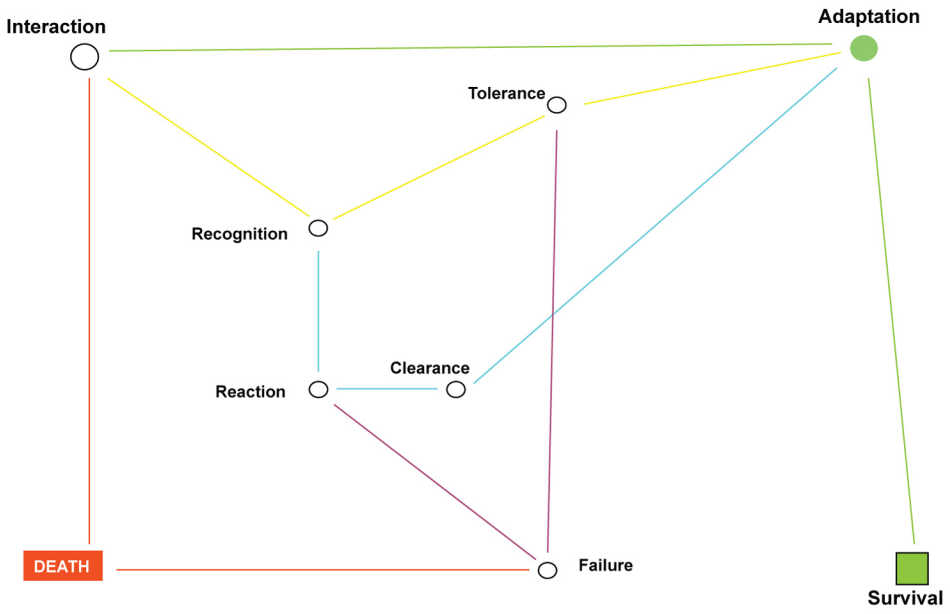
EDITED BY DAVIDE MALAGOLI



THE EVOLUTION OF THE IMMUNE SYSTEM

CONSERVATION AND DIVERSIFICATION

The Evolution of the Immune System

Conservation and Diversification

The Evolution of the Immune System

Conservation and Diversification

Davide Malagoli
Department of Life Sciences
Biology Building, University of Modena
and Reggio Emilia, Modena, Italy

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1800, San Diego, CA 92101-4495, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-801975-7

Typeset by Thomson Digital

Chapter 12

Genomic Instability and Shared Mechanisms for Gene Diversification in Two Distant Immune Gene Families: The Plant *NBS-LRR* Genes and the Echinoid *185/333* Genes

Matan Oren, Megan A. Barela Hudgell, Preethi Golconda, Cheng Man Lun, L. Courtney Smith

Department of Biological Sciences, The George Washington University, Washington DC, United States

1 INTRODUCTION

One of the major challenges faced by immune systems is to generate a protein repertoire that is broad and competent enough to recognize the ever-diversifying array of pathogenic nonself. Eukaryotes have numerous strategies to achieve this. Innate immune systems consist of large families of pattern recognition receptors (PRRs) that identify different pathogen associated molecular patterns (PAMPs) with high specificity. Examples include (1) Toll-like receptors (TLRs)[1] found in most animals from Porifera to humans, with PAMP recognition function demonstrated in some species, including human, mouse, and fruit fly[2]; (2) fibrinogen-related proteins (FREPs) with antiparasite activities in mollusks[3]; (3) Down syndrome cell adhesion molecule (Dscam) in insects[4,5] and crustaceans[6,7] with opsonin function; and (4) variable domain-containing chitin binding proteins (VCBPs) in protochordates[8,9] that respond to gut microbes. The adaptive immune system in jawed vertebrates uses somatic recombination of gene segments to create enormous diversity of T cell and B cell receptors.[10] Alternatively, the adaptive immune system in the jawless vertebrates relies on a copy-choice mechanism to assemble sections of leucine-rich repeat (LRR) cassettes into a germline gene to create similar diversity of variable lymphocyte receptors.[11–13] On the other hand, innate immune systems have been suggested

The Evolution of the Immune System. http://dx.doi.org/10.1016/B978-0-12-801975-7.00012-8

to lack the flexibility of adaptive immunity to identify and respond to novel PAMPs that have either newly appeared, or have been newly introduced into a population, due to either environmental changes or as a result of the arms race with host immunity. Therefore, it is necessary for innate immune systems in eukaryotes to employ other types of swift genomic diversification mechanisms either within the lifespan of the host or between generations, to stay even in the arms race with the pathogens. Here, we discuss different aspects of genome diversification in two very distinct innate immune gene-families: the nucleotide binding site leucine rich repeat (*NBS-LRR)* genes in plants, and the *185/333* genes in echinoids. The first is abundant in many species of plants, is a subset of the resistance (*R*) genes, and appeared early in the plant lineage more than 500 million years ago[14] whereas the second is restricted to the echinoid lineage of echinoderms, and the extant genes are estimated to be only 2.7–10 million years old.[15,16] Although there are many differences between these two gene families, there are some striking similarities in the genomic structure and the gene diversity among and within species, which will be the focus of this review.

2 THE *NBS-LRR* GENE FAMILY IN HIGHER PLANTS

The immune response in plants consists of two arms: PAMP-triggered immunity (PTI) and effector-triggered immunity (ETI).[17,18] PTI relies on cell-surface-membrane mounted PRRs that extend into the apoplast and recognize and respond to microbial molecules. ETI functions most often in the plant cell cytoplasm, either acting directly by detecting pathogen virulence-factors called effectors, or acting indirectly by monitoring host proteins that have been altered by effector activity.[17,19] The guard hypothesis suggests that the indirect detection of effector activity is facilitated through a cytoplasmic complex of an R protein that functions as a guard for a host guardee protein. In normal conditions the guard/guardee complex is stable, but upon injection of effectors into the plant cell by a pathogen, the effectors alter the guardee, which is detected by the R protein guard, and induces a signaling pathway to activate the ETI response.[17,20] The indirect ETI response to changes in the guardee proteins maximizes the capacity of the plant host to detect the activity of a large variety of pathogens with a much smaller number of R proteins.[18,21] The key players in the ETI response are a diverse group of mostly intracellular R proteins[17,22] that are encoded by a few to hundreds of *R* genes that are present typically in clusters in every plant genome (Fig. 12.1A), with an expanded repertoire in flowering plants.[14,22] Most R proteins, although not all, are characterized by the presence of a nucleotide binding site (NBS) domain, a linker region, and a variable number of LRRs (Fig.12.1B, C).[23] There are 151 NBS-LRR proteins in the mouse-ear or thale cress, *Arabidopsis thaliana*, 458 in rice, 459 in wine grape, but only two in the much more primitive plant, the spike moss (reviewed in Ref. [14]). The NBS-LRR type of R proteins are divided into two major structurally distinct sub-groups, defined by the N-terminal domain, which is either a

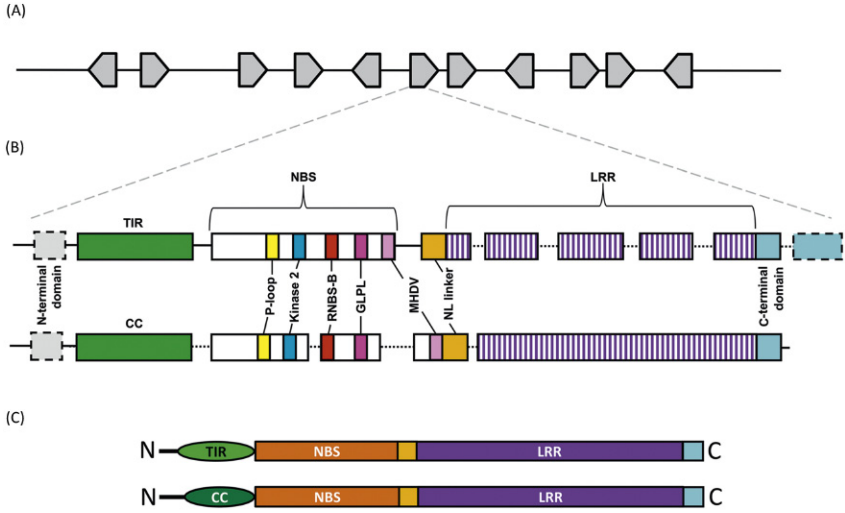


FIGURE 12.1 **An *R* gene cluster and the structures of the TNL and CNL genes and proteins.** (A) A representative homologous *R* gene cluster (not to scale). *R* genes are most often clustered within plant genomes, commonly in homologous clusters, with genes of similar structure and sequence. Each *gray polygon* represents an individual gene (introns and exons are not shown), and gene orientation is indicated by the pointed end of each polygon. Intergenic regions are represented by the *black line* and are not to scale. Gene clusters can vary in size and have different numbers of genes. The majority of genes range in size from 2 to 15 kb, with a maximum size of 44 kb.[25] (B) Representative structures of a Toll/interleukin-1 receptor domain (TIR)-NBS-LRR (TNL) gene and a coiled-coil domain (CC)-NBS-LRR (CNL) gene (not to scale). The structures of *R* genes are highly diverse, with an N-terminal domain (*light gray*, *dashed outline*) in some genes, a TIR domain in TNLs, or a CC domain in CNLs. The NBS domain has five key semi-conserved regions, including a P-loop, a Kinase 2 motif, and a Resistance Nucleotide Binding Site B (RNBS-B) motif,[26] plus two semi-conserved amino-acid motifs, GLPL and MHDV. Between the NBS domain and the LRRs is an NL linker (named for its location between the NBS and LRR regions). The LRRs can be encoded by either a single or by multiple exons, depending on the gene. The C-terminal domain is of variable lengths among genes, the first portion being encoded within the last LRR exon, and additional C-terminal regions can be encoded on multiple following exons *(blue, dashed outline)*. Dotted horizontal lines represent introns that are present in some genes and absent in others. (C) Representative structure of TNL and CNL proteins. The domains that are present in both types of R proteins include the NBS, the NL linker, and the LRRs. The N-terminus is either a TIR or a CC, which defines the TNL or CNL type of R protein, respectively. *(Source: Part B modified from Refs. [27,28].)*

Toll/interleukin-1 receptor (TIR) domain in the TIR-NBS-LRR (TNL) type, or a coiled-coil (CC) domain in the CC-NBS-LRR (CNL) type (Fig. 12.1B, C) (reviewed in[21]). Binding of the LRRs of the TNL and CNL proteins to effector molecules, or to altered guardee proteins, triggers different downstream signaling cascades that lead to the hypersensitivity response (HR) in plants. HR is a rapid apoptotic reaction in infected cells, and those nearby, which functions to remove the availability of cytoplasmic nutrients to pathogens, and thereby restricts their growth and spread.[24]

3 THE *185/333* GENE FAMILY IN ECHINOIDS

The *185/333* gene family encodes a diversified repertoire of immune-response proteins in sea urchins. To date, the *185/333* gene families have only been identified in two species of sea urchins, *Strongylocentrotus purpuratus* (the California purple sea urchin)[15,16] and *Heliocidaris erythrogramma* (the Australian purple sea urchin).[29] However, these genes are likely present in most echinoids, as they have been identified in the genome sequences of *Strongylocentrotus franciscanus* and *Allocentrotus fragilis*[16] and *Lytechinus pictus* (K. Buckley, University of Toronto, personal communication). Among those, the most studied is the *Sp185/333* gene family in the California purple sea urchin, *S. purpuratus*, which was first identified because it showed significant up-regulation in response to immune challenge with heat-killed bacteria and PAMPs, including lipopolysaccharides (LPS), peptidoglycans (PGN), and β-1,3-glucan.[30–33] The family consists of up to 60 members; however, the gene number may vary among individuals[16] and among different species (K. Buckley, personal communication). The *Sp185/333* genes range in size from 1.2 to 2 kb and have only two exons separated by a small intron (380–413 nucleotides).[32,34] The first exon (51–54 nucleotides) encodes the hydrophobic leader, whereas the second encodes the mature protein that shows significant sequence diversity. Optimal alignments of genes and transcripts require the insertion of artificial gaps, which define the presence and absence of short blocks of sequence, known as *elements* (Fig. 12.2A). The combinations of different elements result in recognizable mosaics of elements, called *element patterns.*[31,32,34] This gene structure is persistent among sea urchin species studied to date, although the elements in *185/333* genes from different sea urchin species are not the same.[15,29] The predicted structure of the Sp185/333 proteins is a signal peptide at the N-terminus, a glycine-rich region with an arginine–glycine–aspartic acid (RGD) motif (suggestive of integrin binding), a histidine-rich region, and a C-terminal region (Fig. 12.2B). No secondary structure can be predicted based on the amino-acid sequence for any of the proteins deduced from the cDNA or gene sequences.[31–33,35] The *185/333* genes are expressed in specific subpopulations of sea urchin coelomocytes, and the encoded proteins appear to be localized internally in perinuclear vesicles in some phagocytes, and on the cell surface of small phagocytes.[29,36–38] In *S. purpuratus*, single phagocytes from immune-challenged sea urchins express a single *Sp185/333* message, inferring complex regulation of gene expression from the family and the production of a single Sp185/333 protein per cell.[38] It should be noted that although a genome sequence exists for an individual California purple sea urchin, the *Sp185/333* gene family is artificially underrepresented within this genome, likely due to computational assembly-contraction problems resulting from the variety of repeat sequences that are present between and within the genes (Fig. 12.2A). The size and organization of the *Sp185/333* gene family is currently known, based on gene and message

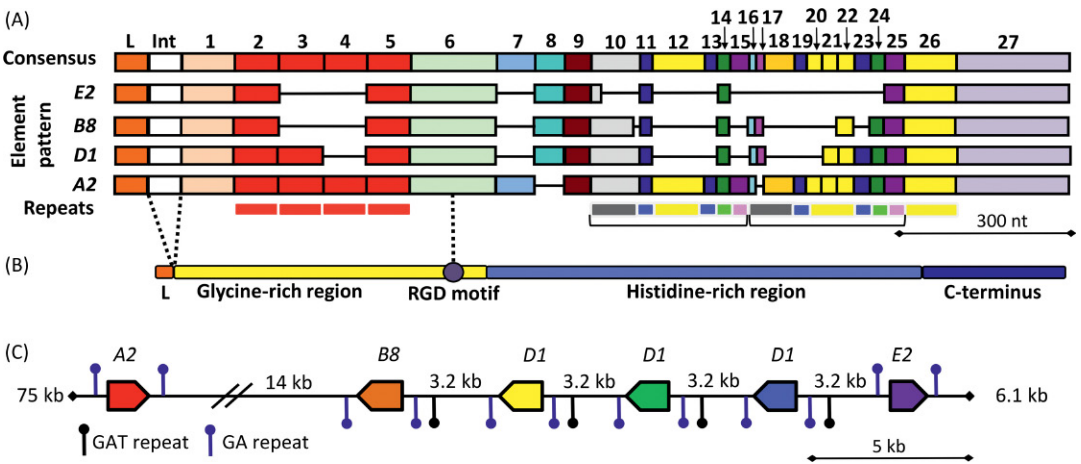


FIGURE 12.2 ***Sp185/333* gene cluster, repeat-based alignment, element patterns and protein structure.** (A) Repeat-based alignment of the *Sp185/333* genes shown in (C). The alignment optimizes correspondence between repeats and elements whenever possible.[34] Optimal alignments require artificial gaps *(horizontal black lines)* that delineate individual *elements* shown as *different-colored rectangles*. The consensus of all possible elements are numbered across the top of the alignment. Each gene is composed of two exons; the first encodes the leader (L) and the second encodes the mature protein. Almost all genes have a single intron (int) of ~400 nt (not to scale). The mosaic combinations of presence or absence of different elements in the second exon defines the element pattern (*E2, B8, D1,* and *A2*). Elements that correlate with each of the six types of repeats are shown in *different-colored rectangles* at the bottom (type 1, *red*; type 2, *blue*; type 3, *yellow*; type 4, *green*; type 5, *pink*; type 6, *dark gray*); the brackets under the type 2 to type 6 repeats indicate the two duplicated regions. (B) The deduced Sp185/333 protein structure. The protein size and regions of the protein are correlated with the gene structure in (A). (C) Six *Sp185/333* genes in a BAC insert (GenBank accession number BK007096) are closely linked. Genes are indicated by element pattern and color; *A2 (red)*, *B8 (orange)*, three *D1 (yellow, green, blue)*, and *E2 (purple)*. The genes are located near the 3' end of the BAC insert within 34 kB. Gene orientations are indicated and spacing is relative to the scale. GA microsatellites flank each gene and GAT microsatellites flank segmental duplications within which are positioned three *D1* genes. *(Sources: Part A modified from [34]; part B modified from [39].)*

sequences, and on the assembled insert for one BAC clone (GenBank accession number BK007096), which contains six tightly clustered genes (Fig. 12.2C).[39]

4 GENE DIVERSIFICATION

There is ample evidence for rapid evolution in the *NBS-LRR* gene family in higher plants[14,40,41] and in the *185/333* gene family in sea urchins.[29,42] One has only to evaluate the variability of both the gene numbers among and within species, and the sequence diversity of the genes, to obtain a general understanding of the pace of diversification. The *NBS-LRR* gene family is one of the largest and most variable gene families in plants.[14] Although the common ancestor for the plant *NBS-LRR* genes is predicted to be much older than the common ancestor of the *Sp185/333* genes, the *NBS-LRR* family has continued to expand and diversify.[14] Many of its members exhibit allelic polymorphism,[18] and for some *NBS-LRR* loci, polymorphism within populations is as great as that characterized for the major histocompatibility complex in vertebrates.[40] The *NBS-LRR* genes show two general types of models for gene evolution: the majority are type I genes that show diversifying selection with a rapid rate of evolution and high sequence exchange among genes, and the rest are type II genes that show a slower rate of diversification correlating with less frequent exchanges.[20,22,27,40,43] These two models of gene evolution are not mutually exclusive, and *NBS-LRR* genes positioned within the same cluster can show signatures of both diversification rates.[27] It is noteworthy that the TNL class tends to show significantly higher evolution rates than the non-TNL genes, including the CNL class.[40] Within the TNL genes, sequences that encode the solvent-exposed regions of the LRRs (Fig. 12.1B) seem to be under the highest positive selection and show the highest levels of genetic diversification.[21] These regions show elevated ratios of the nonsynonymous versus synonymous substitutions (dN/dS). This is likely driven by the shared sequences among the LRRs, together with selection based on the function of the LRRs in pathogen-associated recognition.[43] In contrast, the region encoding the NBS domain undergoes purifying selection and is highly conserved, which is likely based on its functions in nucleotide binding, which is crucial for R protein function to initiate signaling in order to activate the protective HR.[21,23,44]

Similar to the *NBS-LRR* gene family, the *185/333* genes show exceptional diversity both among animals and among sea urchin species.[15,16,29,31,34,45] An unrooted phylogenetic tree of *185/333* sequences from *H. erythrogramma* and *S. purpuratus* shows a complete separation of sequences from the two species into different clades.[29] The recognizable element patterns in the second exon of the *Sp185/333* genes are composed of a mosaic of 25–27 different possible elements (depending on the alignment) that range in size from 12 to 258 nucleotides (Fig. 12.2A) and generate 51 different patterns that have been identified to date.[31,33,34] Similarly, the *He185/333* genes have 26 elements and 31 element patterns, based on the first report on this gene family.[29] The element

patterns of the different *Sp185/333* genes impart high sequence diversity, but paradoxically, because they share element sequences, they are up to 88% identical.[16,34] Furthermore, although element sequences are shared among genes, identical sequences of full-length genes are not shared among individual sea urchins. This is because (1) only subsets of elements are shared among genes and among animals, (2) there is sequence diversity within different versions of the same element, and (3) there are sequence variations among intron from different genes. The *185/333* genes from both species show many nonsynonymous substitutions with respect to synonymous substitutions (dN/dS ratio) for some element sequences, indicating diversifying selection for these regions, whereas for other elements, a low dN/dS ratio, suggesting purifying selection, has been noted.[29,32,34] Furthermore, when *Sp185/333* gene sequences are compared, the level of diversity among the elements shows significant differences.[34] In general, the *185/333* and the *NBS-LRR* gene families portray sequence diversity patterns with exceptionally fast diversification rates and high dN/dS ratios for some regions within the genes, and slow diversification rates and low dN/dS ratios for other regions. For both families, a conserved basic structure of the genes that encode the functional regions of the proteins is maintained.

5 CLUSTERING AND TANDEM REPEATS

The *NBS-LRR* genes are unevenly distributed in the genome, and tend to be present in clusters that vary in size from 2 to 23 genes, with possibly more in single clusters (Fig. 12.1A).[14,28,43,46–51] For example, the rice Xa21 gene cluster has seven paralogs within 230 kb,[52] the tomato I2 cluster has seven paralogs within 90 kb,[53] and the RPW8 cluster in *Arabidopsis* has five paralogs within 13 kb.[54] *NBS-LRR* clusters can be homogeneous, with all members showing similar structure of either TNL or CNL genes (Fig. 12.1B, C), or can be heterogeneous with TNL and CNL genes mixed together.[55] Homogeneous *NBS-LRR* clusters that contain tandemly repeated genes are very common in many plant genomes. For example, ~40 homogeneous clusters are present in the *Arabidopsis* genome, compared to ~10 clusters that are heterogeneous.[49] There is evidence that the clustering of *NBS-LRR* genes is a major factor in the sequence diversification among the members of the family. The cluster size and gene copy number is positively correlated with sequence-exchange frequency among members of the cluster.[40,41,51] Furthermore, there are greater dN/dS ratios for paralogs in clusters compared to isolated paralogs.[51]

The published *Sp185/333* cluster consists of six closely linked *Sp185/333* genes within 34 kb. Five of the genes are tightly clustered within 20 kb and are 3.2 kb apart, whereas a peripheral sixth gene is located at a distance of 14 kb (Fig. 12.2C).[39] The peripheral genes are oriented in the same direction, whereas the four internal genes are oriented in the opposite direction. The cluster is composed of a mixture of homogeneous and heterogeneous genes based on the element patterns of the second exon (Fig. 12.2A). The three central genes all

have a *D1* element pattern, and are positioned within three tandem segmental duplications of ~4.5 kb that show 99.7% sequence identity and are flanked by GAT microsatellites (Fig. 12.2C).[39] The near- identity among the *D1* genes and their flanking regions suggest very recent duplication events.[16,39]

Both the *R* and *185/333* gene families contain several types of repeats. The *NBS-LRR* genes contain exons that encode LRRs of 20–29 amino acids with a consensus sequence of LxxLxLxxNxL(T/S)GxIPxxLGxLxx, in which "L" is Leu, Ile, Val, or Phe, T/S is Thre or Ser, and "x" is any amino acid.[56–58] The number of LRRs can vary among *NBS-LRR* genes, ranging from 4 to 50 repeats.[27] For example, in *Arabidopsis,* the number of LRRs ranges from 8 to 25[28] and the *Resistance Gene Candidate 2* (*RGC2*) genes in lettuce have 40 to 48 LRRs.[27] Although the LRRs have an established function for interaction with PAMPs or pathogen elicitors (reviewed in[17,59]), they also serve as an important component in creating genomic instability due to their repetitive nature, which leads to gene-family diversification. Evidence for the participation of LRRs in gene diversification processes lies within the differences in the LRRs among quickly diversifying type I *R* genes, compared to more slowly diversifying type II genes. The sequence identity of introns within type I genes vary between the 5' region and the 3' region of the gene (Fig. 12.1B). Introns within the LRR region have high sequence-identity when compared to each other, which may reflect higher rates of sequence exchanges within the LRR region. Introns within slowly evolving type II genes have low sequence identity, reflecting their lower rates of sequence-exchange events. TNLs have additional introns within the LRR coding regions that are absent from most known CNLs (Fig. 12.1B), which may be indicative of differences in the evolutionary history of the two gene types.[27,28] The greater number of introns within TNL genes versus CNL genes may indicate that TNL genes originated from a fusion of independent genes and are younger than CNL genes, which have few to no introns.[48] It is noteworthy that, although the CNL genes have lost their modular gene structure over time, the encoded proteins may maintain modular functions.

The repeats within the second exon of the *Sp185/333* genes allow two different alignments that are equally optimal.[34] The initial alignment is based on the cDNA sequences, and did not take into account the positions of the internal repeats.[32,33] The second alignment is repeat-based that optimized the correspondence of elements and repeats.[34] There are six types of imperfect repeats in the second exon that are both tandem and interspersed (Fig. 12.2A).[31,32,34,39] Depending on the gene, there are two to four type-1 repeats at the 5' end of the exon, plus multiple copies of type 2–6 repeats that are present in two duplications of the interspersed repeats, in addition to an extra type-3 repeat (Fig. 12.2A). In addition, there are GA microsatellites positioned on either side of each gene within the intergenic regions, and are located about 430 bp from the 5' end of each gene, and 300–700 bp from the 3' end (Fig. 12.2C).[39] The GAT microsatellites are positioned at the edges of three ~4.5 kb tandem segmental duplications that include three *D1* genes (Fig. 12.2C). Based on their

positions at the edges of the duplicated regions, they may act as mediators of the duplication process.[39] Moreover, pairwise sequence comparisons among the clustered genes identified in the BAC insert show that the sequences between the ends of the coding regions and the nearby flanking GA microsatellite are much more conserved than the regions outside of the GA repeats.[39] This suggests that the microsatellites surrounding the *Sp185/333* genes and those surrounding the segmental duplications may promote diversification of the family through regional instability, including sequence duplication and limiting sequence homogenization from gene conversion.[16] Taken together, both the *185/333* and *R* gene families are characterized by clustering, repeats, and duplications. These features are found abundantly within the genomic structure for each of these innate immune gene-families, and are likely crucial for the processes that lead to gene diversification.

6 SPECULATIONS ON DIVERSIFICATION MECHANISMS OF THE *Sp185/333* GENES

The regions of the genome in which the *NBS-LRR* and the *Sp185/333* gene families are located, are very likely prone to genomic instability, which leads to gene sequence diversification. Gene diversification is initiated by mechanisms that regulate changes in the gene-copy number and organization of the whole family in which entire genes are duplicated, transferred to another location, deleted, or incur changes within the gene sequences (Fig. 12.3). The arms race between host and pathogen drives changes in host immune gene-sequence, which in turn drives functional adaptations in genes encoding effector proteins in pathogens, as demonstrated for the regions of the plant *R* genes that encode the LRRs. The variety of repetitive sequences in the *NBS-LRR* and *Sp185/333* gene families promote genomic instability and nucleotide mismatches that may take place when homologous chromosomes interact either during meiosis or DNA repair processes. Meiotic recombination and homologous DNA repair may be regarded as special events in which homologous chromosomes interact and promote sequence exchange. Several mechanisms that directly and indirectly lead to gene sequence rearrangements have been suggested for the *NBS-LRR* gene family. Based on the structural and diversification similarities of these two immune gene families, we speculate that these mechanisms apply to the *Sp185/333* genes as well. *NBS-LRR* genes are diversified by recombination between alleles and similar family members that result in new *R* genes with altered sequences. This spontaneous allele recombination is combined with selective pressures to detect PAMPs or elicitors, and results in gene variants with altered binding specificity. For example, individual L genes in flax that are derived from intragenic crossing-over show distinct phenotypes with regard to pathogen recognition.[60] Recombination in the *Sp185/333* genes has been detected computationally and is evident, not only between, but within elements and within the intron,[42] suggesting that recombination events can occur at

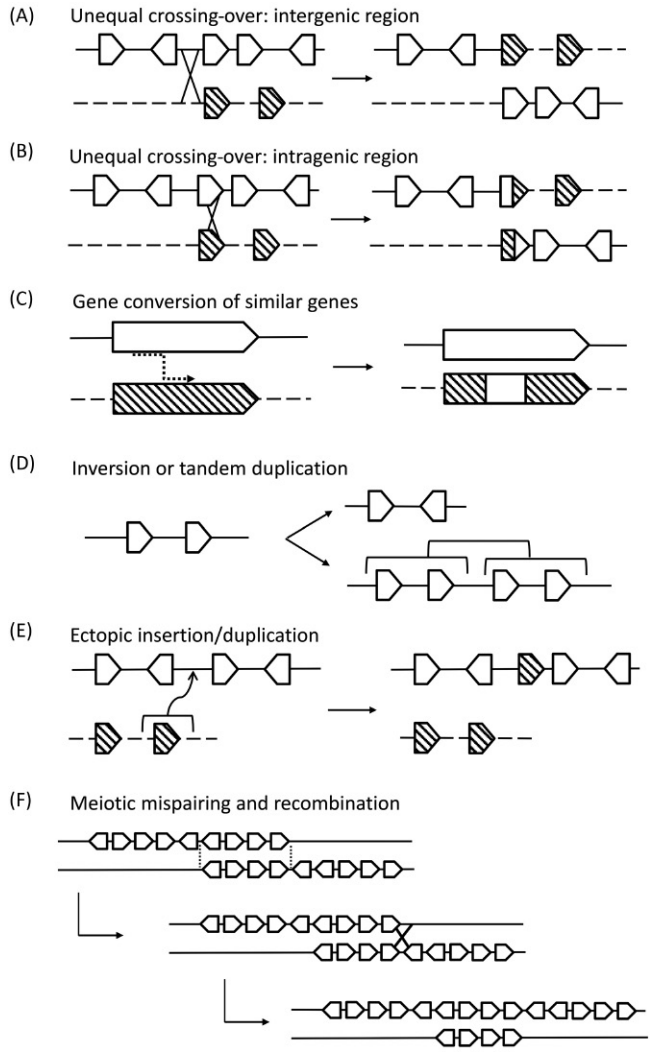


FIGURE 12.3 Genomic modifications that potentially lead to changes in the size of gene families, changes in the organization of clusters, and alterations to gene sequences. Genes are represented as *polygons* (*white* and *striped* genes are in nonallelic clusters), with the pointed end indicating gene orientation. The genomic DNA in which the genes are located is shown as a *solid* or *dashed horizontal line* representing nonallelic regions. The generation of diversity within clusters and sequence diversity within genes is illustrated. (A) An unequal crossing-over in an intergenic region between nonallelic clusters can alter the sizes of the clusters, and result in heterogeneous clusters. (B) Unequal crossing-over within genes in nonallelic clusters can generate recombinant genes, alter cluster sizes, and result in heterogeneous clusters. (C) Gene conversion results when the sequences of one gene are copied into a nonallelic gene of similar sequence. (D) Inversion changes the orientation of a gene within a cluster, whereas the tandem duplication of genes or sets of genes increases the size of a cluster. (E) A duplicated gene can be inserted into an ectopic location, generating a heterogeneous cluster. (F) Meiotic mispairing occurs when chromatids misalign in regions of allelic clusters of highly similar genes, with the outcome of more genes in one cluster and fewer in the allelic cluster. The recombination event is shown between genes, but can occur within genes, as in (B). The processes shown in (A), (B), (E), and (F) can increase and decrease gene-copy numbers in clusters and in gene families.

any point throughout the entire gene sequence and are not focused in hotspots. For example, there is no correlation between the patterns and numbers of the tandem type I repeats in the 5' end of the second exon and the patterns of the interspersed repeats located towards the 3' end of the exon (Fig. 12.1A). It has been suggested that highly similar sequences between duplicated genes within homologous clusters drive further diversification through processes such as unequal crossing-over (Fig. 12.3A, B),[61] resulting in unequal numbers and mispaired linked genes in the progeny, followed by processes that drive further diversification.[55] In both families, shared sequences among paralogs leads to a swift rate of recombination among the genes.

Gene conversion occurs either during meiosis or as a result of DNA repair processes when homologous sites show mismatches in base pairing. These mismatches are recognized and corrected by the DNA repair machinery to convert the sequence of one allele to the sequence of its homologous counterpart (Fig. 12.3C). Gene conversion is an important diversification mechanism in TNL genes that undergo rapid sequence diversification followed by pathogen-driven selection for function (reviewed in[22]). The *RGC2* type I genes in lettuce undergo rapid rates of gene conversion and recombination within the 3' end that encodes the LRRs, which have resulted in a large variety of *RGC2* genes.[27] Bioinformatic analysis of the *Arabadopsis* genome shows that gene-conversion events are driven by genes in clusters with sequence similarity.[61] The *Arabadopsis* gene-conversion events take place most commonly between genes that share 60–70% sequence identity, with most conversion events spanning 60–528 bp.[27,61,62] A greater tendency for gene conversion occurs when genes are proximal to each other, and is rarely found in genes dispersed farther away in the genome. It is not known whether gene conversion is a key mechanism for diversification in the *185/333* gene family. The structural components necessary for promoting gene conversion exist in the family, particularly given the significant sequence identity that is shared among the *Sp185/333* genes, which is based on shared element sequences.[34,39] Within a cluster, the presence of the microsatellites may initiate gene conversion, and then may limit the size of converted regions to block homogenization of the entire cluster.[39] Although it would be expected that higher sequence similarity would be present among tightly linked *Sp185/333* genes based on the likelihood of conversion occurring among proximal genes, comparisons among 121 genes of unknown linkage relationships from three *S. purpuratus* genomes show the same level of sequence similarity as genes of known linkage.[39] This lack of significant differences in the sequence diversity among clustered *Sp185/333* genes and 121 unique unlinked genes suggests that gene conversion may occur within the family among both local and more distant genes,[16] and that it occurs relatively swiftly within the family.

Both gene conversion and unequal crossing-over can drive gene duplication.[20] The most frequent duplication of whole *NBS-LRR* genes are tandem duplications, resulting in two similar genes in close proximity, which leads to the formation of a homogeneous gene cluster (Fig. 12.3D).[14,49,61] Gene duplication

and ectopic insertion of either a small set of genes or single genes to a distant location on the same or on a different chromosome (Fig. 12.3E), may also contribute to the family sequence diversity, which includes the formation of heterogeneous clusters.[28] Chromosomal segmental duplication can affect large portions of plant genomes, and is involved in the expansion of *NBS-LRR* gene families.[28] Small segmental duplications in the *Sp185/333* gene cluster appears to be the source of the *D1* gene duplication.[39] Furthermore, duplications of the tandem type 1 repeats in the *Sp185/333* family (Fig. 12.2A) may have been derived from ancestral sequences through duplications of the repeats, in addition to recombination and deletions, based on a computational estimation of the evolutionary history of this region of the genes.[42] Finally, similar to the *NBS-LRR* gene families, meiotic mispairing (Fig. 12.3F), based on the close proximity of the *Sp185/333* genes within the cluster, in addition to the sequence similarities among the genes, has been speculated to drive changes in the size of the *Sp185/333* gene family.[16,39]

Transposable elements may also contribute to genomic instability, which may drive diversity in both single genes and gene clusters. It has been shown that some *NBS-LRR* genes are associated with transposable elements. For example, the rice Xa21 gene family contains a large number of transposable elements, including LTR-retrotransposons and miniature inverted repeat transposable elements (MITEs).[63] Fragments of transposable elements are also present within the *Sp185/333* gene cluster. A portion of a Gypsy 10 long terminal repeat (LTR) S element is positioned near the 3' end of the *A2* gene in association with the flanking GA microsatellite.[39] In addition, three tandem, incomplete Tc1-N1-SP DNA transposon fragments are positioned at the 5' end of the *E2* gene in association with the GA microsatellite. It is not known whether transposable elements contribute to the diversification of the *NBS-LRR* and the *Sp185/333* gene families. However, we speculate that the transposable elements may contribute to the instability of the genomic regions harboring the gene families, through unequal crossing-over promoted by the duplication of transposable elements in the vicinity of members of the families.

Gene fragments and pseudogenes are commonly found in tightly linked clusters of paralogous genes, including 25% of the sea urchin *SpTLR* genes.[64] It is thought that this is a result of duplication and recombination among similar genes that also promotes sequence diversification. The levels of *NBS-LRR* pseudogenes vary from one species of plant to another, but are generally abundant.[65] In *Arabadopisis*, 8.05% of the *NBS-LRR* genes are pseudogenes,[28] whereas 51.3% of the *NBS-LRR* gene family in two rice subspecies are pseudogenes.[65,66] Contrary to the *NBS-LRR* family and the sea urchin *SpTLR* family, only one pseudogene of 171 sequenced genes has been identified in the *Sp185/333* gene family.[34] The pseudogene had no intron, and had a deletion in part of the coding region in the second exon that introduces a frame shift. Curiously for a gene family with significant levels of shared sequence within and surrounding the genes, no gene fragments have been found in the genome. The unexpectedly

low level of pseudogenes may be the result of rapid gene conversion (see previous sections) that may correct pseudogenes using sequences from nearby (and perhaps distant) genes, or alternatively by an unknown diversification regulation mechanism.

7 CONCLUSIONS

The *185/333* and the *NBS-LRR* gene families share several structural features, including inter- and intra-genic sequence repeats, duplicated genes, clustering, gene conversion, and diversifying selection in response to pathogens. These features are well established in the *NBS-LRR* gene family as components that are necessary for the initiation of a variety of diversification mechanisms. We find that the use of a comparative approach, even between echinoderms and higher plants, can be useful in understanding the biology of immune gene families, or for establishing hypotheses for how innate immune systems diversify and how potentially common mechanisms may function similarly in distantly related eukaryotes.

ACKNOWLEDGMENT

Funding to support the writing of this review was awarded by the United States National Science Foundation (IOS-1146124) to LCS.

REFERENCES

1. Janeway Jr CA, Medzhitov R. Innate immune recognition. *Ann Rev Immunol* 2002;**20**(1):197–216.
2. Leulier F, Lemaitre B. Toll-like receptors—taking an evolutionary approach. *Nat Rev Genet* 2008;**9**(3):165–78.
3. Zhang S-M, Adema CM, Kepler TB, Loker ES. Diversification of Ig superfamily genes in an invertebrate. *Science* 2004;**305**(5681):251–4.
4. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, et al. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 2000;**101**(6):671–84.
5. Watson FL, Püttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, et al. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* 2005;**309**(5742):1874–8.
6. Brites D, McTaggart S, Morris K, Anderson J, Thomas K, Colson I, et al. The Dscam homologue of the crustacean *Daphnia* is diversified by alternative splicing like in insects. *Mol Biol Evol* 2008;**25**(7):1429–39.
7. Ng TH, Chiang TA, Yeh TC, Wang HC. Review of DSCAM-mediated immunity in shrimp and other arthropods. *Dev Comp Immunol* 2014;**46**(2):129–38.
8. Dishaw LJ, Mueller MG, Gwatney N, Cannon JP, Haire RN, Litman RT, et al. Genomic complexity of the variable region-containing chitin-binding proteins in amphioxus. *BMC Genet* 2008;**9**(1):78.

9. Dishaw LJ, Giacomelli S, Melillo D, Zucchetti I, Haire RN, Natale L, et al. A role for variable region-containing chitin-binding proteins (VCBPs) in host gut–bacteria interactions. *Proc Natl Acad Sci* 2011;**108**(40):16747–52.
10. Litman GW, Rast JP, Fugmann SD. The origins of vertebrate adaptive immunity. *Nat Rev Immunol* 2010;**10**(8):543–53.
11. Alder MN, Rogozin IB, Iyer LM, Glazko GV, Cooper MD, Pancer Z. Diversity and function of adaptive immune receptors in a jawless vertebrate. *Science* 2005;**310**(5756):1970–3.
12. Herrin BR, Cooper MD. Alternative adaptive immunity in jawless vertebrates. *J Immunol* 2010;**185**(3):1367–74.
13. Boehm T, McCurley N, Sutoh Y, Schorpp M, Kasahara M, Cooper MD. VLR-based adaptive immunity. *Ann Rev Immunol* 2012;**30**:303–20.
14. Jacob F, Vernaldi S, Maekawa T. Evolution and conservation of plant NLR functions. *Front Immunol* 2013;**4**:297.
15. Ghosh J, Buckley KM, Nair SV, Raftos DA, Miller C, Majeske AJ, et al. Sp185/333: a novel family of genes and proteins involved in the purple sea urchin immune response. *Dev Comp Immunol* 2010;**34**(3):235–45.
16. Smith LC. Innate immune complexity in the purple sea urchin: diversity of the *Sp185/333* system. *Front Immunol* 2012;**3**:70.
17. Jones JD, Dangl JL. The plant immune system. *Nature* 2006;**444**(7117):323–9.
18. Maekawa T, Kufer TA, Schulze-Lefert P. NLR functions in plant and animal immune systems: so far and yet so close. *Nat Immunol* 2011;**12**(9):817–26.
19. Van Der Biezen EA, Jones JD. Plant disease-resistance proteins and the gene-for-gene concept. *Trend Biochem Sci* 1998;**23**(12):454–6.
20. Friedman AR, Baker BJ. The evolution of resistance genes in multi-protein plant resistance systems. *Curr Opin Genet Dev* 2007;**17**(6):493–9.
21. McHale L, Tan X, Koehl P, Michelmore RW. Plant NBS-LRR proteins: adaptable guards. *Genome Biol* 2006;**7**(4):212.
22. McDowell JM, Simon SA. Molecular diversity at the plant–pathogen interface. *Dev Comp Immunol* 2008;**32**(7):736–44.
23. Dangl JL, Jones JD. Plant pathogens and integrated defence responses to infection. *Nature* 2001;**411**(6839):826–33.
24. Lam E, Kato N, Lawton M. Programmed cell death, mitochondria and the plant hypersensitive response. *Nature* 2001;**411**(6839):848–53.
25. Nepal MP, Benson BV. CNL disease resistance genes in soybean and their evolutionary divergence. *Evol Bioinform Online* 2015;**11**:49.
26. Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J* 1999;**20**(3):317–32.
27. Kuang H, Woo S-S, Meyers BC, Nevo E, Michelmore RW. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell Online* 2004;**16**(11):2870–94.
28. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. Genome-wide analysis of NBS-LRR–encoding genes in *Arabidopsis*. *Plant Cell Online* 2003;**15**(4):809–34.
29. Roth MO, Wilkins AG, Cooke GM, Raftos DA, Nair SV. Characterization of the highly variable immune response gene family, *He185/333*, in the sea urchin, *Heliocidaris erythrogramma*. *PLoS ONE* 2014;**9**(10):e62079.
30. Rast JP, Pancer Z, Davidson EH. New approaches towards an understanding of deuterostome immunity. *Origin and evolution of the vertebrate immune system*. Springer; 2000. p. 3–16.

31. Nair SV, Del Valle H, Gross PS, Terwilliger DP, Smith LC. Macroarray analysis of coelomocyte gene expression in response to LPS in the sea urchin. Identification of unexpected immune diversity in an invertebrate. *Physiol Genomics* 2005;**22**(1):33–47.
32. Terwilliger DP, Buckley KM, Mehta D, Moorjani PG, Smith LC. Unexpected diversity displayed in cDNAs expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus*. *Physiol Genomics* 2006;**26**(2):134–44.
33. Terwilliger DP, Buckley KM, Brockton V, Ritter NJ, Smith LC. Distinctive expression patterns of *185/333* genes in the purple sea urchin, *Strongylocentrotus purpuratus*: an unexpectedly diverse family of transcripts in response to LPS, β-1, 3-glucan, and dsRNA. *BMC Mol Biol* 2007;**8**(1):16.
34. Buckley KM, Smith LC. Extraordinary diversity among members of the large gene family, *185/333*, from the purple sea urchin, *Strongylocentrotus purpuratus*. *BMC Mol Biol* 2007;**8**(1):68.
35. Dheilly NM, Nair SV, Smith LC, Raftos DA. Highly variable immune-response proteins (185/333) from the sea urchin, *Strongylocentrotus purpuratus*: proteomic analysis identifies diversity within and between individuals. *J Immunol* 2009;**182**(4):2203–12.
36. Brockton V, Henson JH, Raftos DA, Majeske AJ, Kim Y-O, Smith LC. Localization and diversity of 185/333 proteins from the purple sea urchin–unexpected protein-size range and protein expression in a new coelomocyte type. *J Cell Sci* 2008;**121**(3):339–48.
37. Dheilly NM, Birch D, Nair SV, Raftos DA. Ultrastructural localization of highly variable 185/333 immune response proteins in the coelomocytes of the sea urchin, *Heliocidaris erythrogramma*. *Immunol Cell Biol* 2011;**89**(8):861–9.
38. Majeske AJ, Oren M, Sacchi S, Smith LC. Single sea urchin phagocytes express messages of a single sequence from the diverse *Sp185/333* gene family in response to bacterial challenge. *J Immunol* 2014;**193**(11):5678–88.
39. Miller CA, Buckley KM, Easley RL, Smith LC. An *Sp185/333* gene cluster from the purple sea urchin and putative microsatellite-mediated gene diversification. *BMC Genomics* 2010;**11**(1):575.
40. Chen Q, Han Z, Jiang H, Tian D, Yang S. Strong positive selection drives rapid diversification of *R*-genes in *Arabidopsis* relatives. *J Mol Evol* 2010;**70**(2):137–48.
41. Li J, Ding J, Zhang W, Zhang Y, Tang P, Chen J-Q, et al. Unique evolutionary pattern of numbers of gramineous NBS–LRR genes. *Mol Genet Genomics* 2010;**283**(5):427–38.
42. Buckley KM, Munshaw S, Kepler TB, Smith LC. The *185/333* gene family is a rapidly diversifying host-defense gene cluster in the purple sea urchin *Strongylocentrotus purpuratus*. *J Mol Biol* 2008;**379**(4):912–28.
43. Joshi RK, Nayak S. Perspectives of genomic diversification and molecular recombination towards *R*-gene evolution in plants. *Physiol Mol Biol Plants* 2013;**19**(1):1–9.
44. Krasileva KV, Dahlbeck D, Staskawicz BJ. Activation of an *Arabidopsis* resistance protein is specified by the *in planta* association of its leucine-rich repeat domain with the cognate oomycete effector. *Plant Cell Online* 2010;**22**(7):2444–58.
45. Smith LC. Diversification of innate immune genes: lessons from the purple sea urchin. *Dis Model Mech* 2010;**3**(5–6):274–9.
46. Kanazin V, Marek LF, Shoemaker RC. Resistance gene analogs are conserved and clustered in soybean. *Proc Natl Acad Sci* 1996;**93**(21):11746–50.
47. Michelmore RW, Meyers BC. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res* 1998;**8**(11):1113–30.
48. Shen KA, Meyers BC, Islam-Faridi MN, Chin DB, Stelly DM, Michelmore RW. Resistance gene candidates identified by PCR with degenerate oligonucleotide primers map to clusters of resistance genes in lettuce. *Mol Plant Microbe Interact* 1998;**11**(8):815–23.

49. Leister D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trend Genet* 2004;**20**(3):116–22.
50. Zhou T, Wang Y, Chen J-Q, Araki H, Jing Z, Jiang K, et al. Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol Genet Genomics* 2004;**271**(4):402–15.
51. Guo Y-L, Fitz J, Schneeberger K, Ossowski S, Cao J, Weigel D. Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis*. *Plant Physiol* 2011;**157**(2):757–69.
52. Song W-Y, Pi L-Y, Wang G-L, Gardner J, Holsten T, Ronald PC. Evolution of the rice Xa21 disease resistance gene family. *Plant Cell Online* 1997;**9**(8):1279–87.
53. Simons G, Groenendijk J, Wijbrandi J, Reijans M, Groenen J, Diergaarde P, et al. Dissection of the *Fusarium* I2 gene cluster in tomato reveals six homologs and one active gene copy. *Plant Cell Online* 1998;**10**(6):1055–68.
54. Xiao S, Ellwood S, Calis O, Patrick E, Li T, Coleman M, et al. Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by RPW8. *Science* 2001;**291**(5501):118–20.
55. Baumgarten A, Cannon S, Spangler R, May G. Genome-level evolution of resistance genes in *Arabidopsis thaliana*. *Genetics* 2003;**165**(1):309–19.
56. Kajava A. Structural diversity of leucine-rich repeat proteins. *J Mol Biol* 1998;**277**(3):519–27.
57. Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* 2001;**11**(6):725–32.
58. Matsushima N, Miyashita H. Leucine-rich repeat (LRR) domains containing intervening motifs in plants. *Biomolecules* 2012;**2**(2):288–311.
59. Muthamilarasan M, Prasad M. Plant innate immunity: an updated insight into defense mechanism. *J Biosci* 2013;**38**(2):433–49.
60. Dodds PN, Lawrence GJ, Catanzariti A-M, Teh T, Wang C-I, Ayliffe MA, et al. Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc Natl Acad Sci* 2006;**103**(23):8888–93.
61. Mondragon-Palomino M, Gaut BS. Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. *Mol Biol Evol* 2005;**22**(12):2444–56.
62. Xu S, Clark T, Zheng H, Vang S, Li R, Wong GK, et al. Gene conversion in the rice genome. *BMC Genomics* 2008;**9**(1):93.
63. Richter TE, Ronald PC. The evolution of disease resistance genes. *Plant Mol Evol* 2000;**42**(1):195–204.
64. Rast JP, Smith LC, Loza-Coll M, Hibino T, Litman GW. Genomic insights into the immune system of the sea urchin. *Science* 2006;**314**(5801):952–6.
65. Marone D, Russo MA, Laidò G, De Leonardis AM, Mastrangelo AM. Plant nucleotide binding site–leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. *Int J Mol Sci* 2013;**14**(4):7302–26.
66. Luo S, Zhang Y, Hu Q, Chen J, Li K, Lu C, et al. Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family. *Plant physiol* 2012;**159**(1):197–210.