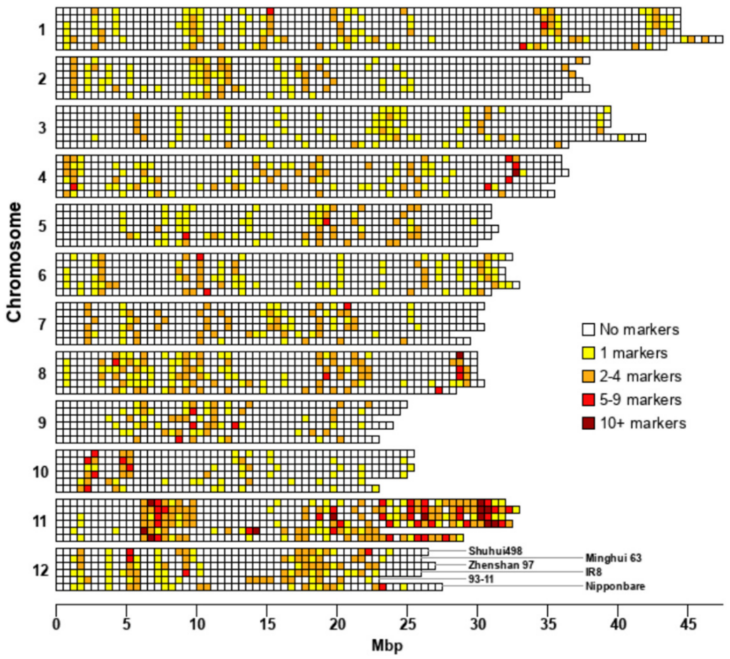

Figure 1. Genomic distribution of identified R-genes and R-pseudogenes. Chromosome lengths differ for each reference genome and adjacent blocks do not represent genomic alignments.

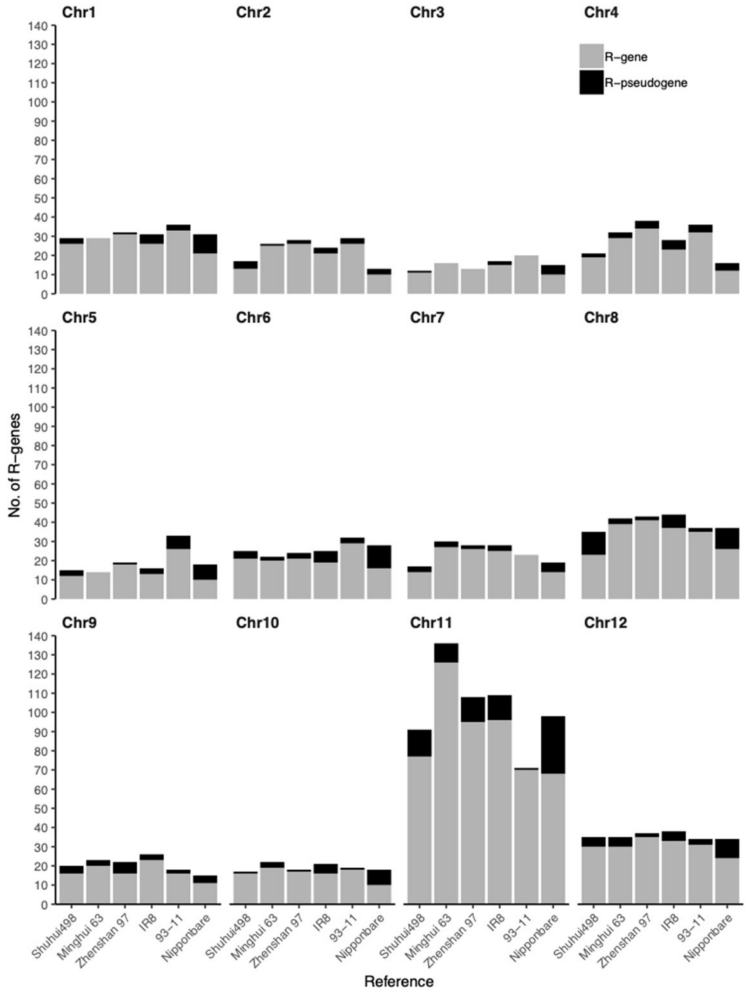

Figure 2. Number of R-genes and R-pseudogenes identified on each chromosome of the six reference genomes analysed.

3.3. R-Gene Orthologs

In this study, allelic or duplicated genes were termed orthologs (due to being detected using the JustOrthologs software), and they were classified as partial R-genes if they had been identified as partial NLR genes by the NLR-parser pipeline, or if HMMER motif searches had identified NBS or LRR domains in the gene, but they had failed to meet the criteria for selection as candidate R-genes or R-pseudogenes. Genes that were not identified at all by these methods were classed as non-R-genes. 76.7% of orthologs to candidate R-genes were also classified as R-genes, with 8.1% classified as R-pseudogenes, 10.3% as partial R-genes and 2.9% as non-R-genes. Only 21.5% of R-pseudogene orthologs were also classified as R-pseudogenes, with 57.3% being classified as R-genes, 10.3% as partial R-genes, and 3.9% as non-R-genes (Table 3).

We identified 654 groups of orthologous genes that contained at least one R-gene or R-pseudogene, of which 612 contained at least one *indica* ortholog (Supplementary File S4). A total of 112 of these groups had a single *indica* member with no orthologs identified in the other *indica* references analysed, 99 of which also lacked a *japonica* ortholog in the Nipponbare genome. The number of R-genes or R-pseudogenes without any identified orthologs ranged from 14 in 93-11 to 30 in Minghui 63 for the *indica* genomes analysed, while 42 were identified in the Nipponbare genome. Orthologs were identified in all 5 *indica* genomes for 309 of the orthologous gene groups, with 268 of these also having a Nipponbare ortholog identified.

Table 3. Number of orthologs of R-genes and R-pseudogenes identified by JustOrthologs search of CDS sequences and reciprocal BLAST of gene DNA sequences.

				Orthologs Identified					
Genome	**Predicted Genes**			**Shuhui498**	**Minghui 63**	**Zhenshan 97**	**IR8**	**93-11**	**Nipponbare**
Shuhui498	R-genes	278	R-genes	-	199	183	173	181	135
			R-pseudogenes	-	9	8	16	14	41
			Partial R-genes	-	9	13	10	28	17
			Non-R-genes	-	7	5	2	3	9
	R-pseudogenes	56	R-genes	-	26	26	16	24	12
			R-pseudogenes	-	10	7	14	5	20
			Partial R-genes	-	3	1	2	3	3
			Non-R-genes	-	2	1	3	2	3
Minghui 63	R-genes	394	R-genes	198	-	257	248	244	125
			R-pseudogenes	24	-	13	25	13	57
			Partial R-genes	58	-	14	15	36	30
			Non-R-genes	9	-	7	1	4	19
	R-pseudogenes	33	R-genes	9	-	16	11	14	7
			R-pseudogenes	10	-	10	8	1	7
			Partial R-genes	3	-	0	1	2	4
			Non-R-genes	2	-	0	3	1	4
Zhenshan 97	R-genes	373	R-genes	184	253	-	243	249	144
			R-pseudogenes	22	14	-	19	11	53
			Partial R-genes	66	18	-	15	39	25
			Non-R-genes	8	6	-	2	5	18
	R-pseudogenes	37	R-genes	8	13	-	11	12	9
			R-pseudogenes	6	10	-	10	4	3
			Partial R-genes	4	3	-	4	3	3
			Non-R-genes	2	2	-	3	0	1
IR8	R-genes	347	R-genes	170	246	238	-	233	130
			R-pseudogenes	15	11	10	-	11	46
			Partial R-genes	51	20	22	-	36	33
			Non-R-genes	17	15	13	-	6	20
	R-pseudogenes	60	R-genes	17	25	20	-	25	14
			R-pseudogenes	13	8	10	-	2	13
			Partial R-genes	10	6	5	-	8	4
			Non-R-genes	1	2	1	-	0	0
93-11	R-genes	371	R-genes	186	247	255	240	-	148
			R-pseudogenes	20	15	13	25	-	58
			Partial R-genes	61	16	19	12	-	28
			Non-R-genes	8	4	3	1	-	13
	R-pseudogenes	32	R-genes	11	13	12	11	-	10
			R-pseudogenes	5	1	4	2	-	7
			Partial R-genes	3	5	4	1	-	3
			Non-R-genes	0	0	1	0	-	1
Nipponbare	R-genes	232	R-genes	136	124	142	128	144	-
			R-pseudogenes	11	8	8	14	10	-
			Partial R-genes	20	18	15	10	25	-
			Non-R-genes	4	5	5	0	1	-
	R-pseudogenes	110	R-genes	45	60	60	49	60	-
			R-pseudogenes	20	8	4	14	6	-
			Partial R-genes	8	4	5	3	8	-
			Non-R-genes	4	1	1	2	1	-

3.4. Cloned R-Genes

BLASTP alignments for the 38 known blast resistance and 12 BB resistance protein sequences (from 31 R-gene loci) against annotated amino acid sequences from all 6 reference genomes are given in Supplementary File S5, along with details of identified orthologs of those genes.

Best hits for the blast resistance protein sequences had an identity of at least 82.7%. All sequences had been identified as R-genes in this study, with the exception of those corresponding to *Pi54* and *Pik-2*, which were identified as partial R-genes, and *Pi21* and *Pid-2*, which could not be identified by the motif searching methods applied here. Best hits for the BB resistance sequences had a minimum identity of 84.2%, with the exception of the best hit for *Xa10*, which had an identity of 52.9%. The only BB resistance locus with a best hit that was identified as an R-gene was *Xa1*. Best hits for *Xa3/Xa26*, *Xa5*, and *Xa23* were classified as partial R-genes, and all remaining BB resistance loci hits as non-R-genes with the in silico analyses.

Eleven protein sequences, from four blast resistance and four BB resistance loci, had joint-best hits in two or more of the six reference genomes. Minghui 63 had the highest number of best hits for both blast resistance and BB resistance loci. Shuhui498 had the highest number of orthologs to cloned blast resistance genes, only lacking orthologs to genes at the *Pi37* and *Pi54* loci. Minghui 63 had the highest number of BB resistance orthologs, with *Xa10* being the only locus without an identified ortholog. Orthologs were identified in all 6 reference genomes for 8 blast resistance and 3 BB resistance loci.

Alignments for protein sequences and their identified orthologs, corresponding to 31 R-gene loci, are given in Supplementary File S6 and summarised in Supplementary File S7. With one exception, all ortholog sequences aligned to the cloned protein sequence to some degree; although, the amount of sequence overlap and percentage identity varied widely across the different loci and, in some cases, between the different orthologs of the same cloned gene. The exception to this was the Minghui 63 ortholog of the best BLAST hit to the *Xa21* protein sequence, from the Zhenshan 97 genome. Although the Minghui 63 ortholog protein sequence aligned very well to the regions of the Zhenshan 97 and IR8 orthologs, the alignment did not overlap with the cloned *Xa21* sequence, nor the 93-11 ortholog.

3.5. Variation in Predicted R-Genes

The number of variations identified within the R-genes and R-pseudogenes of the five *indica* reference genomes ranged from 33,429 in 93-11 to 64,438 in IR8. Alternative variations were identified at a number of these sites, giving between 34,498 and 65,894 unique variations in each *indica* genome. In addition, we identified 35,303 unique homozygous variations at 33,927 sites in the *japonica* Nipponbare genome.

Variation summary metrics, calculated for all Shuhui498 genes (Section 2.7), revealed more variation in R-genes and R-pseudogenes than in non-R-genes (mean V^A of 3.01×10^{-3} for R-genes, 2.19×10^{-3} for non-R-genes, p-value of 6.89×10^{-8}). Significant increases were seen for metrics relating to both functional (mean V^F of 1.81×10^{-3} for R-genes, 1.17×10^{-3} for non-R-genes, p-value of 8.90×10^{-9}) and non-functional variations (mean V^{NF} of 2.09×10^{-3} for R-genes, 1.56×10^{-3} for non-R-genes, p-value of 1.14×10^{-6}) (Figure 3). The amount of non-functional variations was significantly higher in R-genes in both coding and non-coding gene regions (mean V^{NFC} of 1.22×10^{-3} for R-genes, 6.57×10^{-4} for non-R-genes, p-value of 5.28×10^{-6}; mean V^{NFNC} of 3.00×10^{-3} for R-genes, 2.42×10^{-3} for non-R-genes, p-value of 2.20×10^{-3}).

Analysis of the metrics relating to variations within specific gene features showed increased amounts of variation in the 5′ UTR introns of R-genes (mean V^{5UI} of 2.94×10^{-3} for R-genes, 2.21×10^{-3} for non-R-genes, p-value of 7.60×10^{-3}), but no significant differences were observed between amounts of variation in any of the other gene features examined. Comparison of R-genes against R-pseudogenes showed no significant differences for any of the metrics.

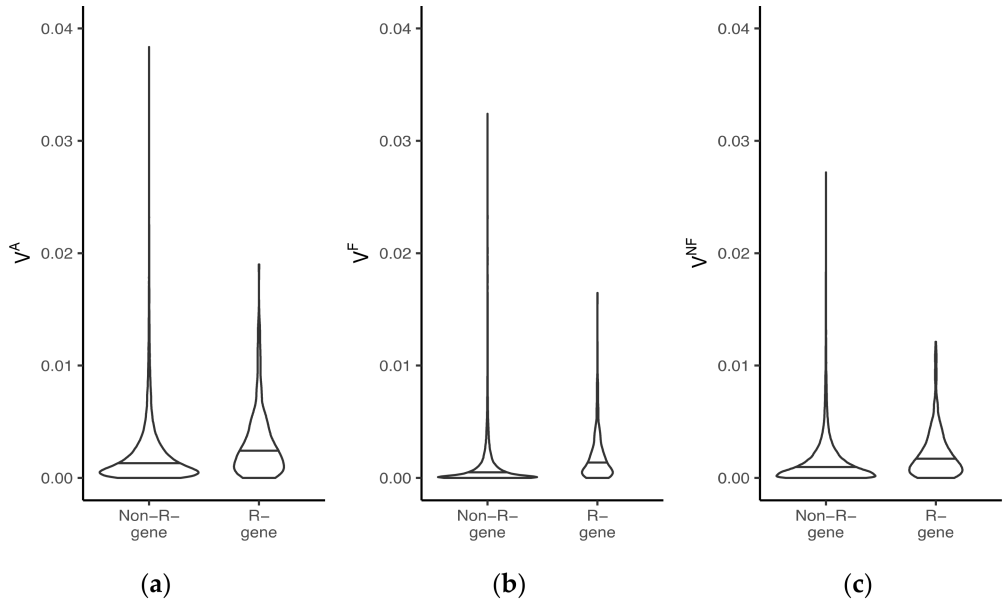

Figure 3. Differences in abundance of (**a**) all variations (V^A), (**b**) functional variations (V^F), and (**c**) non-functional variations (V^{NF}) between R-genes (including R-pseudogenes) and non-R-genes of the Shuhui498 genome in the 75 *indica* rice lines examined. The shapes of the violins represent the distribution of the corresponding summary metric values calculated for the genes. Median values are indicated by the horizontal lines.

The location and alleles of SNP and InDel variants within 334 candidate R-genes and R-pseudogenes (each in a separate worksheet) are presented for 75 rice lines (Supplementary File S8).

Read alignment rates for the 75 resequenced *indica* sequencing reads were at least 84% for all reference genomes, with the exception of the Riz Indetermine line, which had alignment rates of 66–70%. Mean read alignment rates were over 90% for all *indica* reference genomes and 88% for the Nipponbare *japonica* reference. On average, approximately 1 million variations were identified in each of the resequenced lines against each of the *indica* reference genomes, but nearly twice as many were identified against Nipponbare.

3.6. Proximity of SSR Markers to R-Genes

Based on the primer sequence alignments, unique positions could be determined in at least one of the reference genomes for 18,434 (95%) of the SSR (microsatellite) markers, with 17,140 (88%) being aligned to at least one of the five *indica* genomes (Supplementary File S9). Locations of the flanking SSRs for each identified candidate R-gene and R-pseudogene are given in Supplementary File S2. The median distance of the nearest SSR to each R-gene for each of the five *indica* reference genomes ranged from 16.6 kb for Shuhui498 to 23.6 kb for Minghui 63, with these 2 genomes also having the lowest and highest values for maximum distance from the nearest SSR (162.0 kb and 355.2 kb). Between 8.0% (93-11) and 13.3% (Shuhui498) of R-genes and R-pseudogenes identified in the *indica* genomes have an SSR marker located within 1 kb. The median and maximum distances were lower for the *japonica* Nipponbare genome, at 12.1 kb and 149.7 kb, respectively, with 12.9% of Nipponbare R-genes and R-pseudogenes being within 1 kb of an SSR (Figure 4).

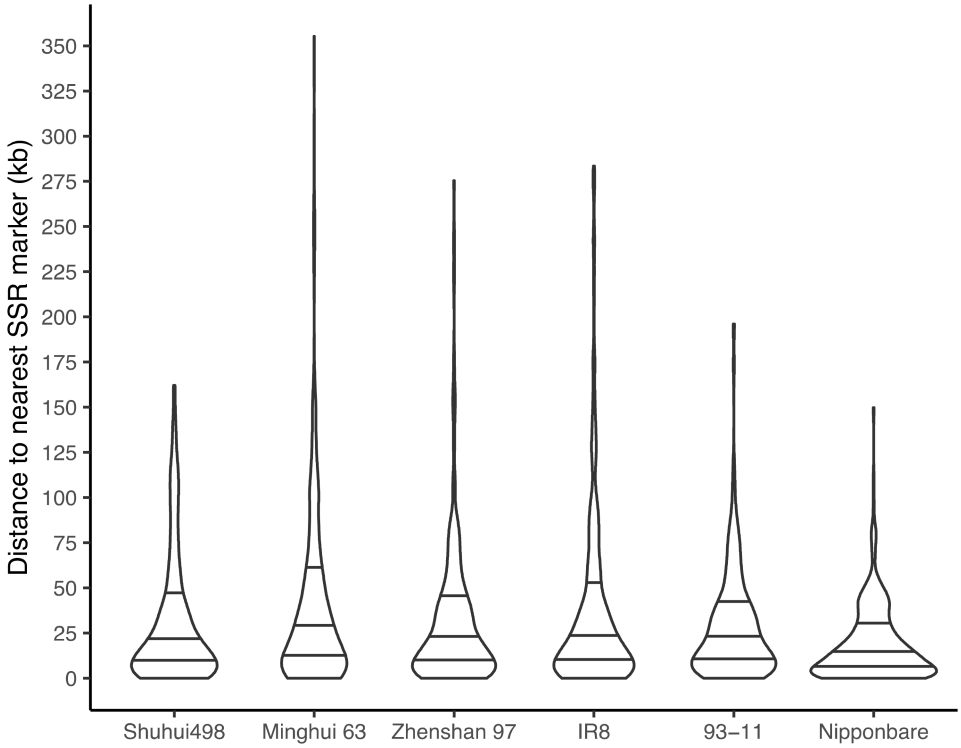

Figure 4. Distance of candidate R-genes and R-pseudogenes from the nearest SSR marker (listed in Gramene, Available online: www.gramene.org (accessed on 10 January 2019) in the six reference genomes analysed. The shape of the violins illustrates the distribution of distances each reference genome, horizontal lines indicate the first quartile, median, and last quartile.

4. Discussion

The predicted R-genes (Figures 1 and 2) were located in similar clusters in each of the six reference genomes and this was consistent with previous studies [8,39–41]. Clustering may be critical for the generation of novel resistance through unequal recombination or gene conversion [39] and could permit the plant to partially negate the fitness costs of reduced growth and reproduction that are associated with expression of defence genes through the coordinated regulation of R-gene clusters [42–44]. Co-expression of clustered genes has also been observed for PTI-associated genes [45]. Variation between homologous cluster members could drive the evolution of new resistances against prevalent races of the pathogen. It is likely that domestication has enabled their duplication in cultivated rice because more NLR family genes were found in Asian cultivated rice than their wild ancestors [15]. The functions of most clustered-R genes are still unknown and only a relatively small number have so far been proven to confer disease resistance. Paralogs may have no function, or a function that could be identified by studying variation between different cultivated varieties.

4.1. Evaluation of In Silico Analysis Methods

One major limiting factor in any in silico analysis of genomic properties is the quality and completeness of the reference genome itself, and another is the number of annotated genes per varietal genome. The genome assembly for the *indica* rice cultivar Shuhui498 is more continuous than the widely used reference genomes of both *japonica* rice Nipponbare (MSU7) and *Arabidopsis thaliana* (TAIR10); with only 5 centromere gaps and up to 14 potential duplication gaps it is estimated to cover more than 99% of the Shuhui498 genome. The assembly has been annotated with 38,714 high-confidence protein-coding genes [26]. Assemblies of the Zhenshan 97 and Minghui 63 genomes are estimated to be 99.99% accurate and over 90% complete, with annotations for 54,831 and 57,174 genes, respectively [27]. The

IR8 reference genome has 35,508 annotated genes and, based on alignment of CEGMA and BUSCO benchmarking gene sets, is estimated to be over 93% complete [28]. The difference in the number of complete R-genes identified in each genome (Table 2) may be influenced by the number of annotated genes in that variety's genome. Although the 93-11 reference genome is much more fragmented and contains many assembly errors [46], its inclusion in the analyses presented here is warranted, due to it being the original *indica* rice reference assembly and to enable comparison of our results with previous research. A recent study using Nipponbare, Shuhui498, Minghui 63, and Tetep identified differences in NLR genes among the four genomes that they consider to be consistent with the high level of diversity found in NLRs [47].

The motif-searching tools used for R-gene prediction rely on annotated gene or protein models and scenario differences in annotation can have a strong impact on the outcome of the prediction. However, they have been shown to be highly effective for R-gene discovery in other plant species such as *A. thaliana*, *Brachypodium distachyon* [30], and *Solanum tuberosum* [23]. BLAST searches with all the candidate gene and protein sequences detected in this study were carried out and a high proportion of the R-gene candidates had hits to resistance genes that have been shown to be involved in pathogen resistance, giving us reasonable confidence in their candidacy.

The use of cloned R-genes to compare with predicted R-genes (Supplementary Files S5–S7) gives confidence in the predicted R-genes being real R-genes; however, the confidence is lower for the predicted orthologs. The best hits identified for genes at all blast resistance loci were classified as R-genes by our searches, with the exception of those at the Pi54 and Pik-2 loci—which were classified as partial R-genes—and the best hits for *Pi21* and *Pid-2*—which were not identified in our searches. Rice *pi21* is a recessive resistance gene that results from a loss-of-function mutation in the wild-type *Pi21* blast susceptibility gene, which encodes a proline-rich protein that has been shown to slow the plant's defence response [48]. *Pid-2* belongs to the receptor-like kinase (RLK) group of R-genes [49]. As such, we would not expect these two genes to be identified in our searches. *Pi54* and *Pik-2* have both been characterised as NLR genes [50,51]. Our searches identified an LRR domain in the best hit for the former, and an NBS domain for the latter, but neither met our criteria for selection as a candidate R-gene or R-pseudogene. Further investigation revealed that the HMMER searches did identify an alignment for the LRR_4 domain sequence in the best hit for *Pik-2*, but that the alignment e-value was below the inclusion threshold. The cloned sequences for *Pi54* utilised here were both partial sequences, and a HMMER search for the NB-ARC domain failed to identify any alignment to either sequence. Most of the BB R-genes encode unique proteins that are not found in other plants. An exception to this, and the only cloned BB R-gene that is a member of the NLR gene family, is *Xa1* [52]. This is consistent with our results, with *Xa1* being the only BB gene for which we identified the best hit as being an R-gene.

In addition to annotated protein sequences, utilisation of the 6-frame-translations for motif searching strategies enabled identification of R-pseudogenes. Although these do not code for complete, functional resistance proteins, they retain the genetic potential to do so in other rice varieties, where variations within the equivalent gene may result in alternative transcript isoforms being expressed. Similarly, many of the functional R-genes in the genomes analysed here will be non-functional in other rice lines. This is borne out by the results of our ortholog identification results, with many R-gene orthologs being classified as R-pseudogenes, partial R-genes, or even non-R-genes. The maintenance of high levels of R-pseudogenes in plant genomes is a recognised phenomenon, thought to result from the fitness benefits associated with having a reservoir of genetic diversity that protects against future pathogen pressure [53]. It is the sum total of the R-genes and R-pseudogenes, and the specific variations within those genes that defines the ability of a population to detect polymorphic pathogen effectors [42].

Using a comparative genomic approach, combining HMMER and BLAST searches, Luo et al. [41] identified 223 full-length R-genes in the 93-11 genome and 276 in Nipponbare,

along with 345 93-11 genes and 347 Nipponbare genes that they classed as R-pseudogenes. Their criteria for classification of genes as R-pseudogenes differed to ours, however, as they considered partial R-genes to be R-pseudogenes, with the R-pseudogenes being further divided into two groups as follows: 146 (93-11) and 138 (Nipponbare), with nonsense point mutations or small frameshift InDels; 199 (93-11) and 209 (Nipponbare), with large deletions (i.e., partial genes). While the former group is likely to overlap with our definition of R-pseudogenes, the latter is not. In our study, all identified R-pseudogenes are considered to have the genetic potential to code for functional R-genes. However, we recognise that partial R-genes, identified by Luo et al. [41], could still have a role to play in the evolution of novel resistance through sequence exchange events.

Of the 223 93-11 R-genes identified by Luo et al. [41], 212 were present in the gene model for this assembly. Of these, we identified 167 as R-genes and 6 as R-pseudogenes, with a further 35 being identified as having NBS or LRR domains but not meeting our criteria for selection as candidate R-genes or R-pseudogenes. Only 4 of the 93-11 genes they identified as R-genes did not appear in any of our search results. Conversely, our study identified 127 R-genes in 93-11 that they classified as R-pseudogenes, 39 that they described as uncharacterised due to incomplete genome assembly coverage, and 39 that were not found in their results.

Comparison of the Nipponbare R-genes identified by Luo et al. [41] with our results is more complicated due to their use of the MSU RGAP gene model, in contrast to our use of the IRGSP model. Of the 276 R-genes they identified, only 159 could be mapped (on the basis of the gene IDs) to the IRGSP gene model used here; of these, 116 were identified as R-genes and 18 were identified as R-pseudogenes. A further 18 genes were identified in our search strategies but not selected as candidate R-genes or R-pseudogenes, and 7 genes did not appear in any of our search results. A total of 103 complete R-genes were identified in our study that Lu et al. classified as R-pseudogenes, and a further 10 were identified in our study that were absent from their results. Very few of the R-pseudogenes described by Luo et al. were completely absent from our search results, yet we identified 991 partial R-genes in 93-11 and 713 in Nipponbare that do not appear in their results. It is worth noting that Stein et al. [28] identified 535 NLR genes in the indica 93-11 genome.

Monosi et al. [17] and Zhou et al. [36] reported the identification of higher numbers of R-genes in the Nipponbare genome. However, gene lists are not available for comparison with our results; furthermore, Monosi et al. include genes lacking an LRR domain.

An unavoidable issue with any in silico, reference-based approach is that there are likely to be a number of *indica* rice R-genes that are simply not present in the reference genomes. High frequencies of presence/absence (P/A) polymorphism in R-genes have been demonstrated between *indica*, *aus*, and *japonica* rice subpopulations [41,54]; although, the occurrence of P/A polymorphisms would be expected to be lower between less diverged *indica* lines. The difference in the numbers of R-genes identified in each of the reference genomes analysed in this study indicates that P/A polymorphisms are common. Approximately half of the 612 *indica* R-gene or pseudogene ortholog groups identified in this study had member genes identified in all *indica* references analysed. That leaves a large number that appear to lack orthologs in at least one of the other genomes tested, with 3.5–7.0% of the R-genes and R-pseudogenes identified in each *indica* genome not having orthologs in any of the other *indica* genomes analysed here. Those R-genes, for which orthologs were not detected in one of more of the other reference genomes analysed, are likely candidates for P/A mutations. Further investigation into individual cases is required to be certain that missing orthologs are not simply due to missing gene annotations, or a failure of the ortholog detection methods applied here to identify true orthologs. It might be expected that differences in the genome assembly and annotation pipelines would introduce biases into the number of orthologs detected between the various references. This could be assessed by either repeating the orthogroup analysis on the genomic sequence instead of gene models, or through a BLAST approach to identify orthologues that are missing in the gene models of individual rice genome annotations.

Identification of paralogous genes within each genome was not carried out in this study. However, combination of the pairwise analyses revealed that a number of the identified groups of orthologous genes contained multiple genes from the same genome. These are likely to be paralogs; although, there are almost certainly additional paralogous relationships involving R-genes and R-pseudogenes that are not captured by these results.

4.2. R-Gene Variation between Rice Varieties

Our results showed significantly higher levels of variation in R-genes compared with non-R-genes in the 75 *indica* rice lines (Figure 3). Variations predicted to cause functional changes in gene products were more prevalent in R-genes, as were synonymous variations. Increased variability in R-genes has been observed in a number of plant species, with particularly high numbers of major-effect polymorphisms [55,56].

The increased variability we observed was not uniform across all portions of the genes. No significant differences in variation levels were observed in intron and exon donor and acceptor splice sites, nor in the introns between coding exons or 3′ UTR exons. A study of the evolution of the *RGC2* family in lettuce, one of the largest clusters of R-genes characterised in plants, revealed homogenisation of intron sequences amongst the gene family members occurring concurrently with diversification of exon sequences. Such evolutionary trends are hypothesised to result from the balance of diversifying selection acting to increase variation, while frequent sequence exchanges amongst the gene family members, combined with genetic drift and a lack of selective restraint, act to homogenise intronic gene portions. The resulting high levels of intron similarity may further facilitate sequence exchange [40].

Many of the R-genes and R-pseudogenes were in close proximity to microsatellite (SSR) markers (Figure 4); therefore, this could facilitate an understanding of their contribution to resistance in mapping populations and their introgression into new genetic backgrounds through marker-assisted selection (MAS). The nearest SSR markers to each of the 2323 genes we identified has been made available for this purpose (Supplementary Files S2 and S9), although nearly 90% of the R-genes and R-pseudogenes do not have a flanking SSR marker within 1 kb. However, the majority R-genes and R-pseudogenes contain SNP and InDel variations (Supplementary Files S8) and many of these could be targeted using appropriate marker technologies such as KASP or chip-based assays, if considered of interest in specific polymorphic lines. Further genotyping (alongside phenotyping) experiments to identify resistance in breeding material, followed by gene expression analysis, is recommended to associate variations with resistance properties and confirm the functionality of candidate R-genes.

5. Conclusions

The increasing availability of whole rice genomes and bioinformatics tools provides better resources for breeders to understand genome organisation of NLR genes and their associated variations. This study produced a catalogue of predicted NLR R-gene candidates, identified variation among 75 *indica* lines, and provided potential selectable SSR, SNPs, and InDel loci that could be useful for further understanding of resistance inheritance and selection across a range of potential *indica* parents. These resourses can inform the design of more precise resistance breeding strategies for rice.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/agronomy11112297/s1, Supplementary File S1: Details of *indica* rice lines used in sequence variation analyses, Supplementary File S2: List of candidate R-gene and R-pseudogene IDs, genomic positions, and closest SSR markers (one sheet per reference genome), Supplementary File S3: Counts of genes identified by the various combinations of methods utilised for R-gene identification, Supplementary File S4: R-gene and R-pseudogene orthologous gene groups identified by JustOrthologs and reciprocal BLAST, Supplementary File S5: Details of best BLAST alignment hits for cloned R-gene sequences and orthologs corresponding to those genes, Supplementary File S6: Multiple sequence alignments of cloned R-genes and identified orthologs, Supplementary File S7: Summary of Clustal

Omega multiple sequence alignments of cloned R-genes and identified orthologs, Supplementary File S8: Read alignment and variant call statistics for 75 *indica* lines (one sheet per gene), Supplementary File S9: Positions of rice SSR markers based on BLASTN alignment of primer pairs.

Author Contributions: M.J.Q.-T. and K.A.S. conceived and designed the study; M.J.Q.-T. developed the theory and carried out the analyses; M.J.Q.-T. wrote the manuscript, with contributions from K.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovate UK (Grant number 103711).

Data Availability Statement: The data presented in this study (all Supplementary Files) are openly available in Dataverse at doi:10.7910/DVN/DENPYE.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Savary, S.; Willocquet, L.; Pethybridge, S.J.; Esker, P.; McRobers, N.; Nelson, A. The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* **2019**, *3*, 430–439. [CrossRef]
2. Nelson, R.; Wiesner-Hanks, T.; Wisser, R.; Balint-Kurti, P. Navigating complexity to breed disease-resistant crops. *Nat. Rev. Genet.* **2018**, *19*, 21–33. [CrossRef]
3. Yoshimura, S.; Yoshimura, A.; Iwata, N.; McCouch, S.R.; Lleva Abenes, M.; Baraoidan, M.R.; Mew, T.W.; Nelson, R.J. Tagging and combining bacterial blight resistance genes in rice using RAPD and RFLP markers. *Mol. Breed.* **1995**, *1*, 375–387. [CrossRef]
4. Balachiranjeevi, C.H.; Bhaskar Naik, S.; Abhilash Kumar, V.; Harika, G.; Mahadev Swamy, H.K.; Hajira, S.; Dilip Kumar, T.; Anila, M.; Kale, R.R.; Yugender, A.; et al. Marker-assisted pyramiding of two major, broad-spectrum bacterial blight resistance genes, Xa21 and Xa33 into an elite maintainer line of rice, DRR17B. *PLoS ONE* **2018**, *13*, e0201271.
5. Ballini, E.; Berruyer, R.; Morel, J.B.; Lebrun, M.H.; Nottéghem, J.L.; Tharreau, D. Modern elite rice varieties of the 'Green Revolution'have retained a large introgression from wild rice around the Pi33 rice blast resistance locus. *New Phytol.* **2007**, *175*, 340–350. [CrossRef]
6. Gottin, C.; Dievart, A.; Summo, M.; Droc, G.; Périn, C.; Ranwez, V.; Chantret, N. A New 'Comprehensive'Annotation of Leucine-Rich Repeat-Containing Receptors in Rice. *Plant J.* **2021**, *108*, 492–508. [CrossRef]
7. McHale, L.; Tan, X.; Koehl, P.; Michelmore, R.W. Plant NLR proteins: Adaptable guards. *Genome Biol.* **2006**, *7*, 212. [CrossRef]
8. Bai, J.; Pennill, L.A.; Ning, J.; Lee, S.W.; Ramalingam, J.; Webb, C.A.; Zhao, B.; Sun, Q.; Nelson, J.C.; Leach, J.E.; et al. Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. *Genome Res.* **2002**, *12*, 1871–1884. [CrossRef]
9. Aarts, M.G.M.; Hekkert, B.L.; Holub, E.B.; Beynon, J.L.; Stiekema, W.J.; Pereira, A. Identification of r-gene homologous DNA fragments linked to disease resistance loci in Arabidopsis thaliana. *Mol. Plant Microbe Interact.* **1998**, *11*, 251–258. [CrossRef]
10. Vossen, J.H.; Dezhsetan, S.; Esselink, D.; Arens, M.; Sanz, M.J.; Verweij, W.; Verzaux, E.; van der Linden, C.G. Novel applications of motif-directed profiling to identify disease resistance genes in plants. *Plant Methods* **2013**, *9*, 37. [CrossRef]
11. Sallaud, C.; Lorieux, M.; Roumen, E.; Tharreau, D.; Berruyer, R.; Svestasrani, P.; Garsmeur, O.; Ghesquiere, A.; Notteghem, J.L. Identification of five new blast resistance genes in the highly blast-resistant rice variety IR64 using a QTL mapping strategy. *Theor. Appl. Genet.* **2003**, *106*, 794–803. [CrossRef]
12. Wan, H.; Yuan, W.; Ye, Q.; Wang, R.; Ruan, M.; Li, Z.; Zhou, G.; Yao, Z.; Zhao, J.; Liu, S.; et al. Analysis of TIR- and non-TIR-NLR disease resistance gene analagous in pepper: Characterization, genetic variation, functional divergence and expression patterns. *BMC Genom.* **2012**, *13*, 502. [CrossRef]
13. Sharma, R.; Rawat, V.; Suresh, C.G. Genome-wide identification and tissue-specific expression analysis of nucleotide binding site-leucine rich repeat gene family in Cicer arietinum (kabuli chickpea). *Genom. Data* **2017**, *14*, 24–31. [CrossRef]
14. de Araújo, A.C.; Fonseca, F.C.D.A.; Cotta, M.G.; Alves, G.S.C.; Miller, R.N.G. Plant NLR receptor proteins and their potential in the development of durable genetic resistance to biotic stresses. *Biotechnol. Res. Innov.* **2019**, *3*, 80–94. [CrossRef]
15. Mizuno, H.; Katagiri, S.; Kanamori, H.; Mukai, Y.; Sasaki, T.; Matsumoto, T.; Wu, J. Evolutionary dynamics and impacts of chromosome regions carrying R-gene clusters in rice. *Sci. Rep.* **2020**, *10*, 872. [CrossRef] [PubMed]
16. Shang, J.; Tao, Y.; Chen, X.; Zou, Y.; Lei, C.; Wang, J.; Li, X.; Zhao, X.; Zhang, M.; Lu, Z.; et al. Identification of a new rice blast resistance gene, Pid3, by genomewide comparison of paired nucleotide-binding site-leucine-rich repeat genes and their pseudogene alleles between the two sequenced rice genomes. *Genetics* **2009**, *182*, 1303–1311. [CrossRef]
17. Monosi, B.; Wisser, R.J.; Pennill, L.; Hulbert, S.H. Full-genome analysis of resistance gene homologues in rice. *Theor. Appl. Genet.* **2004**, *109*, 1434–1447. [CrossRef]
18. Zhou, T.; Wang, Y.; Chen, J.-Q.; Araki, Z.; Jing, K.; Jiang, J.; Shen, J.; Tian, D. Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NLR genes. *Mol. Genet. Genom.* **2004**, *271*, 402–415. [CrossRef] [PubMed]
19. Yu, H.; Shahid, M.Q.; Li, R.; Li, W.; Liu, W.; Ghouri, F.; Liu, X. Genome-wide analysis of genetic variations and the detection of rich variants of NLR encoding genes in common wild rice lines. *Plant Mol. Biol. Report.* **2018**, *36*, 618–630. [CrossRef]

20. Yu, J.; Hu, S.; Wang, J.; Wong, G.K.; Li, S.; Liu, B.; Deng, Y.; Dai, L.; Zhou, Y.; Zhang, X.; et al. A draft sequence of the rice genome (Oryza sativa L. ssp. Indica). *Science* **2002**, *296*, 79–92. [CrossRef] [PubMed]
21. Steele, K.A.; Quinton-Tulloch, M.J.; Amgai, R.B.; Dhakal, R.; Khatiwada, S.P.; Vyas, D.; Heine, M.; Witcombe, J.R. Accelerating public sector rice breeding with high-density KASP markers derived from whole genome sequencing of indica rice. *Mol. Breed.* **2018**, *38*, 38. [CrossRef]
22. The 3,000 rice genomes project. The 3,000 rice genomes project. *GigaScience* **2014**, *3*, 7. [CrossRef] [PubMed]
23. Jupe, F.; Pritchard, L.; Etherington, G.J.; MacKenzie, K.; Cock, P.J.A.; Wright, F.; Sharma, S.K.; Bolser, D.; Bryan, G.J.; Jones, J.D.G.; et al. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genom.* **2012**, *13*, 75. [CrossRef]
24. Bailey, T.; Gribskov, M. Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* **1998**, *14*, 48–54. [CrossRef]
25. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. Meme suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef] [PubMed]
26. Du, H.; Yu, Y.; Ma, Y.; Gao, Q.; Cao, Y.; Chen, Z.; Ma, B.; Qi, M.; Li, Y.; Zhao, X.; et al. Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* **2017**, *8*, 15324. [CrossRef]
27. Zhang, J.; Chen, L.; Xing, F.; Kudrna, D.A.; Yao, W.; Copetti, D.; Mu, T.; Li, W.; Song, J.; Xie, W.; et al. Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E5163–E5171. [CrossRef]
28. Stein, J.C.; Yu, Y.; Copetti, D.; Zwickl, D.J.; Zhang, L.; Zhang, C.; Chougule, K.; Gao, D.; Iwata, A.; Goicoechea, J.L.; et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat. Genet.* **2018**, *50*, 285–296. [CrossRef] [PubMed]
29. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **2000**, *436*, 793–800.
30. Steuernagel, B.; Jupe, F.; Witek, K.; Jones, J.D.G.; Wulff, B.B.H. NLR-parser: Rapid annotation of plant NLR complements. *Bioinformatics* **2015**, *31*, 1665–1667. [CrossRef]
31. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]
32. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
33. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. Blast+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef] [PubMed]
34. 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.; et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* **2013**, *6*, 4. [CrossRef] [PubMed]
35. Sakai, H.; Lee, S.S.; Tanaka, T.; Numa, H.; Kim, J.; Kawahara, Y.; Wakimoto, H.; Yang, C.C.; Iwamoto, M.; Abe, T.; et al. Rice annotation project database (RAP-DB): An integrative and interactive database for rice genomics. *Plant Cell Physiol.* **2013**, *54*, e6. [CrossRef]
36. Zhao, W.; Wang, J.; He, X.; Huang, X.; Jiao, Y.; Dai, M.; Wei, S.; Fu, J.; Chen, Y.; Ren, X.; et al. BGI-RIS: An integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res.* **2004**, *32*, D377–D382. [CrossRef]
37. Miller, J.B.; Pickett, B.D.; Ridge, P.G. JustOrthologs: A fast, accurate and user-friendly ortholog identification algorithm. *Bioinformatics* **2019**, *35*, 546–552. [CrossRef]
38. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scaleable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [CrossRef]
39. Meyers, B.C.; Kozik, A.; Griego, A.; Kuang, H.; Michelmore, R.W. Genome-wide analysis of NLR-encoding genes in Arabidopsis. *Plant Cell* **2003**, *15*, 809–834. [CrossRef]
40. Kuang, H.; Woo, S.S.; Meyers, B.C.; Nevo, E.; Michelmore, R.W. Multiple genetic processes result in heterogenous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell* **2004**, *16*, 2870–2894. [CrossRef]
41. Luo, S.; Zhang, Y.; Hu, Q.; Chen, J.; Li, K.; Lu, C.; Liu, H.; Wang, W.; Kuang, H. Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family. *Plant Physiol.* **2012**, *159*, 197–210. [CrossRef]
42. Zhang, Y.; Xia, R.; Kuang, H.; Meyers, B.C. The diversification of plant NLR defense genes directs the evolution of microRNAs that target them. *Mol. Biol. Evol.* **2016**, *33*, 2692–2705. [CrossRef]
43. Choi, K.; Reinhard, C.; Serra, H.; Ziolkowski, P.A.; Underwood, C.J.; Zhao, X.; Hardcastle, T.J.; Yelina, N.E.; Griffin, C.; Jackson, M.; et al. Recombination rate heterogeneity within Arabidopsis disease resistance genes. *PLoS Genet.* **2016**, *12*, e1006179. [CrossRef] [PubMed]
44. Karasov, T.L.; Chae, E.; Herman, J.J.; Bergelson, J. Mechanisms to mitigate the trade-off between growth and defense. *Plant Cell* **2017**, *29*, 666–680. [CrossRef] [PubMed]
45. Tonnessen, B.W.; Bossa-Castro, A.M.; Mauleon, R.; Alexandrov, N.; Leach, J.E. Shared cis-regulatory architecture identified across defense response genes is associated with broad-spectrum quantitative resistance in rice. *Sci. Rep.* **2019**, *9*, 1536. [CrossRef] [PubMed]

46. Pan, Y.; Deng, Y.; Lin, H.; Kudrna, D.A.; Wing, R.A.; Li, L.; Zhang, Q.; Luo, M. Comparative BAC-based physical mapping of Oryza sativa ssp. indica var. 93-11 and evaluation of the two rice reference sequence assemblies. *Plant J.* **2013**, *77*, 795–805. [CrossRef]
47. Wang, L.; Zhao, L.; Zhang, X.; Zhang, Q.; Jia, Y.; Wang, G.; Li, S.; Tian, D.; Li, W.H.; Yang, S. Large-scale identification and functional analysis of NLR genes in blast resistance in the Tetep rice genome sequence. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 18479–18487. [CrossRef]
48. Fukuoka, S.; Saka, N.; Koga, H.; Ono, K.; Shimizu, T.; Ebana, K.; Hayashi, N.; Takahashi, A.; Hirochika, H.; Okuno, K.; et al. Loss of function of a proline-containing protein confers durable disease resistance in rice. *Science* **2009**, *325*, 998–1001. [CrossRef]
49. Chen, X.; Shang, J.; Chen, D.; Lei, C.; Zou, Y.; Zhai, W.; Liu, G.; Xu, J.; Ling, Z.; Cao, G.; et al. A B-lectin receptor kinase gene conferring rice blast resistance. *Plant J.* **2006**, *46*, 794–804. [CrossRef]
50. Vasudevan, K.; Gruissem, W.; Bhullar, N.K. Identification of novel alleles of the rice blast resistance gene Pi54. *Sci. Rep.* **2015**, *5*, 15678. [CrossRef]
51. Zhai, C.; Lin, F.; Dong, Z.; He, X.; Yuan, B.; Zeng, X.; Wang, L.; Pan, Q. The isolation and characterization of Pik, a rice blast resistance gene which emerged after rice domestication. *New Phytol.* **2011**, *189*, 321–334. [CrossRef]
52. Kim, S.-K.; Reinke, R.F. A novel resistance gene for bacterial blight in rice, Xa43(t) identified by GWAS, confirmed by QTL mapping using a bi-parental population. *PLoS ONE* **2019**, *14*, e0211775. [CrossRef] [PubMed]
53. Marone, D.; Russo, M.A.; Laidò, G.; De Leonardis, A.M.; Mastrangelo, A.M. Plant nucelotide binding site- leucine-rich repeat (NLR) genes: Active guardians in host defense responses 2013. *Int. J. Mol. Sci.* **2013**, *14*, 7302–7326. [CrossRef]
54. Schatz, M.C.; Maron, L.G.; Stein, J.C.; Hernandez Wences, A.; Gurtowski, J.; Biggers, E.; Lee, H.; Kramer, M.; Antoniou, E.; Ghiban, E.; et al. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.* **2014**, *15*, 506. [PubMed]
55. Cao, J.; Schneeberger, K.; Ossowksi, S.; Günther, T.; Bender, S.; Fitz, J.; Koenig, D.; Lanz, C.; Stegle, O.; Lippert, C.; et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **2011**, *43*, 956–963. [CrossRef] [PubMed]
56. Zhou, P.; Silverstein, K.A.T.; Ramaraj, T.; Guhlin, J.; Denny, R.; Liu, J.; Farmer, A.D.; Steele, K.P.; Stupar, R.M.; Miller, J.R.; et al. Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. *BMC Genom.* **2017**, *18*, 261. [CrossRef] [PubMed]