Evolution of antimicrobial cysteine-rich peptides in plants

Huizhen Ma
Chinese Academy of Agricultural Sciences Agricultural Genomes Institute at Shenzhen
Yong Feng
Chinese Academy of Agricultural Sciences Agricultural Genomes Institute at Shenzhen
Qianqian Cao
Chinese Academy of Agricultural Sciences Agricultural Genomes Institute at Shenzhen
Jing Jia
Chinese Academy of Agricultural Sciences Agricultural Genomes Institute at Shenzhen
Muhammad Ali
Sun Yat-Sen University School of Agriculture
Dilip Shah
Donald Danforth Plant Science Center
Blake C. Meyers
Donald Danforth Plant Science Center
Hai He
Sun Yat-Sen University School of Agriculture
Yu Zhang (zhangy2526@sysu.edu.cn)
Sun Yat-Sen University School of Agriculture https://orcid.org/0000-0001-6547-6243

Research Article

Keywords: Antimicrobial peptides, Cysteine-rich peptides, Gene duplication, Plant disease resistance

Posted Date: March 1st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2598984/v1

Version of Record: A version of this preprint was published at Plant Cell Reports on June 28th, 2023. See the published version at https://doi.org/10.1007/s00299-023-03044-3.

Abstract

To defend against diverse groups of pathogens, plants produce cysteine-rich peptides (CRPs) with long-lasting broad-spectrum antimicrobial activity. We analyzed 240 plant genomes ranging from algae to eudicots and our comparative genomics results revealed that CRPs are widely distributed in plants. Further, we found that *CRP* genes have been amplified through both whole genome and local tandem duplication. Their copy number varied significantly across lineages and was associated with the plant ecotype, which may be the result of resistance to changing pathogenic environments. The conserved and lineage-specific CRP families contribute to diverse antimicrobial activities. Moreover, we investigated the unique bi-domain CRPs that result from unequal crossover events. Our findings provide a unique evolutionary perspective on CRPs and provide insights into their antimicrobial and symbiosis characteristics.

Key Message

We analyzed the evolutionary pattern of cysteine-rich peptides (CRPs) to infer the relationship between CRP copy number and plant ecotype, and the origin of bi-domains CRPs.

Introduction

As sessile organisms, plants are frequently exposed to diverse pathogenic microorganisms. Understanding the mechanisms of how plants defend against these biotic factors is fundamental to both plant biotechnology and sustainable agriculture. Plants possess two layers of an innate immune system, called pattern-triggered immunity (PTI) and effector-triggered immunity (ETI) (Jones and Dangl 2006). PTI is triggered by microbe-associated molecular patterns (MAMPs) which are sensed by cell surface-localized pattern-recognition receptors. In addition to activation of PTI at the cell surface, ETI is activated by intracellular pathogen effector proteins which are specifically secreted into the host cell by pathogens and subsequently recognized by intracellularly localized NOD-like receptor proteins which include Toll/Interleukin-1 receptor (TIR) motifs (the TIR-NBS-LRRs) or coiled-coil (CC) domains at the N-terminus (the CC-NBS-LRRs) (Spoel and Dong 2012; Cui et al. 2015; Couto and Zipfel 2016; Yu et al. 2017). In recent years, studies have revealed the role of plant antimicrobial peptides (AMPs) as endogenous factors in plant immune response (Nawrot et al. 2014; Tam et al. 2015; Srivastava et al. 2021).

AMPs are small proteins with potent antibacterial, antiviral, and antifungal activities and are ubiquitous among multicellular eukaryotes. Most plant and animal species express dozens of distinct AMP genes. Plant AMPs are components of plant barrier defenses present in different tissues of diverse plants; they have significant structural and functional diversity (Nawrot et al. 2014)(Srivastava et al. 2021). Plant-derived AMPs exhibit broad-spectrum antimicrobial activity that modulates the innate immune system of different life forms such as plant and animal pathogens, protozoans, and insects; AMPs can even function against cancer cells (Campos et al. 2018; Srivastava et al. 2021). Based on their amino acid sequence identity, number of cysteine residues and their spacing, AMPs can be classified into different

groups (Lay and Anderson 2005). AMPs rich in cysteine residues form multiple disulfide bridges and thus are protected from chemical and proteolytic degradation (Haag et al. 2012). These cysteine-rich peptides (CRPs), namely defensins, comprise one of largest families of AMPs (Kovaleva et al. 2020). The conserved N-terminal region includes a secretion peptide signal that guides CRPs to specific destinations through the secretory pathway, and the C-terminal region which includes a cysteine-rich domain (Zhu et al. 2005; Silverstein et al. 2007; Marshall et al. 2011; Tam et al. 2015; Srivastava et al. 2021). Based on amino acid sequence homology, small variable CRPs were classified mostly as nodule-specific cysteine-rich (NCR) peptides (Scheres et al. 1990), defensins (Florack and Stiekema 1994; Broekaert et al. 1995), lipid transfer proteins (Terras et al. 1992; Molina et al. 1993), hevein-like peptides (Van Parijs et al. 1991), knottin-type peptides (Cammue et al. 1992), snakins family members (Berrocal Lobo et al. 2002), and others (Silverstein et al. 2007; Hammami et al. 2009; Srivastava et al. 2021).

CRPs are not well conserved as the copy number and domain arrangement of these peptides vary significantly among closely related species. For example, NCRs have been reported to drive the terminal differentiation of the rhizobia into nitrogen-fixing bacteroids within nodules of inverted repeat-lacking clade (IRLC) legumes (Terras et al. 1992; Velde et al. 2010; Czernic et al. 2015; Alunni and Gourion 2016). There are ~700 NCR peptides in *Medicago truncatula* whereas only seven have been identified in *Glycyrrhiza uralensis* (Montiel et al. 2017; Roy et al. 2020). Another study has reported a bi-domain defensin, MtDef5, and a number of domains associated with their function (Islam et al. 2017). MtDef5, containing two domains, has more potent antifungal activity than single-domain MtDef5A or MtDef5B, suggesting that the linker peptide APKKVEP contributes to the potent antifungal activity of the bi-domain MtDef5 (Islam et al. 2017). The evolutionary history of how these complexities of CRPs formed remains unknown. Recent advances in comparative genomics have greatly improved our understanding of CRPs and suggest that they are worthy of increased attention. In this study, we performed a search for CRPs in the genomes of 240 plant species using HMMER and Cysmotif. We then clustered, annotated, and analyzed the evolution of the CRPs. Our study demonstrates the pattern of cysteines in the identified CRPs and explores the potential value of AMPs in future research.

Methods

Data used in this study

The data for 240 plant genomes were downloaded from Phytozome, Refseq, and other publicly available databases (Table S1). The branch, order, and family of the 240 plant genomes are referenced from the APG IV system (Chase et al. 2016), published plant genomes (https://plabipd.de/plant_genomes_pa.ep) and PlantRep (Luo et al. 2022). The species tree is mainly referenced from the TimeTree of Life (Kumar et al. 2022), phyloT v2, which is based on NCBI or GTD taxonomy (Schoch et al. 2020), and the literature (Sanjur et al. 2002; Knapp et al. 2004; Lim et al. 2007; Mahelka et al. 2011; Arias and Pires 2012; Yang et al. 2013; Takahashi et al. 2016; Liu et al. 2020; Zhuang et al. 2020). The ecotypes of species were divided into herbaceous and woody species, land and aquatic species, and nodulating and non-nodulating species (Table S1).

Identification of CRP genes

We identified CRP candidate genes in the 240 plant genomes using two different procedures as described previously — HMMER (Johnson et al. 2010) and Cysmotif (Shelenkov et al. 2018). We downloaded the subfamily domain we identified as CRPs from the Pfam database (http://pfam.xfam.org/). Then, we screened all the protein sequences in each genome to identify these subfamily domains: Albumin I (PF08027), Plant antimicrobial peptide (PF14861), Chitin recognition protein (PF00187), Cyclotide (PF03784), Defensin-like (PF10868), Gamma thionin (PF00304), Gibberellin regulated protein (PF02704), Probable lipid transfer protein (PF14368), MiAMP1 (PF09117), Late nodulin protein (PF07127), Potato type II proteinase inhibitor family (PF02428), Thionin (PF00321), Plant lipid transfer proteins (PF00234), Vicilin N-terminal region (PF04702)) using an HMM search, as implemented in hmmer3.3.2 (Finn et al. 2011), and using a threshold value E=1e-5 (available online: http://hmmer.org/). In addition, we downloaded the reported plant CRP motifs, removed the obvious errors, and filtered the eligible ones. Next, we identified various CRPs by the Cysmotif searcher pipeline based on credible CRP motifs (Shelenkov et al. 2018). Finally, we combined the results of these two pipelines to generate a more complete repertoire of CRP candidate genes.

Orthologous/paralogous groups identification, classification, and annotation

We used OrthoFinder (Emms and Kelly 2015, 2019) to identify orthologous/paralogous groups, with default parameters based on protein similarity. We defined these groups as gene families. Then, we classified these families into three categories by R scripts: “highly conserved” families are those present in more than 80% of species in each evolutionary branch; “species-specific” families are those present only in a particular species not in other species; and others. The identified CRPs were functionally annotated using eggnog-mapper (Cantalapiedra et al. 2021) (Table S5), then the annotation of highly conserved and species-specific families was done using shell script.

Definition and classification of *CRP* gene clusters

The *CRP* gene clusters in each family were calculated using customized Perl scripts. In one species, if two *CRP* genes were separated by less than or equal to eight other genes, we considered them to be located in the same *CRP* gene cluster (Richly et al. 2002). If a gene cluster contained *CRP* genes from multiple families, we defined it as a heterogeneous *CRP* cluster. If a cluster contained *CRP* genes only from one family, we defined it as a homogeneous *CRP* cluster. The location of homogeneous CRP clusters on the chromosome of *Arabidopsis thaliana* was shown with the assistance of TBtools (Chen et al. 2020).

Identification of patterns of cysteines

To accurately identify patterns of cysteines in the CRP proteins, the CRP families were clustered into smaller groups by BLASTP and mcl (van Dongen 2000). Next, we aligned them within the small group using mafft (Rozewicki et al. 2019). We masked the non-cysteine positions and generated the consensus sequence of each alignment with customized scripts (schematic diagram in Fig. S3a). In addition, we calculated the distance between two conserved cysteines and counted their frequencies using custom Perl script.

Identification of bi-domain CRPs

For each protein sequence, we used an HMM search (Johnson et al. 2010) to identify the CRP domain and select the high quality or 'credible' sequences containing CRP domains – those that had an alignment proportion greater than 80% with the reference CRP domain. Then we defined the credible sequences with two similar domains as "bi-domain" and the credible sequences with different domains as "single domain". In addition, the identified bi-domains were aligned to single domains in the same species to predict the formation of bi-domains.

Results

Amplification of *CRP*s through whole genome and local tandem duplication

We collected a total of 240 plant genomes from Phytozome, Refseq, and other publicly available databases to identify CRPs. The plant species ranged from lower plants such as algae to higher plants i.e., seed plants, including diversified orders and families (Table S1). We obtained 80,953 CRP candidate genes using HMM and Cysmotif, based on a search for conserved protein domains and sequence homology, and we used these for our subsequent analysis (Table S2). An overview of the process used in our study is shown in Figure S1.

The distribution and counts of *CRPs* varied significantly among the 240 plant genomes (Fig. 1a). To determine if there is a correlation between the *CRP* copy number and plant genome size, we analyzed the linear regression of *CRP* copy number and genome size (Fig. 1b). We found that there is a positive correlation relationship – *CRP* copy number increases with genome size. The positive correlation relationship between *CRP*s and genome size suggested that *CRP*s were duplicated with whole genome duplication events. We hypothesize that whole genome duplication is the factor driving the expansion of the *CRP* gene family members.

Next, we analyzed the clustering of *CRP* genes in the plant genomes. We defined gene clusters as genes on the same chromosome with a distance between two *CRP* genes that is no greater than eight other non-*CRP* genes (Richly et al. 2002). Based on our definition, more than half of *CRP*s formed gene clusters, and gene clusters accounted for more than 50% of the total *CRP*s in more than 30% of the plant species

(Table S3). Among them, all *CRP* genes in *Monoraphidium neglectum*, *Triticum aestivum,* and *Humulus lupulus* could be placed into gene clusters. Gene clusters form as a result of localized gene duplication – tandem duplication and translocation – typically followed by subsequent divergence. The tendency of *CRP*s to cluster on chromosomes is consistent with the evolution of *CRP*s by local tandem duplication.

We next assessed the homogeneity of these *CRP* clusters. We defined as “homogeneous” the genes belonging to the same gene cluster that were orthologous, while we defined as “heterogeneous” the genes belonging to the same gene cluster but not orthologous (classified in Methods, consistent with earlier definitions (Zhang et al. 2016). Here, we take *A. thaliana* as an example to show the position of homogeneous cluster on the chromosomes (Fig. S2). We calculated the proportion of homogeneous cluster in the *CRP* gene clusters; this demonstrated that more than a third of species have more homogeneous than heterogeneous clusters, while the *CRP*s of *Trifolium pratense* were completely homogeneous (Table S3). *CRP* genes of these species were composed of tandem duplications, while other species’ levels of heterogeneous genes were relatively high and represent translocations (Table S3). We conclude that the tendency of *CRP* genes to cluster on chromosomes shows that *CRP* duplication was influenced by local tandem duplication, and the pattern of *CRP* gene clusters on the chromosomes indicates that the duplication of *CRP*s reflects a complex evolutionary history.

CRP copy number variation among evolutionary branches and ecotypes

Statistical analysis of the distribution of *CRP*s in 240 plant genomes revealed that the copy number of *CRP*s varies highly among the species. These species were distributed across 78 families in 50 orders, and in 15 major branches (Fig. 1a, Table S1). A comparison of *CRP*s to a species tree indicated that *CRP*s exist in all 240 species, but their copy number per plant genome ranged from 13 to 2445. Notably, even in the same branch, the difference in the number of *CRP*s was significant; for example, in monocots, *Thinopyrum intermedium* had 2445 CRPs while *Zostera marina* had only 147 (Table S2).

The scattered distribution of species with unusually high or low copy number of *CRP*s among the 240 plant species caused us to speculate that rapid *CRP* expansion or contraction occurred during evolution. First, across different evolutionary branches (Fig. 1a), the copy number of *CRP*s was highly diverse. For example, more than 76% of the species in the algae had *CRP* gene copy numbers fewer than 100, while more than 92% of the species in the Malvids (a eudicot clade) had copy numbers greater than 200. The highest CRP copy number was found in the monocots, but six species of the top ten with the greatest copy number were in the Malvids. This shows that *CRP* copy number varies highly among different branches. Secondly, at the level of families, the number of *CRP*s also significantly differed. For better representation (the number of species > 3), we examined the distribution of *CRP*s among families, which revealed that the CRP copy number of all families significantly differed from each other (Fig. 2a). There was a significant difference in *CRP* gene copy numbers among different families after performing the Student’s t-test (Table S4). In addition, the copy number of *CRP*s varies substantially within the same

family; for example, in the Poaceae, a 2- to 8-fold difference was recorded in gene copy number. The varied gene numbers suggested that *CRP* gene gain or loss occurs rapidly during speciation, which might serve to help plants respond to diversified environments.

We found that the diversity of *CRP*s was also reflected in the different ecotypes of plants and it was closely related to the environment in which they live. A previous study on *NBS-LRR* genes in angiosperms found that *NBS-LRR* gene reduction is associated with ecological specialization (Liu et al. 2021). In our current study, we speculated that there may be a relationship between the *CRP* copy number and different ecological phenotypes. We found that *CRP* distribution was correlated with three groups of different ecological phenotypes: herbaceous and woody species, land and aquatic species, and nodulating and non-nodulating species (Table S4). This shows that there is a substantial difference in the distribution of *CRP*s among each group of ecotypes. The average number of *CRP* genes in aquatic species was significantly lower than land species; herbaceous species had higher *CRP* copy numbers on average relative to woody species (Fig. 2b, Table S4). These results are consistent with the observation that the distribution of *NBS-LRR* genes in specialized lifestyles or environments was lower than common lifestyles or environments. Further, we hypothesized that the living environment of the species affects the *CRP* copy numbers.

Previous studies have shown that CRPs have an antimicrobial function, can defend against the invasion of pathogens, and can respond quickly to changing living conditions (Terras et al. 1995). To investigate our hypothesis that *CRP* copy numbers are affected by the living environment with different ecotypes, defensin-like genes were selected as representative relevant disease-resistance gene subfamily. We analyzed the number of these subfamilies of *CRP*s which displayed extremely significant differences in ecotypic groups (Fig. 2c) as verified by Student's t-test (Table S4). The observed differences in *CRP* copy number between ecological conditions is consistent with the hypothesis that *CRPs* may have evolved to better cope with threats posed by different environments.

Conserved and specific CRP families involved in antimicrobial activities

To examine additional characteristics of CRPs, we analyzed the presence/absence of *CRPs* in each species by clustering *CRPs* into gene families (see Methods). We found that the highly conserved gene families accounted for only 0.74% of the total (Fig. 3a). Here, we took the highly conserved gene family of annotated chitin recognition protein as an example (Fig. 3b). These chitin recognition proteins were found in most species and only missing in 15 species of algae and 9 other species (Fig. 3b, Table S5). Chitin recognition proteins are the basis for recognition of fungi and resisting them via induction of a series of immune responses. This demonstrated that these families of chitin recognition protein are highly conserved.

On the other hand, the species-specific gene families accounted for 22.1% of total (Fig. 3a). For example, we observed a striking peak in the distribution of NCRs and found that most NCRs were present in *M. truncatula,* yet only a few presents in other species, further verifying that NCRs are species-specific (Fig. 3b, Table S5). NCRs are responsible for the control of bacteroid differentiation in IRLC legumes (Velde et al. 2010; Alunni and Gourion 2016), and maintaining the working balance during nitrogen-fixing symbiosis (Wang et al. 2010; Pan and Wang 2017). These results indicated that CRPs also contributed to species-specific plant-microbial interaction.

Despite genomic variation of CRPs, in particular defensins, their three-dimensional structures are relatively conserved, including a stable backbone structure formed by the connection of an α helix with three antiparallel β-strands through four conserved disulfide bonds (Kovaleva et al. 2020). Conserved cysteines form disulfide bonds and play an important role in maintaining stable, three-dimensional protein structures. We identified conserved cysteines in CRP subfamilies (schematic diagram in Fig. S3a), and explored the evolutionary pattern of conserved cysteines to infer the evolution of CRPs. We found that conserved cysteines were prevalent in CRP subfamilies. Further, we calculated the number of other amino acids between two conserved cysteines; we found that a distance of three amino acids was the highest frequency, followed by one and two amino acids (Fig. S3b). The frequency of the distance between cysteines implies that conserved cysteines are concentrated and highly localized.

The unique bi-domain CRPs generated through un-equal crossover event

Some CRPs showed unique characteristics of having two identical cysteine-rich domains; we investigated them. As shown in Figure 4A, there was one sequence (*Medtr8g012775*) that encoded two similar domains on chromosome 8, and another sequence (*Medtr8g012795*) that encoded the corresponding single domain elsewhere on chromosome 8. In the sequence with the bi-domain, exon-2 and its preceding sequence were duplicated to exon-3 and its preceding sequence and formed the bi-domain (Fig. 4a). Correspondingly, we predict the molecular mechanism by which genes encoding bi-domain proteins are formed is that homologous chromosomes were cross-exchanged which caused the single domain to be transformed into a bi-domain. In this process, the functional efficacy of the bi-domain may have become stronger than the single domain. A recent study demonstrated that the bi-domain MtDef5 has more potent antifungal activity against the fungal pathogen *Botrytis cinerea* than its two single domains, MtDef5A and MtDef5B (Islam et al. 2017). This suggests that, in at least this case, the bi-domain has stronger antifungal potential or efficacy than the single domain.

The linker peptide between two single domains may contribute to the potent antifungal activity of the bi-domain MtDef5 (Islam et al. 2017). We were curious about the origin of the linker that connects the two single domains. Thus, we aligned the linker sequence to the other sequences in this species, and found the linker matched the sequence on chromosome 8. This indicated that the linker sequences might be formed by homologous repair during the unequal cross-over of two single domains.

This provides a unique opportunity to explore the evolution of the bi-domain CRP proteins. Based on the sequence characteristics of MtDef5, we also identified additional distinct bi-domain proteins in other CRP sequences, via pfam annotation (Fig. 4b). The genes encoding these bi-domain proteins were derived from homologs encoding single domain proteins in each of the genomes in which they were found. The identification of some bi-domain proteins in Figure 4b suggests that they did not occur by accident in one particular species, but is widespread among plant species and might result from regular events over the course of evolution. It is interesting to hypothesize whether increasing the number of CRPs with bi-domains in plants could be used as a novel and more effective antimicrobial measure. Alternatively, another form of bi-domain proteins may result from the fusion of genes encoding different single domains. Perhaps these gene fusion events increase the efficiency or efficacy of interactions that would otherwise have to occur between separate molecules.

Discussion

With the increasing availability of sequenced plant genomes, our understanding of plant CRP evolution is enhanced. Here, we constructed a dataset including 80,953 CRPs derived from 240 plant genomes (Table S1) to explore the evolutionary pattern of CRPs and the factors that influence their evolution.

CRPs, as a polygenic family, have significant differences in the distribution of their copy number in plant genomes (Fig. 1a), and have positive distribution relationship to genome size (Fig. 1b). *CRP* copy number increases with the plant genomes size. It was speculated that the copy number of polygenic family will expand as the genome duplication as observed for *CRPs*.

CRPs play different roles in different species by their multiple subfamilies (Silverstein et al. 2007), in which *NCRs* are one of the prominent characteristics. *NCRs* play a key role in IRLC legumes and are indispensable peptides in the process of nodule formation and nitrogen fixation (Mergaert et al. 2003; Velde et al. 2010; Alunni and Gourion 2016). This is consistent with the fact that most of NCRs are distributed in *M. truncatula* in our study. Interestingly, it may be that the resulting balance between the effect of NCR peptides and the rhizobia's ability to resist NCR peptides is related to the content of NCR peptides (Wang et al. 2010; Pan and Wang 2017; Pan 2019). When the concentration of NCR peptides becomes too low, symbiosis may fail. Conversely, an overexpression of NCR peptides also caused bacterial death and led to the failure of symbiosis (Pan and Wang 2017). Only an optimal range and concentration of NCR peptides may give one partner some advantage while still maintaining a working symbiosis (Pan and Wang 2017).

Domains are considered the basic and evolutionary units of CRPs, and we identified multiple CRP family members that exist in the form of double or bi-domain proteins. It was reported that MtDef5, the cysteine-rich plant defensin in the genome of the model legume *M. truncatula*, is a unique bi-domain defensin (Islam et al. 2017; Velivelli et al. 2018). This 107-amino acid protein contains two domains, 50 amino acids each, linked by a short peptide APKKVEP (Islam et al. 2017). A previous study hypothesized that the formation of this bi-domain of MtDef5 encoded by a single gene is a fusion of two recently-

duplicated genes encoding single domain defensins in the genomes of *M. truncatula* and *M. sativa* (Islam et al. 2017). In this study, we found that MtDef5 is formed by unequal crossover followed by homologous repair. According to the formation and evolution of bi-domain proteins, it is possible to explore the fusion of multiple functional domains onto one sequence to promote functional diversity.

The structure of CRP is composed of domain units, but it is the presence of disulfide bonds that maintains its stability. Disulfide bonds play an important role in protein folding and structural stability, due to the creation of covalent bonds between cysteine pairs in protein structures (Matsumura et al. 1989). We hypothesize that the conserved cysteines are more likely to form disulfide bonds. In future studies, more sophisticated algorithms are needed to predict disulfide bonds and map the predicted disulfide bonds to the conserved cysteines in the peptide. This will further provide insight into the evolution of disulfide bonds and the conserved cysteines in CRP.

In conclusion, we constructed a dataset containing 240 plant genomes providing a resource to explore various aspects of CRP evolution. Using this large dataset, we studied the evolution of CRP and provided insights into the antimicrobial properties and other functions of CRPs. Further, we studied the cysteine pattern of CRPs to explore the potential value of conserved cysteine for CRP. It is conceivable that with the more comprehensive understanding of the functions of CRPs, we may be able to develop new environmentally friendly pesticides for agriculture or even antibiotics for medicine.

Declarations

Funding

This work was supported by the National Natural Science Foundation of China (32070250), the Natural Science Foundation of Guangdong Province (2020A1515011030) and the open research project of “Cross-Cooperative Team” of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences.

Acknowledgments

We thank Joanna Friesner for helping to revise the grammar.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

YZ, HH, BCM, and DS planned and designed the research. HM and YF performed the bioinformatic analysis. HM and YF analyzed the data. QC contributed to data collection. HM wrote the manuscript. JJ contributed to sort out the literature. MA contributed to revise the grammar. All authors read and approved the manuscript.

Data Availability

All collected data used for this project were taken from available public databases. All other analysis scripts are available at https://github.com/Ma-hz/Evolution-of-plant-CRPs.

References

1. Alunni B, Gourion B (2016) Terminal bacteroid differentiation in the legume-rhizobium symbiosis: nodule-specific cysteine-rich peptides and beyond. New Phytol 211:411–417. https://doi.org/10.1111/nph.14025
2. Arias T, Pires C (2012) A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications. Taxon 61:980–988. https://doi.org/10.1002/tax.615005
3. Berrocal Lobo M, Segura A, Moreno M, et al (2002) Snakin-2, an antimicrobial peptide from potato whose gene is locally induced by wounding and responds to pathogen infection. Plant Physiol 128:951–961. https://doi.org/10.1104/pp.010685
4. Broekaert WF, Terras FRG, Cammue BPA, Osborn RW (1995) Plant defensins: Novel antimicrobial peptides as components of the host defense system. Plant Physiol 108:1353–1358. https://doi.org/10.1104/pp.108.4.1353
5. Cammue BPA, De Bolle MFC, Terras FRG, et al (1992) Isolation and characterization of a novel class of plant antimicrobial peptides from *Mirabilis jalapa* L. seeds. J Biol Chem 267:2228–2233. https://doi.org/10.1016/s0021-9258(18)45866-8
6. Campos ML, De Souza CM, De Oliveira KBS, et al (2018) The role of antimicrobial peptides in plant immunity. J Exp Bot 69:4997–5011. https://doi.org/10.1093/jxb/ery294
7. Cantalapiedra CP, Hern andez-Plaza A, Letunic I, et al (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825–5829. https://doi.org/10.1093/molbev/msab293
8. Chase MW, Christenhusz MJM, Fay MF, et al (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 181:1–20. https://doi.org/10.1111/boj.12385
9. Chen C, Chen H, Zhang Y, et al (2020) TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. Mol Plant 13:1194–1202. https://doi.org/10.1016/j.molp.2020.06.009

10. Couto D, Zipfel C (2016) Regulation of pattern recognition receptor signalling in plants. Nat Rev Immunol 16:537–552. https://doi.org/10.1038/nri.2016.77
11. Cui H, Tsuda K, Parker JE (2015) Effector-triggered immunity: From pathogen perception to robust defense. Annu Rev Plant Biol 66:487–511. https://doi.org/10.1146/annurev-arplant-050213-040012
12. Czernic P, Gully D, Cartieaux F, et al (2015) Convergent evolution of endosymbiont differentiation in dalbergioid and inverted repeat-lacking clade legumes mediated by nodule-specific cysteine-rich peptides. Plant Physiol 169:1254–1265. https://doi.org/10.1104/pp.15.00584
13. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:1–14. https://doi.org/10.1186/s13059-015-0721-2
14. Emms DM, Kelly S (2019) OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol 20:1–14. https://doi.org/10.1186/s13059-019-1832-y
15. Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res 39:29–37. https://doi.org/10.1093/nar/gkr367
16. Florack DEA, Stiekema WJ (1994) Thionins: properties, possible biological roles and mechanisms of action. Plant Mol Biol 26:25–37. https://doi.org/10.1007/BF00039517
17. Haag AF, Kerscher B, Dall'Angelo S, et al (2012) Role of cysteine residues and disulfide bonds in the activity of a legume root nodule-specific, cysteine-rich peptide. J Biol Chem 287:10791–10798. https://doi.org/10.1074/jbc.M111.311316
18. Hammami R, Ben Hamida J, Vergoten G, Fliss I (2009) PhytAMP: A database dedicated to antimicrobial plant peptides. Nucleic Acids Res 37:963–968. https://doi.org/10.1093/nar/gkn655
19. Islam KT, Velivelli SLS, Berg RH, et al (2017) A novel bi-domain plant defensin MtDef5 with potent broad-spectrum antifungal activity binds to multiple phospholipids and forms oligomers. Sci Rep 7:1–13. https://doi.org/10.1038/s41598-017-16508-w
20. Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:1471–2105. https://doi.org/10.1186/1471-2105-11-431
21. Jones JDG, Dangl JL (2006) The plant immune system. Nature 444:323–329. https://doi.org/10.1038/nature05286
22. Knapp S, Chase MW, Clarkson JJ (2004) Nomenclatural changes and a new sectional classification in Nicotiana (Solanaceae). Taxon 53:73–82. https://doi.org/10.2307/4135490
23. Kovaleva V, Bukhteeva I, Kit OY, Nesmelova I V. (2020) Plant defensins from a structural perspective. Int J Mol Sci 21:1–23. https://doi.org/10.3390/ijms21155307
24. Kumar S, Suleski M, Craig JM, et al (2022) TimeTree 5: An Expanded Resource for Species Divergence Times. Mol Biol Evol 39:1–6. https://doi.org/10.1093/molbev/msac174
25. Lay F, Anderson M (2005) Defensins - Components of the Innate Immune System in Plants. Curr Protein Pept Sci 6:85–101. https://doi.org/10.2174/1389203053027575

26. Lim KY, Kovarik A, Matyasek R, et al (2007) Sequence of events leading to near-complete genome turnover in allopolyploid Nicotiana within five million years. New Phytol 175:756–763. https://doi.org/10.1111/j.1469-8137.2007.02121.x
27. Liu C, Yuan D, Liu T, et al (2020) Characterization and Comparative Analysis of RWP-RK Proteins from *Arachis duranensis*, *Arachis ipaensis*, and *Arachis hypogaea*. Int J Genomics 2020:. https://doi.org/10.1155/2020/2568640
28. Liu Y, Zeng Z, Zhang YM, et al (2021) An angiosperm NLR Atlas reveals that NLR gene reduction is associated with ecological specialization and signal transduction component deletion. Mol Plant 14:2015–2031. https://doi.org/10.1016/j.molp.2021.08.001
29. Luo X, Chen S, Zhang Y (2022) PlantRep: a database of plant repetitive elements. Plant Cell Rep 41:1163–1166. https://doi.org/10.1007/s00299-021-02817-y
30. Mahelka V, Kopeck D, Patová L (2011) On the genome constitution and evolution of intermediate wheatgrass (*Thinopyrum intermedium*: Poaceae, Triticeae). BMC Evol Biol 11:1–17. https://doi.org/10.1186/1471-2148-11-127
31. Marshall E, Costa LM, Gutierrez-Marcos J (2011) Cysteine-Rich Peptides (CRPs) mediate diverse aspects of cell-cell communication in plant reproduction and development. J Exp Bot 62:1677–1686. https://doi.org/10.1093/jxb/err002
32. Matsumura M, Signor G, Matthews BW (1989) Substantial increase of protein stability by multiple disulphide bonds. Nature 342:291–293. https://doi.org/10.1038/342291a0
33. Mergaert P, Nikovics K, Kelemen Z, et al (2003) A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. Plant Physiol 132:161–173. https://doi.org/10.1104/pp.102.018192
34. Molina A, Segura A, García-Olmedo F (1993) Lipid transfer proteins (nsLTPs) from barley and maize leaves are potent inhibitors of bacterial and fungal plant pathogens. FEBS Lett 316:119–122. https://doi.org/10.1016/0014-5793(93)81198-9
35. Montiel J, Downie JA, Farkas A, et al (2017) Morphotype of bacteroids in different legumes correlates with the number and type of symbiotic NCR peptides. Proc Natl Acad Sci U S A 114:5041–5046. https://doi.org/10.1073/pnas.1704217114
36. Nawrot R, Barylski J, Nowicki G, et al (2014) Plant antimicrobial peptides. Folia Microbiol (Praha) 59:181–196. https://doi.org/10.1007/s12223-013-0280-4
37. Pan H (2019) More than antimicrobial: Nodule cysteine-rich peptides maintain a working balance between legume plant hosts and rhizobia bacteria during nitrogen-fixing symbiosis. In: de Bruijn FJ (ed) The Model Legume Medicago truncatula, 1st edn. Wiley, Changsha. pp 617–626
38. Pan H, Wang D (2017) Nodule cysteine-rich peptides maintain a working balance during nitrogen-fixing symbiosis. Nat Plants 3:1–6. https://doi.org/10.1038/nplants.2017.48
39. Richly E, Kurth J, Leister D (2002) Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution. Mol Biol Evol 19:76–84. https://doi.org/10.1093/oxfordjournals.molbev.a003984

40. Roy P, Achom M, Wilkinson H, et al (2020) Symbiotic Outcome Modified by the Diversification from 7 to over 700 Nodule-Specific Cysteine-Rich Peptides. Genes (Basel) 11:1–16
41. Rozewicki J, Li S, Amada KM, et al (2019) MAFFT-DASH: Integrated protein sequence and structural alignment. Nucleic Acids Res 47:W5–W10. https://doi.org/10.1093/nar/gkz342
42. Sanjur OI, Piperno DR, Andres TC, Wessel-Beaver L (2002) Phylogenetic relationships among domesticated and wild species of Cucurbita (Cucurbitaceae) inferred from a mitochondrial gene: Implications for crop plant evolution and areas of origin. Proc Natl Acad Sci U S A 99:535–540. https://doi.org/10.1073/pnas.012577299
43. Scheres B, van Engelen F, van der Knaap E, et al (1990) Sequential induction of nodulin gene expression in the developing pea nodule. Plant Cell 2:687–700. https://doi.org/10.1105/tpc.2.8.687
44. Schoch CL, Ciufo S, Domrachev M, et al (2020) NCBI Taxonomy: A comprehensive update on curation, resources and tools. Database 2020:1–21. https://doi.org/10.1093/database/baaa062
45. Shelenkov AA, Slavokhotova AA, Odintsova TI (2018) Cysmotif Searcher Pipeline for Antimicrobial Peptide Identification in Plant Transcriptomes. Biochem 83:1424–1432. https://doi.org/10.1134/S0006297918110135
46. Silverstein KAT, Moskal WA, Wu HC, et al (2007) Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. Plant J 51:262–280. https://doi.org/10.1111/j.1365-313X.2007.03136.x
47. Spoel SH, Dong X (2012) How do plants achieve immunity? Defence without specialized immune cells. Nat Rev Immunol 12:89–100. https://doi.org/10.1038/nri3141
48. Srivastava S, Dashora K, Ameta KL, et al (2021) Cysteine-rich antimicrobial peptides from plants: The future of antimicrobial therapy. Phyther Res 35:256–277. https://doi.org/10.1002/ptr.6823
49. Takahashi Y, Somta P, Muto C, et al (2016) Novel genetic resources in the genus vigna unveiled from gene bank accessions. PLoS One 11:1–18. https://doi.org/10.1371/journal.pone.0147568
50. Tam JP, Wang S, Wong KH, Tan WL (2015) Antimicrobial peptides from plants. Pharmaceuticals 8:711–757. https://doi.org/10.3390/ph8040711
51. Terras FRG, Eggermont K, Kovaleva V, et al (1995) Small cysteine-rich antifungal proteins from radish: Their role in host defense. Plant Cell 7:573–588. https://doi.org/10.2307/3870116
52. Terras FRG, Goderis IJ, Van Leuven F, et al (1992) In Vitro antifungal activity of a radish (*Raphanus sativus* L.) seed protein homologous to nonspecific lipid transfer proteins. Plant Physiol 100:1055–1058. https://doi.org/10.1104/pp.100.2.1055
53. van Dongen S (2000) A cluster algorithm for graphs. Inf Syst [INS] R 0010:1–40
54. Van Parijs J, Broekaert WF, Goldstein IJ, Peumans WJ (1991) Hevein: an antifungal protein from rubber-tree (Hevea brasiliensis) latex. Planta 183:258–264. https://doi.org/10.1007/BF00197797
55. Velde W Van de, Zehirov G, Szatmari A, et al (2010) Plant Peptides Govern Terminal Differentiation of Bacteria in Symbiosis. Science (80-) 327:1122–1126. https://doi.org/10.4159/harvard.9780674333987.c22

56. Velivelli SLS, Islam KT, Hobson E, Shah DM (2018) Modes of action of a Bi-domain plant defensin MtDef5 against a bacterial pathogen *Xanthomonas campestris*. Front Microbiol 9:1–9. https://doi.org/10.3389/fmicb.2018.00934
57. Wang D, Griffitts J, Starker C, et al (2010) A nodule-specific protein secretory pathway required for nitrogen-fixing symbiosis. Scienece 327:1126–1130. https://doi.org/10.1126/science.1184096
58. Yang MQ, van Velzen R V., Bakker FT, et al (2013) Molecular Phylogenetics and character evolution of Cannabaceae. Taxon 62:473–485. https://doi.org/10.12705/623.9
59. Yu X, Feng B, He P, Shan L (2017) From Chaos to Harmony: Responses and Signaling upon Microbial Pattern Recognition. Annu Rev Phytopathol 55:109–137. https://doi.org/10.1146/annurev-phyto-080516-035649
60. Zhang Y, Xia R, Kuang H, Meyers BC (2016) The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them. Mol Biol Evol 33:2692–2705. https://doi.org/10.1093/molbev/msw154
61. Zhu S, Gao B, Tytgat J (2005) Phylogenetic distribution, functional epitopes and evolution of the CSαβ superfamily. Cell Mol Life Sci 62:2257–2269. https://doi.org/10.1007/s00018-005-5200-6
62. Zhuang W, Shu X, Zhang M, et al (2020) Characterization of the complete chloroplast genome of *Populus deltoides* Zhonglin 2025. Mitochondrial DNA Part B Resour 5:3723–3724. https://doi.org/10.1080/23802359.2020.1833773

Figures

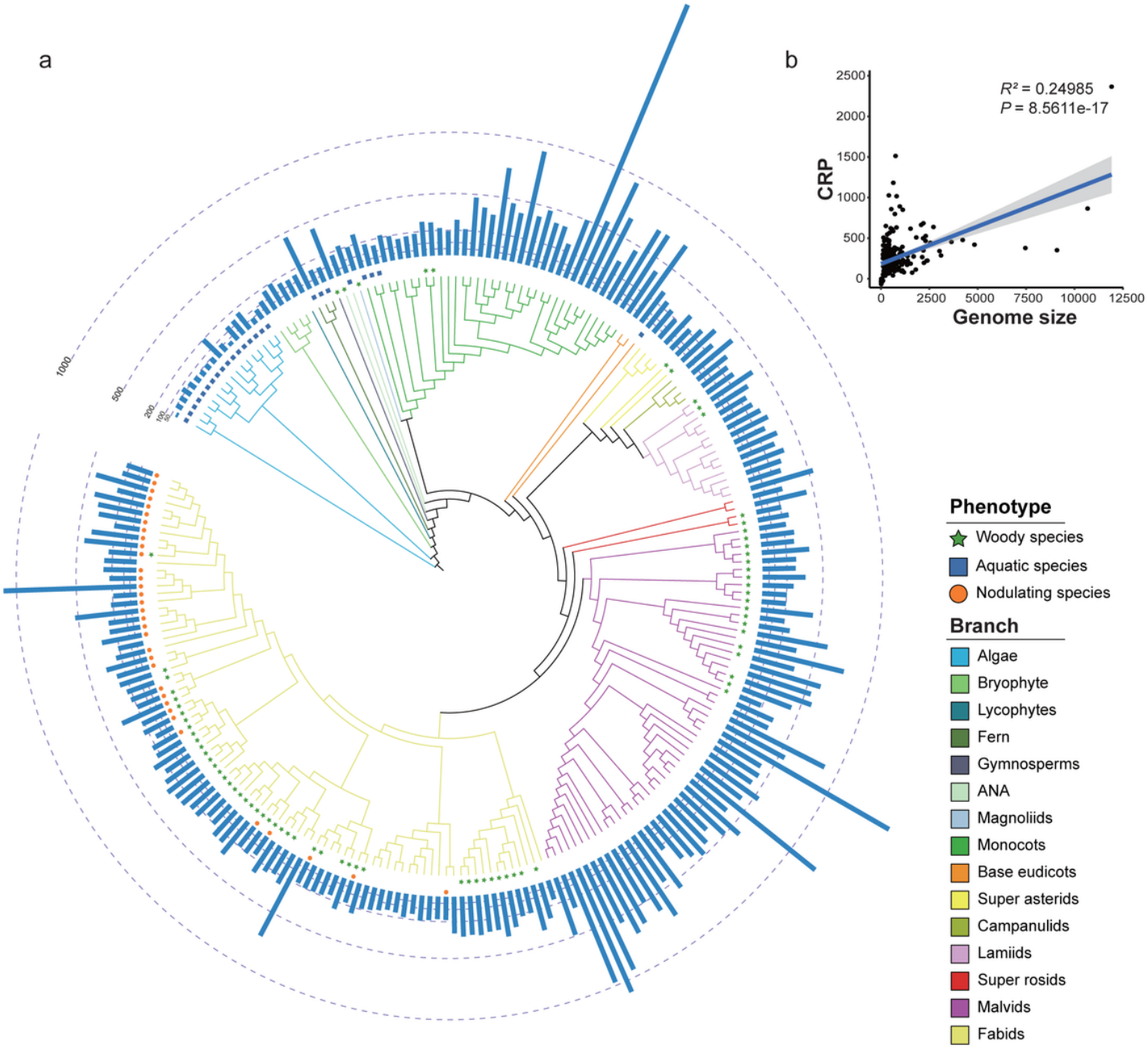

Figure 1

Variations in CRP numbers among 240 plant genomes. (a) Phylogenetic tree was constructed using TimeTree (http://www.timetree.org/), Phylot (https://phylot.biobyte.de/), and published methods (Sanjur et al., 2002; Knapp et al., 2004; Lim et al., 2007; Mahelka et al., 2011; Arias & Pires, 2012; Yang et al., 2013; Takahashi et al., 2016; Liu et al., 2020; Zhuang et al., 2020). Different colors distinguish different evolutionary branches and ecotypes. The numbers of CRP genes in each genome are shown by the blue bars. (b) Correlation between CRP copy numbers and genome size.

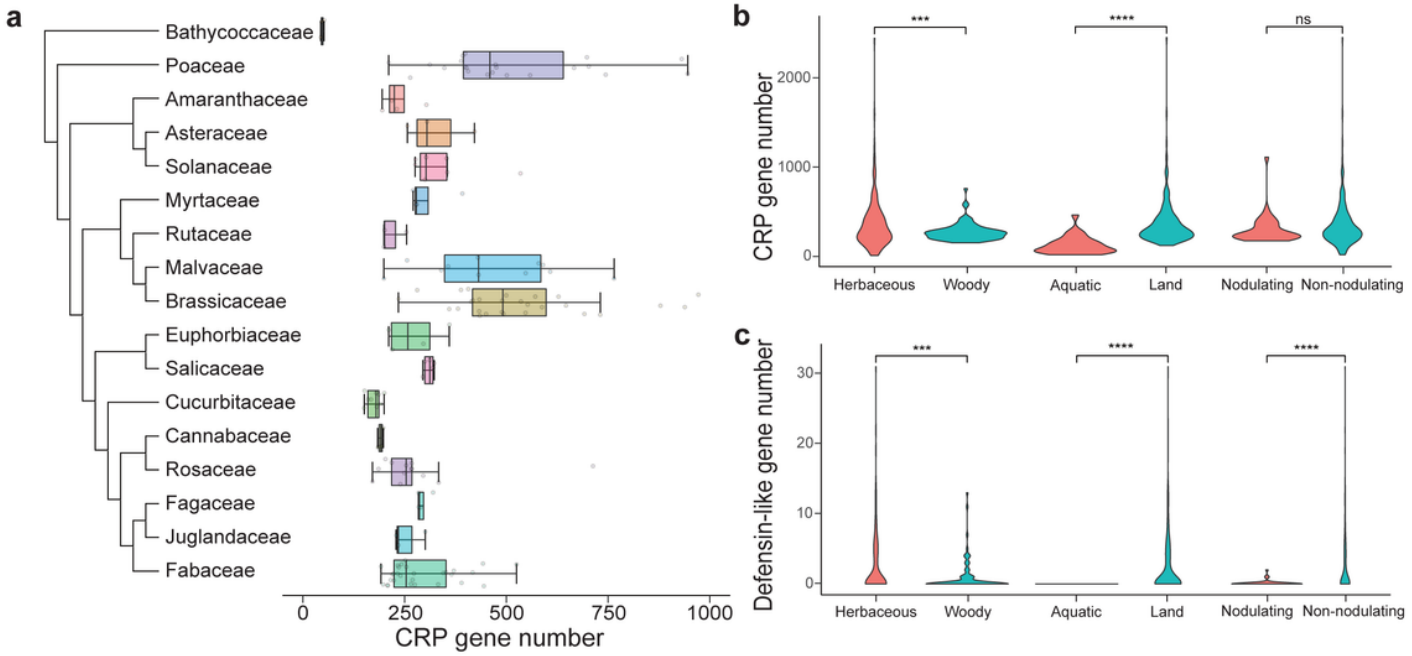

Figure 2

Differential distribution of CRP number. (a) Copy number variation in different plant families. (b) The violin plot of CRP copy numbers in different ecotype groups (herbaceous and woody species, aquatic and land species, nodulating and non-nodulating species). (c) The violin plot of defensin-like subfamily in different ecotype groups.

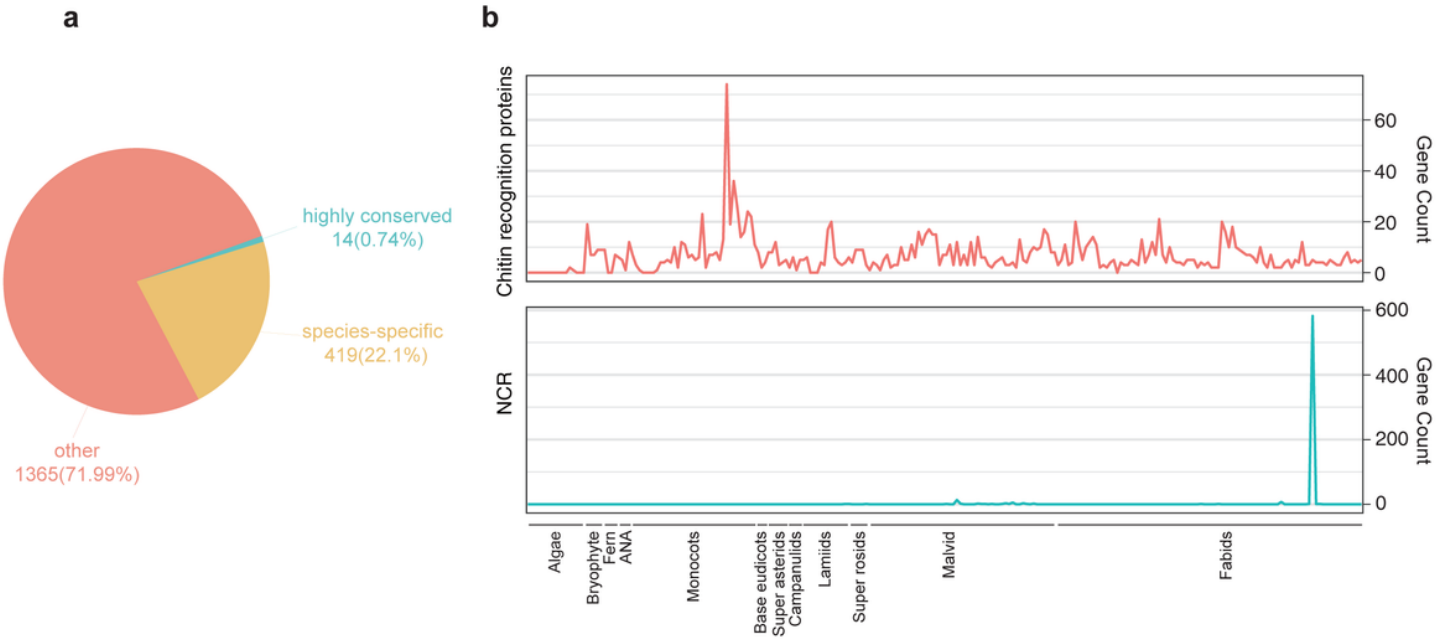

Figure 3

The classification of conservation and specificity of CRP. (a) The pie chart represents ratio of highly conserved gene families, species-specific gene families, and other. (b) The distribution of highly conserved chitin recognition proteins (above) and species-specific NCRs (below) among species from lower plants to higher plants.

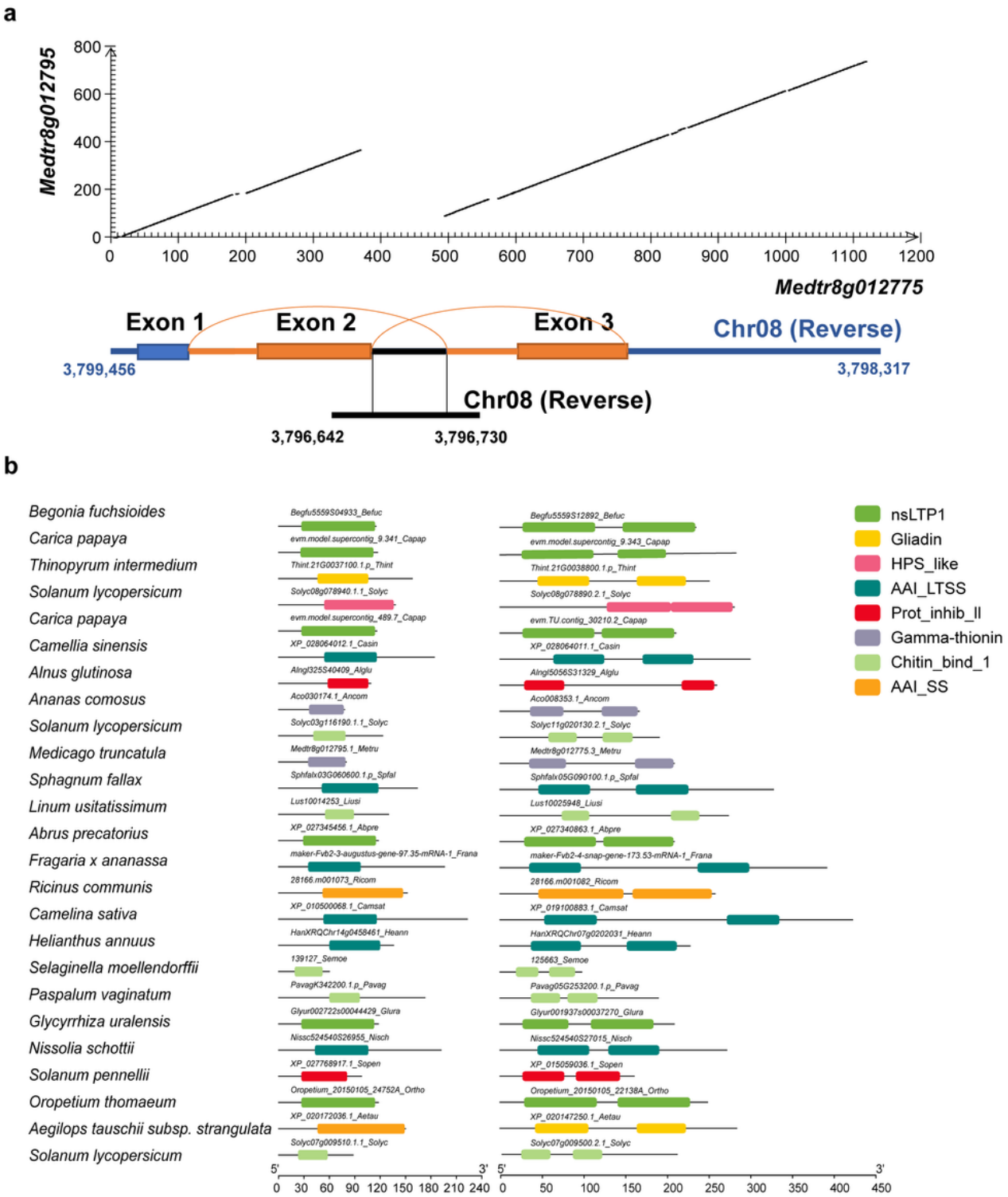

Figure 4

The mechanism of formation of CRP bi-domain-encoding genes. (a) The formation mechanism of bi-domain in Medicago truncatula. The dot plot showed the alignment of the bi-domain sequence (Medtr8g012775) with the single domain sequence (Medtr8g01279) in Medicago truncatula. (b) Other distinct bi-domains protein-encoding genes were found in other species. Filled boxes indicate open reading frames, thin lines indicate introns and flanking sequences.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureSupplement.pdf
- TableS1.Genomesusedinthisstudy.xlsx
- TableS2.CRPandNCRnumbersidentifiedinplantgenomes.xlsx
- TableS3.CRPclustersonthechromosome.xlsx
- TableS4.Phenotypeandsubfamilyamong240species.xlsx
- TableS5.CRPorthologousgroupsandannotation.xlsx