Evolutionary pattern of the presence and absence genes in Fragaria species

Authors: Zhong, Yan, Wang, Ping, Shi, Qinglong, and Cheng, Zong-Ming

Source: Canadian Journal of Plant Science, 102(2) : 427-436

Published By: Canadian Science Publishing

URL: https://doi.org/10.1139/CJPS-2020-0316

Canadian Science Publishing

ARTICLE

Evolutionary pattern of the presence and absence genes in *Fragaria* species

Yan Zhong, Ping Wang, Qinglong Shi, and Zong-Ming Cheng

Abstract: Presence and absence polymorphisms (PAPs) exist extensively and have been investigated in different organisms. However, PAPs have rarely been detected between strawberry species at the genome level. This study identified the presence and absence genes (P/A genes) between wild strawberry species (*Fragaria vesca*) and octoploid cultivated species (*F.* × *ananassa*) under a relatively strict criterion. In total, 333 P/A genes present in the wild strawberry but absent in the cultivated strawberry were detected. Of the P/A genes, 91.89% (306/333) were single genes, and only 8.11% were confirmed as multi-genes. The majority of the identified P/A genes in *Fragaria* were generated by tandem duplications. The P/A genes were unevenly distributed on the seven chromosomes of woodland strawberry, and they clustered preferentially near the telomeric regions of the chromosomes. The P/A genes tended to encode proteins with domains closely associated with responses to varying ecological factors, such as PPR, Protein kinases (PKs), NB-ARC, F-box and EF-hand domains. This indicated that the P/A genes were associated with coping with biotic and abiotic stresses to improve the adaptability of plants to changing environments.

Key words: strawberry, presence and absence genes, gene duplication, biotic and abiotic stresses.

Résumé : Les polymorphismes d'absence et de présence se rencontrent fréquemment; on les a étudiés chez différents organismes. Cependant, ils ont rarement été décelés au niveau du génome chez le fraisier. Les auteurs ont étudié les gènes de présence et d'absence (P/A) chez le fraisier sauvage (*Fragaria vesca*) et les espèces octoploïdes cultivées (*F.* × *ananassa*) selon des critères relativement rigoureux. En tout, 333 gènes P/A identifiés chez le fraisier sauvage n'ont pu être retrouvés chez les variétés cultivées. Sur ce nombre, 91,89 % (306/333) étaient des gènes simples et 8,11 %, seulement, des gènes multiples. La plupart des gènes P/A de *Fragaria* dérivent d'une répétition en tandem. Les gènes P/A sont répartis de manière inégale sur les sept chromosomes de la fraise des bois et se regroupent essentiellement près des télomères. Ils ont tendance à coder des protéines aux domaines étroitement liés à la réaction à divers paramètres écologiques, notamment les protéines PPR, les protéines kinases, les protéines NB-ARC, les protéines à boîte F et celles à main EF. Ces résultats indiquent que les gènes P/A sont associés à la réaction aux stress biotiques et abiotiques, et aident la plante à s'adapter aux changements environnementaux. [Traduit par la Rédaction]

Mots-clés : fraisier, gènes de présence et d'absence, répétition des gènes, stress biotiques et abiotiques.

Introduction

Presence and absence polymorphisms (PAPs) are genomic structural variations that have led to genetic diversity over the course of evolutionary processes in many species (Springer et al. 2009; Conrad et al. 2010; Jiang et al. 2015; Hartmann et al. 2018; Hurgobin et al. 2018). Presence and absence genes (P/A genes) are genes for which the protein-encoding sequences are present in one species but missing in another closely related species (Tan et al. 2012). P/A genes can be generated by duplications, homologous exchanges (HEs) and transposable elements (TEs) (Tan et al. 2012; Darracq et al. 2018; Hurgobin et al. 2018). Gene duplications, which include whole-genome duplications (WGDs), tandem duplications and transposed duplications, provide one of the main sources of genetic variation in plants and animals

Received 8 December 2020. Accepted 11 October 2021.

Y. Zhong,* P. Wang,* and Q. Shi. College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China.

Z.-M. Cheng. College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China; Department of Plant Science, University of Tennessee, Knoxville TN 37996, USA.

Corresponding authors: Yan Zhong (email: yzhong@njau.edu.cn) and Zong-Ming Cheng (email: zcheng@utk.edu).

*These authors contributed equally to this work.

(Houzelstein et al. 2008; Kern and Begun 2008; Gonzalez et al. 2013; Hartmann et al. 2018). WGDs rapidly generate new gene copies, increase the genome size, and are typically followed by extensive loss and specialization of duplicated genes (Bhattacharya et al. 2000; Magadum et al. 2013; Panchy et al. 2016). Tandem duplications are the main contributors to PAP emergence in *Arabidopsis* accessions (Tan et al. 2012). Transposed duplications are associated with TEs that mediate transposition mechanism (Cusack and Wolfe 2007; Wang et al. 2012). HEs chiefly drive the P/A genes arising in the amphidiploid *Brassica napus* (Hurgobin et al. 2018), and TEs are the primary genetic sources of PAPs in maize lines (Darracq et al. 2018). An increasing number of PAPs have been discovered in different organisms, such as two closely related castrating anther-smut fungi (Hartmann et al. 2018), *Acropora digitifera* (Takahashi-Kariyazono et al. 2020), *Drosophila melanogaster* (Schrider et al. 2011), tomatoes (Gao et al. 2019), *Arabidopsis* species (Tan et al. 2012) and maize inbred lines (Darracq et al. 2018). In addition, the P/A genes actively participate in regulating plant growth and development and responding to biotic and abiotic stresses in some species, including *B. napus* (Gabur et al. 2020), oysters (Rosa et al. 2015), *Arabidopsis thaliana* (Shen et al. 2006), *Cucumis melo* (Gonzalez et al. 2013) and pepper (Ou et al. 2018).

The diploid wild woodland strawberry, *F. vesca*, is a typical model plant from the Rosaceae family, and *F.* × *ananassa* is an octoploid cultivated species. *F. vesca* is considered as one of the donors of *Fragaria* species. Previous studies have published six versions of the wild strawberry, including v1.1 (32 831 genes) (Shulaev et al. 2011), v1.1.a2 (33 496 genes) (Darwish et al. 2015), v2.0.a1 (33 673 genes) (Tennessen et al. 2014), v2.0.a2 (33 538 genes) (Li et al. 2018b), v4.0.a1 (28 588 genes) (Edger et al. 2017), and the latest version v4.0.a2 (34 007 genes) (Li et al. 2019*b*). However, the v4.0.a2 of *F. vesca* with 34 006 genes was used as reference genotype to perform BLAST searches. Compared with the six versions (v1.1, v1.1.a2, v2.0.a1, v2.0.a2, v4.0.a1, and v4.0.a2) of genome annotations for wild strawberry, *F.* × *ananassa* has been annotated and assembled only two versions contain v1.0.a1 (Edger et al. 2019) and v1.0.a2 (Liu *et al.* 2021). There were 108 447 genes in v1.0.a2, which showed 360 new genes than the prior annotations (v1.0.a1, 108 087 genes). Nonetheless, the whole-genome sequences of *F.* × *ananassa* was still the previous version (v1.0.a1) (Edger *et al.* 2019) *when* we used it as the database file to perform the BLAST searches. The published *F. vesca* and *F.* × *ananassa* genomes support the identification and analysis of P/A genes between the two species at the genome-wide level. In this study, *F. vesca* Hawaii 4 was used as the reference genotype, and 333 candidate P/A genes were uncovered as present in the *F. vesca* genome but absent from *F.* × *ananassa* Camarosa. Subsequently, chromosome distribution, protein domain, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses of these P/A genes were performed. The results suggest that the P/A genes help strawberries adapt to biotic and abiotic stresses.

Materials and Methods

Identification of the P/A genes

The whole-genome sequences and annotations of *F. vesca* Hawaii 4 (v4.0.a2) (Li et al. 2019*b*) and *F.* × *ananassa* Camarosa (v1.0.a1) (Edger et al. 2019) were downloaded from the Genome Database for Rosaceae (GDR) (https://www.rosaceae.org). Subsequently, the whole-genome nucleotide coding sequences (CDSs, 34 006 CDSs with 2953 bp average length) of *F. vesca* (Li et al. 2019*b*) were used as query sequences for BLASTN against the whole-genome sequences of *F.* × *ananassa* with a default e-value by using local BLAST+. The query CDSs without BLAST hits were determined to be candidate P/A genes; that is, present in *F. vesca* but absent in *F.* × *ananassa*. In contrast, the CDSs with BLAST hits were considered as candidate non-P/A genes. Furthermore, the distributions of these non-P/A genes in *F. nipponica*, *F. nilgerrensis*, *F. iinumae*, *F. viridis* and *F. nubicola* of *Fragaria*, and *Rosa chinensis*, *Rubus occidentalis*, *Prunus persica*, *P. mira*, *Malus* × domestica, *Pyrus pyrifolia* and *P. betulifolia* were analyzed by BLASTN searches with a default e-value using non-P/A gene CDSs as the query sequences and every genome sequence as the database (Hirakawa et al. 2014; Edger et al. 2020; Zhang et al. 2020; Feng et al. 2021; Hardigan et al. 2021).

PCR examinations of the P/A genes

To verify the accuracy of identification of the candidate P/A genes, 32 of them were randomly chosen for three-primer PCR amplifications. Two of the three primers for each selected P/A gene were designed to be located in the two flanking sequences of the P/A regions, and the third primer was located in the middle of the P/A regions (Supplementary Fig. S1[1]). Subsequently, the genomic DNA of *F.* × *ananassa* Camarosa was extracted and used as the PCR template, and that of *F. vesca* Hawaii 4 was used as the positive control template. According to the detection on the gel electrophoresis, the true P/A genes could be confirmed based on the production of two PCR products in *F. vesca* but only one product with a different nucleotide length in *F.* × *ananassa*. In contrast, non-P/A genes exhibited the same PCR results in both species.

Classification of multi-genes and single genes

An all-vs.-all BLASTN search was performed on the whole-genome CDSs of *F. vesca* with a default e-value.

[1]Supplementary data are available with the article at https://doi.org/10.1139/cjps-2020-0316.

The genes were divided into multi-genes families based on the BLAST results with the criteria of coverage ≥60% and identity ≥60%. The remaining genes were defined as single genes. Subsequently, the P/A genes were classified as multi-genes and single genes based on the classification of the genes within the whole genome.

Duplication types of the P/A genes

The different genome-wide duplication types of the genes in *F. vesca* were defined by using the comparative genomic tool DupGen_finder (Qiao et al. 2019). Afterwards, the duplication types of the P/A genes were detected based on their genome-wide duplication types.

Functional analysis of the P/A genes

The proteins encoded by the identified P/A genes were found by searching the Pfam database (http://pfam.xfam.org/) with an e-value cutoff equal to 1.0. GO analysis was performed according to the sequencing annotations of *F. vesca*. KEGG analysis was performed by using the KEGG Automatic Annotation Server (KAAS, https://www.genome.jp/tools/kaas/) and KEGG Mapper (https://www.kegg.jp/kegg/mapper.html).

Chromosomal locations of the P/A genes

The chromosomal positions of the P/A genes were determined from the sequencing annotations of *F. vesca*. Each of the seven chromosomes was divided into different windows in units of 1 Mb, and the gene numbers in each window of all seven chromosomes were counted. The chromosome locations of the P/A genes were performed by MapChart v2.32 software.

Results

Identification and domain preference of the P/A genes

A total of 334 genes were discovered to be present in the *F. vesca* genome but missing in the *F.* × *ananassa* genome. This analysis also demonstrated that 0.98% of the *F. vesca* genes were absent from the *F.* × *ananassa* Camarosa genome. The genomes of the octoploid *F.* × *ananassa* were donated by the diploid ancestor species A, B, C, and D (Hardigan et al. 2021). However, except for *F. vesca* (A) and *F. iinumae* (B) have been verified as the diploid ancestors of *F.* × *ananassa*, the other two subgenomes still remain unknown (Edger et al. 2019; Feng et al. 2021; Hardigan et al. 2021). In this study, the distribution of non-P/A genes was investigated by BLAST+ analysis for some diploid *Fragaria* species used as reference genomes. The non-P/A genes exhibited similar distribution trends in these diploids *Fragaria* species, including 98.31% (33 104/33 673) of the non-P/A genes in *F. nipponica*, 96.80% (32 597/33 673) in *F. nilgerrensis*, 96.52% (32 501/33 673) in *F. iinumae*, 95.85% (32 275/33 673) in *F. viridis* and 94.39% (31 784/33 673) in *F. nubicola*. These results indicated that the non-P/A genes might be commonly prevalent among *Fragaria* species. In addition, to further excavate whether the non-P/A genes were widely distributed in Rosaceae family, some Rosaceae species also served as the reference genomes to perform BLAST + searches. There were 26 600 (79.00%) non-P/A genes in *R. chinensis*, 23 382 (69.44%) in *R. occidentalis*, 16 293 (48.39%) in *P. persica*, 16 096 (47.80%) in *P. mira*, 16 020 (47.58%) in *M. domestica*, 15 769 (46.83%) in *P. pyrifolia* and 15 659 (46.50%) in *P. betulifolia*. Therefore, about 46.50%–79.00% of the non-P/A genes widely distributed in the Rosaceae family, and approximately 15.39% to 19.31% of the non-P/A genes were exclusive to *Fragaria* species.

To verify the true situation of these candidate P/A genes, three-primer PCR amplifications were performed for the 32 randomly selected genes (Supplementary Table S1; Fig. S2[1]). Among the 32 genes, 31 were found to be true P/A genes, and only one was a false P/A gene. For this false P/A gene, we checked the whole-genome sequences of *F.* × *ananassa* and found no corresponding nucleotide sequence of this gene. This indicated that the misjudgment of this gene was due to sequencing or assembly errors rather than to our methods. Therefore, the 333 genes were regarded as the candidate P/A genes for further analysis (Supplementary Table S2[1]).

The P/A genes were involved in a total of 208 protein domains, in which some specific domains showed a relative high occurrence frequency (Fig. 1). Among the top 20 domains in occurrence frequency, the PPR, PKs, NB-ARC, F-box and EF-hand domains are all involved in the response to biotic or abiotic stresses and adaptation to the environment. For example, an *Arabidopsis* PPR protein is closely related to ecological adaption (Zsigmond et al. 2008); PKs and NBS-LRR genes are key plant disease-resistance genes (Li et al. 2016; Zhong et al. 2018; Yang et al. 2019); and EF-hand proteins are important in response to ecological stress in the soybean genome (Zeng et al. 2017).

Multi-gene families and duplication types of the P/A genes

Based on the criteria of coverage and identity both being larger than 60%, a total of 2667 multi-gene families were defined among the whole genome CDSs of *F. vesca*. These families contain 7578 multi-genes, representing 22.28% of the whole-genome genes sorted in multi-gene families. Only 27 P/A genes belonged to multi-gene families, indicating that 8.11% (27/333) of the identified P/A genes were multi-genes. The vast majority of them (306/333) were single genes without homologs in *F.* × *ananassa*. However, 7551 non-P/A genes were classified as multi-genes, and 26 122 non-P/A genes were single genes. The non-P/A genes classified as multi-genes accounted for 22.42% of all non-P/A genes, which was larger than the proportion of multi-genes among the P/A genes.

The duplication types of *F. vesca* genes were uncovered at the genome-wide level and included WGDs, tandem duplications and transposed duplications. In total, the smallest number of genes (3038) were WGDs, and 3061

Fig. 1. The top 20 protein domains of the P/A genes.

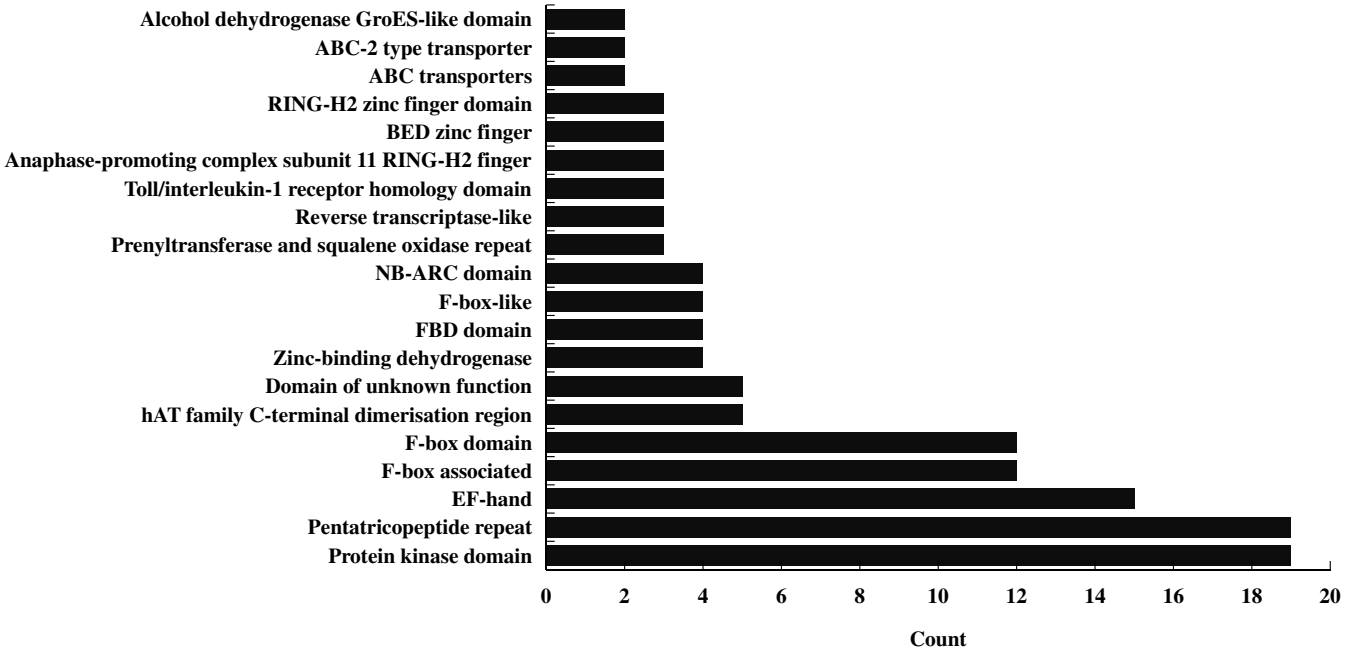


Table 1. Number of P/A genes and non-P/A genes in different duplication types.

	Duplication types			
Gene types	Whole-genome duplications	Tandem duplications	Transposed duplications	Non-duplicated genes
P/A genes	5	34	15	279
Non-P/A genes	3033	3351	3046	24 243
Total	3038	3385	3061	24 522

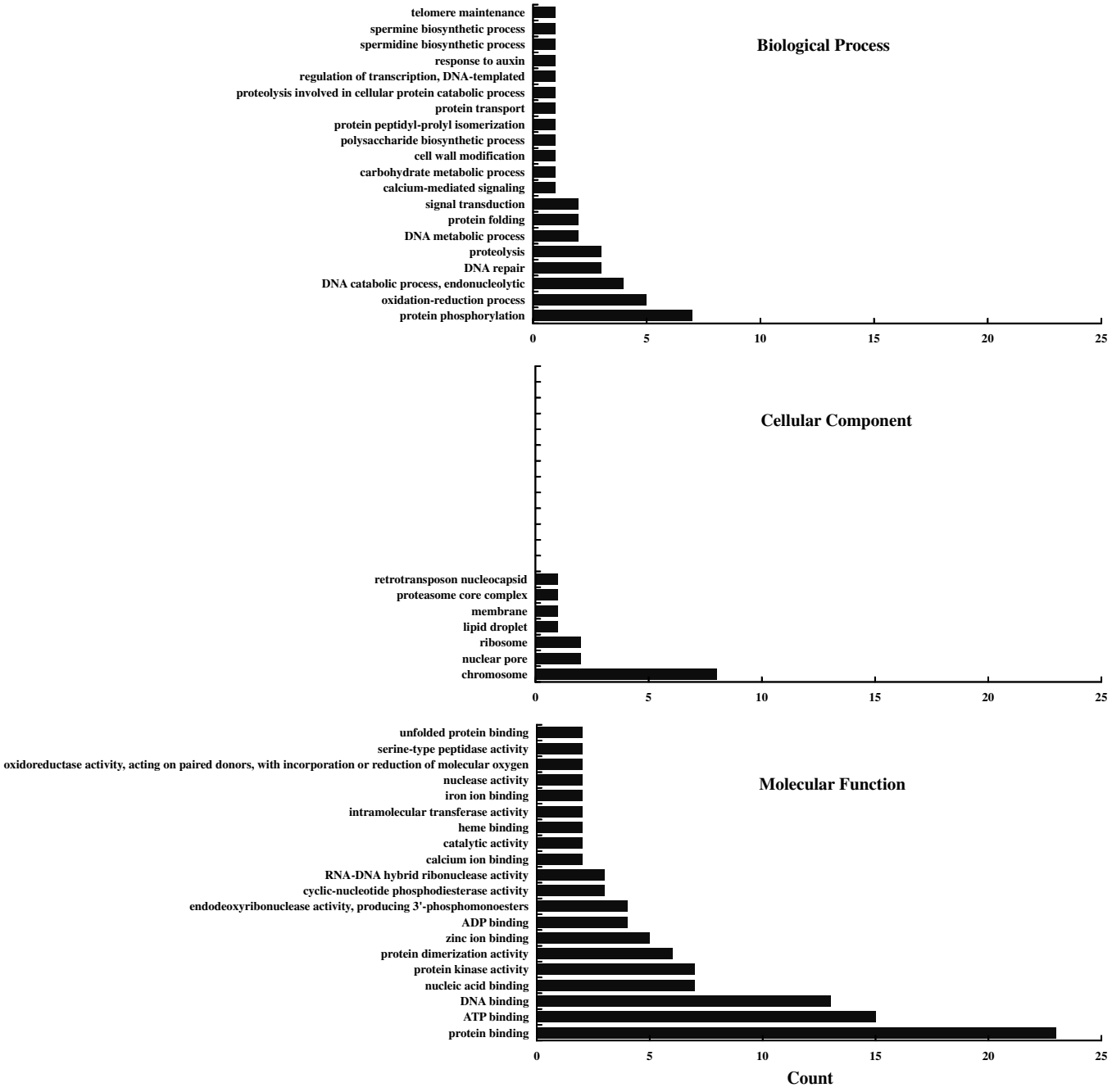
genes were considered transposed duplications. Finally, the maximum number of genes (3385) in the *F. vesca* genome were found to be tandem duplications (Table 1). Among the identified P/A genes, there were 5, 34 and 15 genes from WGDs, tandem duplications and transposed duplications, respectively. These results demonstrated that 1.50% (5/333), 10.21% (34/333) and 4.50% (15/333) of all the P/A genes were generated by WGDs, tandem duplications and transposed duplications, respectively. In addition, 9.00% (3033/33 673), 9.95% (3351/33 673) and 9.05% (3046/33 673) of all the non-P/A genes were involved in WGDs, tandem duplications and transposed duplications. The majority of P/A genes were tandem duplications compared with the other two duplication types. This demonstrated that tandem duplications might play more important roles in P/A gene expansions than WGDs and transposed duplications.

Functional analysis of the P/A genes

Three categories of GO terms, biological process, cellular component and molecular function, were analyzed. Most P/A genes were categorized in terms of molecular function, followed by the biological process and cellular component categories (Fig. 2). In the category of molecular function, many of the P/A genes exhibited binding functions with proteins, nucleotides, metal ions and other molecules, such as the dominant subcategory of protein binding (23, 6.91%), as well as ATP binding (15, 4.50%), DNA binding (13, 3.90%), nucleic acid binding (7, 2.10%) and zinc ion binding (5, 1.50%). A variety of activities were also in this category, including protein kinase activity (7, 2.10%), oxidoreductase activity (3, 0.90%) and catalytic activity (2, 0.60%), etc. In the category of biological process, the P/A genes seemed to be enriched in protein phosphorylation (7, 2.10%) and oxidation–reduction process (5, 1.50%). Within the cellular component category, more P/A genes were associated with the subcategories of chromosome (8, 2.40%), nuclear pore (2, 0.60%) and ribosome (2, 0.60%).

KEGG pathway analysis showed that the P/A genes were preferentially involved in signal transduction pathways such as the MAPK signaling pathway, phosphatidylinositol signaling system, and plant hormone signal transduction (Table 2). The P/A genes also participated

Fig. 2. Gene Ontology (GO) analysis of the P/A genes. The abscissa represents the P/A gene numbers in each category.



in metabolism-related pathways, including amino sugar and nucleotide sugar metabolism, biosynthesis of secondary metabolites and metabolic pathways. In addition, some P/A genes might take part in the response to pathogens for their clustering in the plant–pathogen interaction pathway.

Chromosomal locations of the P/A genes

Except for the two genes (FvH4_c10g00010.1 and FvH4_c1g00300.1) that have not been assembled on the chromosome, the remaining 331 P/A genes were unevenly distributed among the seven chromosomes. The most P/A genes were located on the chromosome 3 (59), 4 (59) and 6 (62), respectively, and the fewest of 24 P/A genes were on the chromosome 7. Some P/A genes were discovered clustering near the telomeric regions (Fig. 3). In particular, fewer P/A genes were in the bottom telomeres than in the upper telomeres on the chromosomes 2, 3, 4 and 6. The P/A genes displayed no conspicuous gene cluster on chromosome 7. However, the P/A genes located in different sizes of gene clusters on the other six chromosomes. For instance, 13 P/A gene

Table 2. KEGG pathways of the P/A genes.

Pathway ID	Pathway	Gene numbers
fve00040	Pentose and glucuronate interconversions	2
fve00053	Ascorbate and aldarate metabolism	1
fve00520	Amino sugar and nucleotide sugar metabolism	1
fve00909	Sesquiterpenoid and triterpenoid biosynthesis	1
fve01100	Metabolic pathways	2
fve01110	Biosynthesis of secondary metabolites	1
fve03010	Ribosome	1
fve03013	RNA transport	3
fve03040	Spliceosome	2
fve03430	Mismatch repair	1
fve04016	MAPK signaling pathway - plant	1
fve04070	Phosphatidylinositol signaling system	1
fve04075	Plant hormone signal transduction	1
fve04141	Protein processing in endoplasmic reticulum	1
fve04144	Endocytosis	1
fve04626	Plant–pathogen interaction	3

constituting a cluster were found on chromosomes 3 around 31–32Mb; a cluster containing 9 P/A genes was near 4–5Mb on chromosome 2.

Discussion

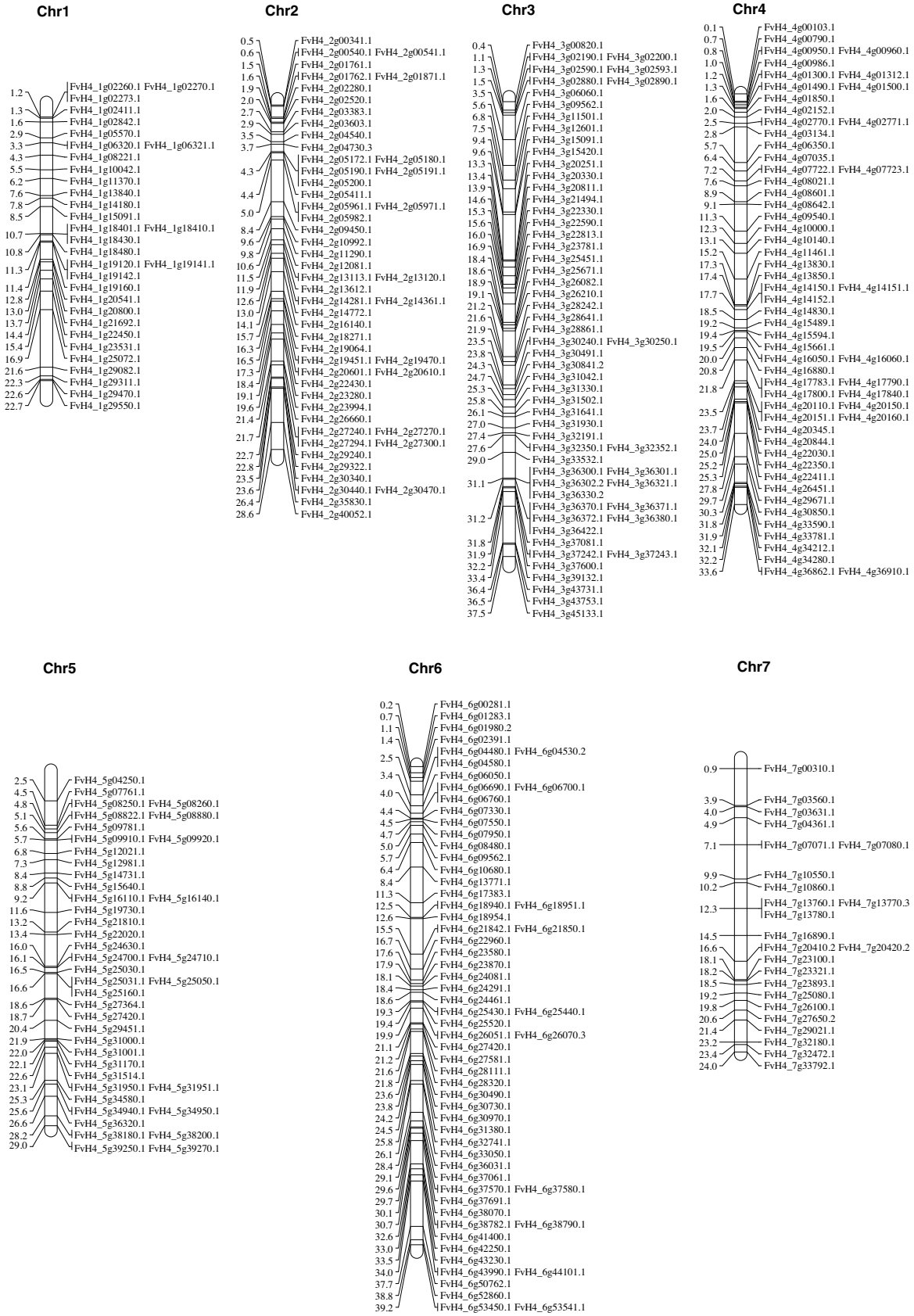
Number variations of the P/A genes in different species

F. virginiana and *F. chiloensis* were evolved from the four diploids of ancestral species (subgenomes A, B, C, and D) since about 1 million years (Njuguna et al. 2013; Hardigan et al. 2021). The allo-octoploid *F.* × *ananassa* species were originated from spontaneous hybrids between *F. virginiana* and *F. chiloensis* and underwent about 300 years of domestication and breeding selection (Edger et al. 2019). *F. virginiana* and *F. chiloensis* might be the intermediate genomes between the diploid and octoploid species and most progenitor genes from these two octoploid ancestral species were remained during the domestication process of *Fragaria* species (Edger et al. 2019; Feng et al. 2021). However, except *F. vesca* (A) and *F. iinumae* (B) were verified as the diploid ancestors of *F.* × *ananassa*, the other two subgenomes still remain unknown (Edger et al. 2020; Hardigan et al. 2021). Our study identified 333 candidate P/A genes in *F. vesca* Hawaii 4 and *F.* × *ananassa* Camarosa genomes, and the remaining 33 673 non-P/A genes were commonly distributed in *Fragaria* genus even across Rosaceae family. This finding was supported by massive highly conservative genes existed across the diploid *Fragaria* species (Hardigan et al. 2019; Feng et al. 2021). Although *F. vesca* genome contributed a large number of genes for the *F.* × *ananassa* genome, the *F.* × *ananassa* chromosomes also maintained more genes from the other ancestral diploids (Edger et al. 2019). The subgenome *F. vesca* accounts for dominant roles in the speciation of the allopolyploid cultivated strawberries, and follows its own independent evolutionary mechanism and stricter selection restrictions (Comai 2005; Edger et al. 2019). However, the diploid species have undergone widespread recombination in intermediate polyploids and a variety of changes have taken place during the evolution of the ancestral species (Edger et al. 2018; Hardigan et al. 2021). As for the 333 candidate P/A genes present in *F. vesca* and absent in *F.* × *ananassa*, the functions could be acquired by the other subgenomes.

The octoploid *F.* × *ananassa* contains four subgenomes, A, B, C, and D (Hardigan et al. 2021). *F. vesca* is considered to be the diploid ancestor of subgenome A of *F.* × *ananassa* and has made great contributions to the adaptation of cultivated strawberry to various environments (Edger et al. 2019; Edger et al. 2020; Hardigan et al. 2021). PAPs commonly appear in nature, but different species have different P/A gene numbers. Using the melon cultivar DHL92 as the reference genome, 1.5% of the genes in the analyzed varieties of *C. melo* were PAPs (Gonzalez et al. 2013). Out of the 10 255 single-copy genes surveyed in the stony coral (*A. digitifera*), 5% of them were determined to be P/A genes (Takahashi-Kariyazono et al. 2020). There was 2% of the genes exhibiting PAPs in 38 strains of *Microbotryum lychnidis-dioicae*, while only 0.6% of the autosomal genes were identified as P/A genes in 19 *M. silenes-dioicae* strains (Hartmann et al. 2018). Furthermore, among the 94 013 genes in *B. napus* genome, 38% of them exhibited PAPs (Hurgobin et al. 2018).

In our study, 333 genes were identified as PAPs between the two *Fragaria* species by using a strict criterion. The results demonstrated that approximately 0.98% of the *F. vesca* genes were absent from the *F.* × *ananassa* Camarosa genome. Only one gene was misidentified among the 32 randomly selected P/A genes checked for correctness by PCR. This false identification was due to a sequencing or assembly error in the

Fig. 3. The locations of the P/A genes on chromosomes. Gene IDs of the P/A genes were shown on the right of each chromosome, and the physical distance is listed on the left of each chromosome (Mb).



whole-genome sequences of *F.* × *ananassa*. In previous studies, 75% of the P/A genes identified by PCR approaches in *Arabidopsis* were categorized correctly (Tan et al. 2012), and the accuracy of P/A gene identification by PCR amplification in *B. napus* was 86% (Gabur et al. 2020). The strict standards used here led to a higher identification accuracy in the studied strawberry species.

Gene duplications are significant for evolution and adaptation to environmental changes in organisms (Magadum et al. 2013). The emergence of P/A genes was closely related to gene duplications in many species. For example, a number of P/A multi-genes were caused by duplications in fungi (Hartmann et al. 2018), *C. melon* (Gonzalez et al. 2013) and *Arabidopsis* (Tan et al. 2012). In our results, 8.11% of the P/A genes were classified into multi-gene families, showing a relatively lower percentage of the P/A genes derived from duplications. This might be due to the small number of genome-wide duplication events in the woodland strawberry genome (Shulaev et al. 2011). Moreover, tandem duplications contributed to the formation of P/A gene clusters on chromosomes in the *Arabidopsis* genome (Kaul et al. 2000; Tan et al. 2012). In this study, 10.21% of the P/A genes were produced by tandem duplication, higher than the proportion of genes from WGDs and transposed duplications.

Distributions of the P/A genes on chromosomes

The genes with PAPs were commonly found to be non-randomly located on each chromosome, with a particular tendency to cluster at centromeres and/or telomeres. A large number of P/A genes are situated near centromeres in *Arabidopsis* (Tan et al. 2012) and at subtelomeric regions and centromeres in the fungal genome (Hartmann et al. 2018). P/A genes also occurred at the telomeric regions on the chromosomes (Winzeler et al. 2003). The locations of P/A genes between maize lines were not randomly distributed on chromosomes, and more P/A genes were located in the telomeric regions than in the centromeric regions (Darracq et al. 2018). The P/A genes tended to cluster in telomeres on the 3L chromosomal arms of *D. melanogaster* and *D. simulans* (Kern and Begun 2008). Similarly, the studied P/A genes in *F. vesca* were nonrandomly distributed along the seven chromosomes. Some of them tended to cluster near the telomeres, especially the chromosomes 2, 3, 4 and 6, with more P/A genes in the upper telomere regions than in the bottom telomeres.

The P/A genes in response to biotic and abiotic stresses

The GO analysis showed that the investigated P/A genes were involved in a variety of biological processes, molecular functions and cellular components. The P/A genes were relatively well-annotated for molecular functions and were especially involved in the binding of proteins, ATP and DNA. Moreover, among the biological process categories, most P/A genes participated in protein phosphorylation and oxidation–reduction processes (redox). Protein phosphorylation is involved in many cells physiological activities and actively participates in cell information transmission (Watanabe and Osada 2016). For example, protein phosphorylation regulates antioxidant enzyme activities to meet the challenges of waterlogging in maize (Xu et al. 2009). Phosphorylation of photorespiration enzymes is closely associated with a series of metabolic reactions in *Arabidopsis* (Hodges et al. 2013). Stimulating redox reactions enhances the response to biotic and abiotic stresses in plants (Cornic and Fresneau 2002; Gonzalez-Bosch 2018). Therefore, it could be inferred that some of the P/A genes might be related to adaptability to environmental stresses in strawberries.

The studied P/A genes encoding NB-ARC, PPR, PKs, F-box and EF-hand domains were functionally enriched for increased stress-related resistance (Kepinski and Leyser 2005; Wang et al. 2016; Zhu 2016; Zeng et al. 2017; Zhang et al. 2019). Similarly, some of the P/A genes were previously shown to belong to the plant resistance gene families (Bush et al. 2014; Rosa et al. 2015; Weisweiler et al. 2019; Gabur et al. 2020), such as NBS-LRR genes in *Arabidopsis* and *C. melon* (Shen et al. 2006; Tan et al. 2012; Gonzalez et al. 2013). In addition, PPR genes improved the resistance to cold stress in rice organs, and PKs enhanced drought tolerance in maize (Chang et al. 2017; Li et al. 2018a; Li et al. 2019a). F-box proteins can increase the abiotic stress resistance of peppers (Yu et al. 2007; Chen et al. 2014), and EF-hand proteins regulate calcium ions (Ca^{2+}) to improve the plant defense and adaptability to environmental stresses in soybeans (Zeng et al. 2017). Therefore, it could be hypothesized that the P/A genes in strawberries might participate in the response to biotic and abiotic stresses.

Conclusions

In summary, a total of 333 genes were detected present in *F. vesca* but absent from *F.* × *ananassa* Camarosa genome. About 10% of the candidate P/A genes were generated by tandem duplications, suggesting that tandem duplications exerted more important roles than other duplication types. The majority of P/A genes were related to the molecular function, followed by biological process and cellular component for GO categories. Moreover, some of the P/A genes preferred to encode the protein domains like PPR, PKs, NB-ARC, F-box and EF-hand domains, which might be closely related to response to biotic and abiotic stresses. The P/A genes were unevenly located on the seven chromosomes. Finally, the non-P/A genes were commonly distributed in the *Fragaria* species and even in Rosaceae plants. This work helps us to better understand the generation types and functions of the P/A genes and provides a novel viewpoint for the evolution of them in plant genomes.

Competing Interests

The authors declare no competing interests.

Contributors' Statement

YZ and ZMC designed and initiated this study. YZ carried out the bioinformatic analysis. YZ and PW wrote the manuscript. QLS designed the primers and performed the PCR amplification. YZ and ZMC critically revised the manuscript. All authors read and approved the final manuscript.

Data Availability

The 333 presence and absence sequences of *Fragaria* species will be available from the corresponding author on reasonable request.

Acknowledgements

This study was supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions, China.

References

Bhattacharya, D., Aubry, J., Twait, E.C., and Jurk, S. 2000. Actin gene duplication and the evolution of morphological complexity in land plants. J. Phycol. **36**: 813–820.

Bush, S.J., Castillo-Morales, A., Tovar-Corona, J.M., Chen, L., Kover, P.X., and Urrutia, A.O. 2014. Presence-absence variation in A-thaliana is primarily associated with Genomic signatures consistent with relaxed selective constraints. Mol. Biol. Evol. **31**: 59–69.

Chang, Y., Yang, H.L., Ren, D.T., and Li, Y. 2017. Activation of ZmMKK10, a maize mitogen-activated protein kinase kinase, induces ethylene-dependent cell death. Plant Sci. **264**: 129–137.

Chen, R.G., Guo, W.L., Yin, Y.X., and Gong, Z.H. 2014. A Novel F- box protein CaF- box is involved in responses to plant hormones and Abiotic Stress in Pepper (*Capsicum annuum* L.). Int. J. Mol. Sci. **15**: 2413–2430.

Comai, L. 2005. The advantages and disadvantages of being polyploid. Nat. Rev. Genet. **6**: 836–846.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y.J., et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature, **464**: 704–712.

Cornic, G., and Fresneau, C. 2002. Photosynthetic carbon reduction and carbon oxidation cycles are the main electron sinks for photosystem II activity during a mild drought. Ann. Bot-Lond. **89**: 887–894.

Cusack, B.P., and Wolfe, K.H. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol. Biol. Evol. **24**: 679–686.

Darracq, A., Vitte, C., Nicolas, S., Duarte, J., Pichon, J.P., Mary-Huard, T., et al. 2018. Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genomics, **19**: 119.

Darwish, O., Shahan, R., Liu, Z.C., Slovin, J.P., and Alkharouf, N.W. 2015. Re-annotation of the woodland strawberry (Fragaria vesca) genome. BMC Genomics, **16**: 29. PMID:25623424.

Edger, P.P., VanBuren, R., Colle, M., Poorten, T.J., Wai, C.M., Niederhuth, C.E., et al. 2017. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. Gigascience, **7**: 1–7.

Edger, P.P., McKain, M.R., Bird, K.A., and VanBuren, R. 2018. Subgenome assignment in allopolyploids: challenges and future directions. Curr. Opin. Plant Biol. **42**: 76–80.

Edger, P.P., Poorten, T., VanBuren, R., Hardigan, M.A., Colle, M., McKain, M.R., et al. 2019. Origin and evolution of the octoploid strawberry genome. Nat. Genet. **51**: 541–547.

Edger, P.P., McKain, M.R., Yocca, A.E., Knapp, S.J., Qiao, Q., and Zhang, T.C. 2020. Reply to: revisiting the origin of octoploid strawberry. Nat. Genet. **52**: 5–7.

Feng, C., Wang, J., Harris, A.J., Folta, K.M., Zhao, M.Z., and Kang, M. 2021. Tracing the diploid ancestry of the cultivated Octoploid strawberry. Mol. Biol. Evol. **38**: 478–485.

Gabur, I., Chawla, H.S., Lopisso, D.T., von Tiedemann, A., Snowdon, R., and Obermeier, C. 2020. Gene presence-absence variation associates with quantitative Verticillium longisporum disease resistance in Brassica napus. Sci. Rep. Uk. **10**: 4131.

Gao, L., Gonda, I., Sun, H.H., Ma, Q.Y., Bao, K., Tieman, D.M., et al. 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat. Genet. **51**: 1044–1051.

Gonzalez, V.M., Aventin, N., Centeno, E., and Puigdomenech, P. 2013. High presence/absence gene variability in defense-related gene clusters of Cucumis melo. BMC Genomics, **14**: 782.

Gonzalez-Bosch, C. 2018. Priming plant resistance by activation of redox-sensitive genes. Free Radical. Bio. Med. **122**: 171–180.

Hardigan, M.A., Feldmann, M.J., Lorant, A., Bird, K.A., Famula, R., Acharya, C., et al. 2019. Genome Synteny has been conserved among the Octoploid progenitors of cultivated strawberry over millions of years of evolution. Front. Plant Sci. **10**: 1789.

Hardigan, M.A., Lorant, A., Pincot, D.D.A., Feldmann, M.J., Famula, R.A., Acharya, C.B., et al. 2021. Unraveling the complex hybrid ancestry and domestication history of cultivated strawberry. Mol. Biol. Evol. **38**: 2285–2305.

Hartmann, F.E., de la Vega, R.C.R.I., Brandenburg, J.T., Carpentier, F., and Giraud, T. 2018. Gene presence-absence polymorphism in castrating Anther-Smut fungi: recent gene gains and phylogeographic structure. Genome Biol. Evol. **10**: 1298–1314.

Hirakawa, H., Shirasawa, K., Kosugi, S., Tashiro, K., Nakayama, S., Yamada, M., et al. 2014. Dissection of the Octoploid strawberry genome by deep sequencing of the genomes of Fragaria species. DNA Res. **21**: 169–181.

Hodges, M., Jossier, M., Boex-Fontvieille, E., and Tcherkez, G. 2013. Protein phosphorylation and photorespiration. Plant Biol. **15**: 694–706.

Houzelstein, D., Goncalves, I.R., Orth, A., Bonhomme, F., and Netter, P. 2008. Lgals6, a 2-million-year-old gene in mice: a case of positive Darwinian selection and presence/absence polymorphism. Genetics, **178**: 1533–1545.

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.K.K., Tirnaz, S., Dolatabadian, A., et al. 2018. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnol. J. **16**: 1265–1274.

Jiang, L., Lv, Y.D., Li, T., Zhao, H., and Zhang, T.F. 2015. Identification and characterization of presence/absence variation in maize genotype Mo17. Genes Genom. **37**: 503–515.

Kaul, S., Koo, H.L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L.J., et al. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, **408**: 796–815.

Kepinski, S., and Leyser, O. 2005. The Arabidopsis F-box protein TIR1 is an auxin receptor. Nature, **435**: 446–451.

Kern, A.D., and Begun, D.J. 2008. Current deletion and gene presence/absence polymorphism: Telomere dynamics dominate evolution at the tip of 3L in Drosophila melanogaster and D-simulans. Genetics, **179**: 1021–1027.

Li, Y.J., Zhong, Y., Huang, K.H., and Cheng, Z.M. 2016. Genomewide analysis of NBS-encoding genes in kiwi fruit (Actinidia chinensis). J. Genet. **95**: 997–1001.

Li, J., Li, Y.H., Deng, Y.L., Chen, P., Feng, F., Chen, W.W., et al. 2018a. A calcium-dependent protein kinase, ZmCPK32, specifically expressed in maize pollen to regulate pollen tube growth. PLoS ONE, **13**: e0195787.

Li, Y.P., Wei, W., Feng, J., Luo, H.F., Pi, M.T., Liu, Z.C., and Kang, C.Y. 2018b. Genome re-annotation of the wild strawberry Fragaria vesca using extensive Illumina- and SMRT-based RNA-seq datasets. DNA Res. **25**: 61–70.

Li, H.S., Han, X.D., Liu, X.X., Zhou, M.Y., Ren, W., Zhao, B.B., et al. 2019a. A leucine-rich repeat-receptor-like kinase gene SbER2–1 from sorghum (*Sorghum bicolor* L.) confers drought tolerance in maize. BMC Genomics, **20**: 737.

Li, Y.P., Pi, M.T., Gao, Q., Liu, Z.C., and Kang, C.Y. 2019b. Updated annotation of the wild strawberry Fragaria vesca V4 genome. Hortic. Res. Engl. **6**: 61.

Liu, T.J., Li, M.Z., Liu, Z.C., Ai, X.Y., and Li, Y.P. 2021. Reannotation of the cultivated strawberry genome and establishment of a strawberry genome database. Hortic. Res. Engl. **8**: 41.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. 2013. Gene duplication as a major force in evolution. J. Genet. **92**: 155–161.

Njuguna, W., Liston, A., Cronn, R., Ashman, T.L., and Bassil, N. 2013. Insights into phylogeny, sex function and age of Fragaria based on whole chloroplast genome sequencing. Mol. Phylogenet. Evol. **66**: 17–29.

Ou, L.J., Li, D., Lv, J.H., Chen, W.C., Zhang, Z.Q., Li, X.F., et al. 2018. Pan-genome of cultivated pepper (Capsicum) and its use in gene presence-absence variation analyses. New Phytol. **220**: 360–363.

Panchy, N., Lehti-Shiu, M., and Shiu, S.H. 2016. Evolution of gene duplication in plants. Plant Physiol. **171**: 2294–2316.

Qiao, X., Li, Q.H., Yin, H., Qi, K.J., Li, L.T., Wang, R.Z., et al. 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. Genome Biol. **20**: 38.

Rosa, R.D., Alonso, P., Santini, A., Vergnes, A., and Bachere, E. 2015. High polymorphism in big defensin gene expression reveals presence-absence gene variability (PAV) in the oyster Crassostrea gigas. Dev. Comp. Immunol. **49**: 231–238.

Schrider, D.R., Stevens, K., Cardeno, C.M., Langley, C.H., and Hahn, M.W. 2011. Genome-wide analysis of retrogene polymorphisms in Drosophila melanogaster. Genome Res. **21**: 2087–2095. PMID:22135405.

Shen, J.D., Araki, H., Chen, L.L., Chen, J.Q., and Tian, D.C. 2006. Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in Arabidopsis thaliana. Genetics, **172**: 1243–1250.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., et al. 2011. The genome of woodland strawberry (Fragaria vesca). Nat. Genet. **43**: 109–116.

Springer, N.M., Ying, K., Fu, Y., Ji, T.M., Yeh, C.T., Jia, Y., et al. 2009. Maize inbreds exhibit high levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome content. Plos Genet. **5**: e1000734.

Takahashi-Kariyazono, S., Sakai, K., and Terai, Y. 2020. Presence-absence polymorphisms of single-copy genes in the stony coral Acropora digitifera. BMC Genomics, **21**: 158. PMID:32054446.

Tan, S.J., Zhong, Y., Hou, H., Yang, S.H., and Tian, D.C. 2012. Variation of presence/absence genes among Arabidopsis populations. BMC Evol. Biol. **12**: 86.

Tennessen, J.A., Govindarajulu, R., Ashman, T.L., and Liston, A. 2014. Evolutionary origins and dynamics of Octoploid strawberry subgenomes revealed by dense targeted capture linkage maps. Genome Biol. Evol. **6**: 3295–3313.

Wang, W.B., Li, X.L., Chen, S.X., Song, S.Y., Gai, J.Y., and Zhao, T.J. 2016. Using presence/absence variation markers to identify the QTL/allele system that confers the small seed trait in wild soybean (Glycine soja Sieb. & Zucc.). Euphytica, **208**: 101–111.

Wang, Y., Wang, X., and Paterson, A.H. 2012. Genome and gene duplications and gene expression divergence: a view from plants. Ann. N Y Acad. Sci. **1256**: 1–14.

Watanabe, N., and Osada, H. 2016. Small molecules that target phosphorylation dependent protein-protein interaction. Bioorgan. Med. Chem. **24**: 3246–3254.

Weisweiler, M., de Montaigu, A., Ries, D., Pfeifer, M., and Stich, B. 2019. Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mRNA sequencing and their power to predict phenotypic traits. BMC Genomics, **20**: 787.

Winzeler, E.A., Castillo-Davis, C.I., Oshiro, G., Liang, D., Richards, D.R., Zhou, Y.Y., and Hartl, D.L. 2003. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. Genetics, **163**: 79–89.

Xu, S.C., Ding, H.D., Su, F.X., Zhang, A.Y., and Jiang, M.Y. 2009. Involvement of protein phosphorylation in water stress-induced antioxidant defense in Maize leaves. J. Integr. Plant Biol. **51**: 654–662.

Yang, P., Praz, C., Li, B.B., Singla, J., Robert, C.A.M., Kessel, B., et al. 2019. Fungal resistance mediated by maize wall-associated kinase ZmWAK-RLK1 correlates with reduced benzoxazinoid content. New Phytol. **221**: 976–987.

Yu, H.C., Wu, J., Xu, N.F., and Peng, M. 2007. Roles of F-box proteins in plant hormone responses. Acta Bioch. Bioph. Sin. **39**: 915–922.

Zeng, H.Q., Zhang, Y.X., Zhang, X.J., Pi, E.X., and Zhu, Y.Y. 2017. Analysis of EF-hand proteins in Soybean Genome suggests their potential roles in environmental and nutritional stress signaling. Front. Plant Sci. **8**: 877.

Zhang, J.X., Lei, Y.Y., Wang, B.T., Li, S., Yu, S., Wang, Y., et al. 2020. The high-quality genome of diploid strawberry (Fragaria nilgerrensis) provides new insights into anthocyanin accumulation. Plant Biotechnol. J. **18**: 1908–1924.

Zhang, S.L., Tian, Z.L., Li, H.P., Guo, Y.T., Zhang, Y.Q., Roberts, J.A., et al. 2019. Genome-wide analysis and characterization of F-box gene family in *Gossypium hirsutum* L. BMC Genomics, **20**: 993. PMID:31856713.

Zhong, Y., Zhang, X.H., and Cheng, Z.M. 2018. Lineage-specific duplications of NBS-LRR genes occurring before the divergence of six Fragaria species. BMC Genomics, **19**: 128. PMID:29422035.

Zhu, J.K. 2016. Abiotic stress signaling and responses in plants. Cell, **167**: 313–324.

Zsigmond, L., Rigo, G., Szarka, A., Szekely, G., Otvos, K., Darula, Z., et al. 2008. Arabidopsis PPR40 connects abiotic stress responses to mitochondrial electron transport. Plant Physiol. **146**: 1721–1737.