BMC Genomics

RESEARCH

Open Access

Common bean pan-genome reveals abundant variation patterns and relationships of stress response genes and pathways

Xu Wang[1,3], Ming Yan[1], Shanshan Cui[1], Fang Li[1], Qingqing Zhao[1], Qingnan Wang[1], Bin Jiang[1], Yixin Huang[2,3], Yang Sun[1*] and Xiangdong Kong[4*]

Abstract

Long-term geographical isolation and the different directions of domestication can cause a large number of genome variations. Population genetic analysis based on a single reference genome cannot capture all the variation information. Pan-genome construction is an effective way to overcome this problem. Resequencing data from 683 common bean landraces and breeding lines provided a pan-genome construction data resource. For the first time, for common bean pan-genome construction, 305 Mb non-reference contigs and 10,452 novel genes were identified. Among these new genes, 373 resistance gene analogs containing 372 variable genes were identified and used to narrow down the candidate genes in *Pseudomonas syringae pv. phaseolicola* resistance quantitative trait locus interval of the common bean. Transcriptome analysis of multiple biotic and abiotic stresses reveals that gene expression patterns are organ-, stress-, and gene conservation-specific. Core and shell genes may be co-expressed in all samples and may have functional complementarity to maintain the stability of plant growth. Within pathways, 8990 and 30,272 mutual exclusivity and co-occurrence gene presence-absence variations (PAVs) were discovered respectively, providing further insights into the functional complementarity of genes. In conclusion, our study provides a comprehensive genome resource, which will be useful for further common bean breeding and study.

Keywords Common bean, Pan-genome, Presence-absence variations, Resistance gene analogs, Mutual exclusivity

Introduction

Common beans are a group of legume crops grown and used by humans as a source of protein and other nutrients [1]. Common bean (*Phaseolus vulgaris L.*) is the most widely grown bean in the *Phaseolus* genus because of its high protein content and low fat content. Common bean is grown in several countries, particularly in Asia and the USA, where domestication and environmental effects have resulted in the formation of many different varieties [2]. These varieties are landraces or breeding lines that have been improved from landraces. The use of these data resources requires large-scale collection and sequencing work. Wu et al. [3] performed whole-genome resequencing of 683 lines located worldwide, providing valuable data for common bean breeding.

*Correspondence:
Yang Sun
2018259@ahnu.edu.cn
Xiangdong Kong
xdkong@zju.edu.cn
[1] Anhui Provincial Key Laboratory of the Conservation and Exploitation of Biological Resources, Key Laboratory of Biotic Environment and Ecological Safety in Anhui Province, College of Life Sciences, Anhui Normal University, Wuhu, Anhui 241000, China
[2] Collaborative Innovation Center of Recovery and Reconstruction of Degraded Ecosystem in Wanjiang Basin Co-Founded By Anhui Province and Ministry of Education, School of Ecology and Environment, Anhui Normal University, Wuhu, Anhui, China
[3] Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China
[4] JiguangGene Biotechnology Co. Ltd, Nanjing, China

For resequencing data analysis, most studies performed single-nucleotide polymorphism (SNP) or insertion and deletion (indel) calling based on the reference genome and used SNP as markers for a series of population genetic analyses [4, 5]. However, the reference genome was obtained from a single strain within the species; for example, the reference genome of common bean was obtained by sequencing an inbred landrace line of *P. vulgaris* (G19833) from the Andean pool [6]. A single genome cannot capture the total genes of a species; therefore, pan-genome construction could solve this problem. Currently, there are two main strategies for pan-genome construction: iterative assembly or de novo assembly using whole-genome resequencing data [7–9] and assembling multiple high-quality genomes by third-generation sequencing and HiC (a high-throughput chromosome conformation capture technique) sequencing to further construct a graphical pan-genome [10, 11]. Because of the high cost of constructing multiple high-quality genomes, using next-generation sequencing (NGS) data to construct pan-genomes is a rapid and low-cost method that offers the possibility of pan-genome sequencing in multiple species. Therefore, many second-generation sequencing-based pan-genome construction tools have been developed [12–14].

Genes and genomic sequences that are lost in the reference genome can be identified by pan-genome construction; for example, the pan-genome of mung bean has 287.73 Mb sequences, which contains 3337 protein-encoding genes, absent in the reference genome [15]. Analysis of presence-absence variation (PAV) information can reveal the absence of promoters and genes that may affect the phenotypes [16, 17] or explain the missing heritability [18]. In pan-genome studies, the gene absences identified in the population can also be used together with other variant types, such as a missense variant, to evaluate the selection pressure on a gene; for example, a candidate gene in a quantitative trait loci (QTL) interval can be further screened by population-wide variation information [19, 20]. Therefore, the pan-genome is an important resource for species-wide studies, and many studies have constructed pan-genomic databases that provide such resources [21, 22].

In the present study, a common bean pan-genome containing 305 Mb non-reference contigs and 10,452 protein-coding genes was constructed from the whole-genome resequencing data of 683 common beans from the study by Wu et al. [3]. The PAV information of all genes in the pan-genome was then constructed. Based on the gene PAV, the PAV of RGAs was analysed to provide a more comprehensive understanding of the mutation numbers and types of RGAs in the resistance QTL interval to narrow down candidate genes. According to the transcriptional profiles of a variety of biotic and abiotic stresses, the relationship between PAV and stress-responsive genes revealed the expression characteristics of core/softcore, shell, and cloud genes in different tissues and growth environments. Finally, by identifying the mutual exclusivity and co-occurrence of gene PAVs within and between pathways, we can gain a more comprehensive understanding of functional complementarity and co-selection of genes. In conclusion, the pan-genome provides a valuable resource for further study and breeding of common beans.

Materials and methods

Resequencing and RNA-seq data retrieval

Wu et al. [3] resequenced 683 common bean samples, obtaining a total of 4.27 tera base pairs (bp) of NGS data. We obtained the raw whole genome sequencing data for all samples from the National Centre for Biotechnology Information (NCBI) (PRJNA515107). In addition, biotic and abiotic stress-related RNA-seq data from NCBI (PRJNA288189, PRJNA648388, PRJNA691982, PRJNA746732, PRJNA793687, PRJNA311998, PRJNA656794, PRJNA741786, and PRJNA758821) were obtained (Table S1).

Pan-genome construction

The raw paired-end WGS reads were trimmed by Fatsp v0.23.4 [23] with default parameters. Megahit v1.2.9 [24] was used to assemble the genome for each sample with default parameters. After removing contigs shorter than 500 bp, a tool called nucmer in Mummer v4.0 [25] was used to align the remaining contigs with the reference genome with default parameters. The unaligned contigs were classified into two types: 1) fully unaligned contigs, meaning the contigs contained sequence identity and length exceeding 90% and 300 bp, respectively, with the reference genome; 2) partially unaligned contigs, meaning the contigs contained regions that were longer than 500 bp with an identity of < 90%. Unaligned sequences of partially unaligned contigs were extracted. All non-references were compared to the NCBI NT database (https://ftp.ncbi.nlm.nih.gov:/blast/db/FASTA/nt.gz) using blastn v2.14.1 +with parameters "-evalue 1E-5 -max_target_seqs 1", and sequences not belonging to Eukaryota or sequences belonging to Eukaryota but not Viridiplantae were removed. Clean sequences were merged, and redundancy was removed using cd-hit v4.8.1 [26] with parameters "-i 0.9 -M 200000". In addition, we used two other strategies to further remove redundancy: 1) blastn-based all vs. all comparisons and 2) nucmer-based all vs. all comparisons. A threshold of 90% for regions with 90% sequence identity was used.

Pan-genome annotation

A de novo repeat sequence database was built using RepeatModeler v2.0.5 [27] with parameters "-engine ncbi". RepeatMasker v4.1.7 was used to annotate non-references based on the repeat sequence database [28]. RepeatProteinMask v4.1.7 was used to search for repeated sequences in the TE protein database with parameters "-engine ncbi -noLowSimple -pvalue 0.0001". Tandem repeats finder version 3 was used to identify tandem repeats in the non-reference contigs [29].

The MARKER2 v3.01.03 [30] was used to predict the gene structure of the genome based on repeat sequence-masked contigs. We used Augustus v3.4.0 [31] for de novo gene prediction with default parameters, and the model was trained by the *Phaseolus vulgaris* reference annotation. RNA-seq data from 135 accessions (Table S2) were used as the evidence for transcription. Fastp v0.23.4 was used to remove low quality sequences with default parameters [23]. We used hisat2 v2.2.1 [32] to map clean reads to non-reference contigs, and samtools v1.17 were used to extract reads that could be aligned with non-reference sequences with parameters "-f 12/-f 68/-f 132" [33]. Trinity v2.15.2 [34] was used to de novo assemble reads that mapped to non-reference sequences in each sample with parameters "–seqType fq". After merging the assembled transcript sequences, we used cd-hit-est [26] to remove redundancy (with default parameters). Finally, annotation of the non-reference sequence was obtained using maker2. We removed the genes which overlapped 50% with the repeat annotations. Furthermore, Interproscan v5.55–88.0 [35] was used to annotate gene sequences with parameters "-t p -f gff3 -appl PfamA -goterms -pa -iprlookup", and genes annotated with interpro domains were retained.

The predicted gene sequences were compared with the NT database using blastn and the NR, uniport, and swissport databases using blastx. Simultaneously, GO and KEGG annotations were obtained through correspondence between the databases.

Presence-absence variation analysis

All the reference genome and non-reference sequences were merged into the common bean pan-genome sequence. Bwa was used to align the 683 resequencing data to the pan-genome. Ccov in HUPAN [36] was used to calculate the ratio of each gene to the coding sequence (CDS) region covered by the reads. A gene whose region was covered by the reads of a sample by >80% and <80% was considered to exist and not exist in the sample, respectively. As reported by Gao et al. [16], genes that existed in all accessions were defined as core genes, genes that existed in 99%–100% of accessions (676–683 accessions) were defined as softcore genes, genes that existed in 1%–99% of accessions (68–676 accessions) were defined as shell genes, and genes that existed in less than 1% of accessions (68 accessions) were defined as cloud genes.

RGA gene prediction and QTL integration

RGAugury pipeline [37] was used to identify RGAs in the common bean pan-genome with default parameters. RGA includes nucleotide-binding site-leucine-rich repeat (NBS-LRR), receptor-like kinases (RLK), receptor-like proteins (RLP), and transmembrane coiled-coil domain protein (TM-CC) candidate genes, which can be divided into 12 subfamilies. These resistance genes were divided into two groups (core and variable genes) based on the results of the PAV analysis. SNP information of the common bean population was obtained (https://doi.org/https://doi.org/10.5281/zenodo.3236786) and annotated using the Variant Effect Predictor v99 with parameters "–fork 8 –force_overwrite –no_intergenic" [38]. The overlapping was determined using bedtools v2.16.2 intersect [39]. González et al. [40] identified several intervals of resistance to *Pseudomonas syringae pv. Phaseolicola* based on QTL localisation, four of which contained RGA genes. Waterfall plots with SNP and PAV information were drawn using GenVisR v1.11.3 [41].

Protease inhibitor gene cluster

The genomes of *V. unguiculata*, *G. max* and *A. hypogaea* were obtained from the NCBI under the accession numbers GCF_004118075, GCF_000004515 and GCF_003086295 respectively. With the keyword "Protease inhibitor (PI), the PI sequences in the NR database were extracted. The legume protein sequences were compared with the sequences for candidate PI gene identification. Subsequently, the interproscan annotation was combined with the literature information, which was used to select genes that were supported as PI genes. Based on the identified candidate PI genes, a genome-wide scan of the legume genome was performed using an in-house perl script. A 150 kb window was used to scan for the presence of multiple PI genes, and the adjacent eligible genome regions were then merged. Mcscanx was used to identify the synteny or collinearity of this gene cluster across multiple species genomes [42] with default parameters. The CDS files and GFF format annotation files of *V. unguiculata*, *G. max, A. hypogaea,* and *Phaseolus vulgaris L.* were used for analysis. The Python version of Mcscanx was used in the present study, which was obtained from https://github.com/tanghaibao/jcvi. The figure showing the genome synteny relationship was plotted using jcvi.

RNA-seq data analysis

The raw paired-end reads were trimmed and quality was controlled by fatsp [23]. The clean reads were mapped to the reference genome using hisat2 [32]. To identify differentially expressed genes (DEGs) between sample groups, DEseq2 was used [43] and | log2 Fold Change| of ≥ 1 and FDR of < 0.05 were used as a cut-off for significant differential expression. HTSeq was performed to count the number of reads for each gene [44]. Gene expression levels were calculated as fragments per kilobase of exon per million mapped reads (FRKM). The FPKM values of the DEGs were used to perform principal components analysis (PCA) using the vegan R package [45].

All DEGs were used for weighted gene co-expression network analysis (WGCNA) analysis, and their expression was calculated using FPKM values [46]. The enrichment significance of the PAV and GO terms of genes in each module was calculated using the hypergeometric distribution test.

Mutually exclusive and co-occurrence PAV analysis

Using the binary gene PAV data, mutually exclusive and co-occurrences between PAVs were calculated using Rediscover [47]. Genes with mutually exclusive and co-occurring PAVs within and between pathways were extracted based on the KEGG annotation information of the common bean pan-genome.

Results

Pan-genome construction and PAV analysis

We obtained the 683 whole genome sequencing data from Wu et al. [3] which were sequenced between 4.73 ×and 21.8x (Table S3). All the samples were de novo assembled. After removing the reference sequences, decontamination, and redundancy, 305 Mb novel non-reference contigs were obtained, and 10,452 protein-coding genes were predicted in the novel sequences. In addition to the 473 Mb reference genome sequences and 28,125 reference genes, 778 Mb genome sequences and 38,577 genes in common bean pan-genomes were obtained.

As in the study by Gao et al. [40], we classified genes into core (11,290), softcore (9809), shell (11,024), and cloud (6454) genes (Fig. 1A, D, Table S4). The numbers of these gene types were related to the number of samples, particularly when the number of samples was small, changed rapidly with the number of samples, and stabilised when the number of samples reached a high level. We constructed a phylogenetic tree of 683 samples based on the binary gene PAV information. The results showed that the tree constructed by gene PAV exhibited the same affinities of varieties as the SNP-based phylogenetic trees constructed by Wu et al. [3] (Fig. 1B). This indicates that gene-based PAV information can also be used to accurately analyse the relationships between varieties. We also calculated gene frequencies in the Andean and Mesoamerican populations and used Fisher's test to assess the significance of frequency differences between the two groups. The results showed that 842 genes had significantly higher frequencies in the Andean population, while 1,662 genes had significantly lower frequencies (Figure S2 A, Table S5). Genes with high frequencies in the Andean population were enriched in pathways such as flavone and flavonol biosynthesis and flavonoid biosynthesis (Figure S2B), indicating population-specific frequency differences in genes related to the synthesis of flavonoids and other secondary metabolites. These results provide insights for exploring phenotypic diversity in common bean metabolic traits. With an increase in the number of samples, the number of pan-genome genes increased continually, whereas the number of core genes decreased continually (Fig. 1E). Shell genes were enriched in some pfam domains, such as polysaccharide biosynthesis and the carboxylesterase family (Fig. 1C).

Identification of RGA genes in pan-genome and QTL integration

RGAs play a crucial role in plant defense mechanisms, as they are often associated with responses to biotic stresses such as pathogen infection. A total of 1902 RGAs were identified in the pan-genome, of which 1529 were located on the reference and 373 on non-reference contigs (Table 1). Of these RGAs, 806 were core genes and 1096 were variable genes, and 372 of the 373 were located in non-reference contigs. This indicates that RGAs are frequently selected in the breeding process, with a large number of RGA genes gained and lost during the domestication and breeding process. RGAs contain multiple resistance gene types, and the number of genes identified in each resistance gene family is inconsistent across the pan-genome owing to the different number of gene family members. For example, 36 RLK genes were identified in the pan-genome additional contigs, and a total of 843 were identified in the pan-genome, which contained 163 variable genes (shell and cloud genes) and 680 core and softcore genes. Compared to 95.7% of RLK genes located in the reference genome, only 66.9% of NL-like resistance genes were found. Moreover, the number of core/softcore genes (99 genes) in the NL class of resistance genes was lower than the number of shell/cloud genes (49 genes). This suggests that different types of resistance genes have different conservation levels, which may be related to their functional differentiation. Additionally, the distribution of resistance genes on the chromosomes was not uniform (Fig. 2A), which may be related to the different ways in which the genes replicated.

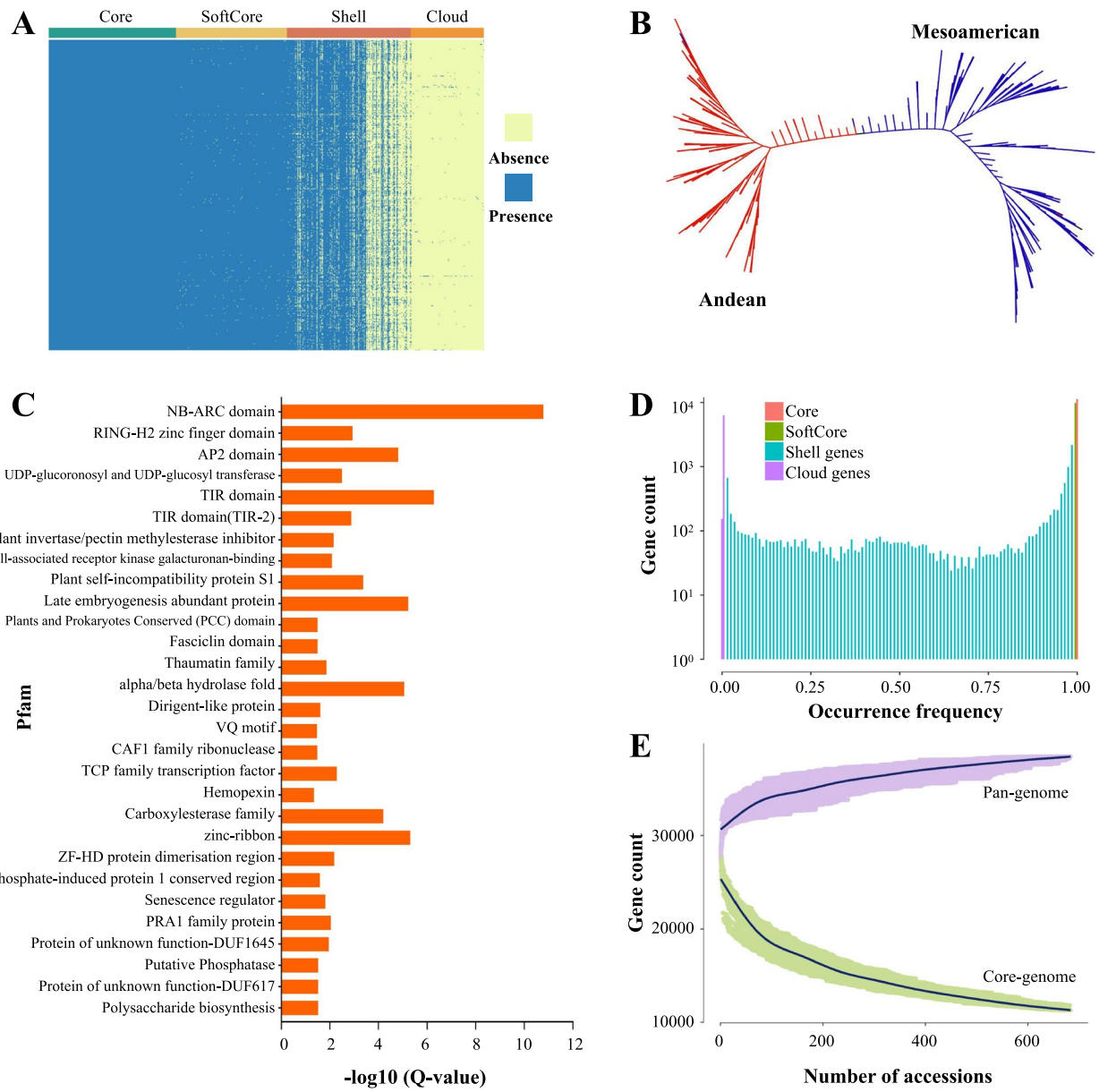

Fig. 1 Construction of the common bean pan-genome and presence-absence variations analysis. **A** PAV heatmap of core, softcore, shell, and cloud genes. **B** Phylogenetic tree construction based on binary PAV data. **C** Results of pfam enrichment analysis in shell genes. **D** Distribution of the number of core, softcore, shell, and cloud genes. **E** Trend of an increasing number of genes within the pan-genome and decreasing number of core genes in the pan-genome of common bean after increasing sample size

González et al. identified several QTL intervals for resistance to *Pseudomonas syringae pv. phaseolicola*, four of which contained RGAs. These four QTL intervals were located between 45.49 to 48.5 Mb for chr2, 45.7 to 48.3 Mb for chr8, 15.1 to 16.7 Mb for chr9, and 0 to 1.5 Mb for chr11, respectively. The numbers of RGAs located in these four QTL intervals were 16, 6, 3, and 11, respectively. Based on the variation information of RGAs within these QTL intervals, we found that the number and type of RGA variations were different. We ranked the effects of different variant types on genes, and the most influential variant type was gene loss; there were also stop gained, missense variants, synonymous variants, etc.

Within the QTL of chr2, the mutational load of the RGA genes varied greatly. For example, PHAVU_002G324000 g (SNRPD2, small nuclear ribonucleoprotein D2), PHAVU_002G301900 g (Pkinase_Tyr, Protein tyrosine kinase), and PHAVU_002G3235001

Table 1 The number of different types of resistance gene analog (RGA) candidates and subfamilies found on the reference genomes and pan-genome additional contigs

RGAs	Reference	Pangenome additional contigs	Pangenome
CN	13 (12, 1)	36 (36, 0)	49 (48, 1)
CNL	172 (99, 73)	81 (81, 0)	253 (180, 73)
NBS	7 (4, 3)	40 (40, 0)	47 (44, 3)
NL	99 (50, 49)	49 (49, 0)	148 (99, 49)
RLK	807 (128, 679)	36 (35, 1)	843 (163, 680)
RLP	132 (60, 72)	51 (51, 0)	183 (111, 72)
TMCC	159 (16, 143)	28 (28, 0)	187 (44, 143)
TN	15 (8, 7)	6 (6, 0)	21 (14, 7)
TNL	73 (30, 43)	6 (6, 0)	79 (36, 43)
TX	31 (21, 10)	40 (40, 0)	71 (61, 10)
OTHER	21 (6, 15)	0 (0, 0)	21 (6, 15)
Total	1529 (434,1095)	373 (372,1)	1902 (806,1096)

g (TMVRN_NICGU, TMV resistance protein N) showed minor variation across the population, indicating that they are conserved genes. The genes PHAVU_002G323200 g (RPP1, Probable disease resistance protein) and PHAVU_002G323100 g (SNC1, Protein SUPPRESSOR OF npr1-1, CONSTITUTIVE 1) had extremely high mutation loads, and the variant types were mainly gene loss and missense variants, indicating that these genes were under strong positive selection pressure. There were also differences between samples; for example, samples on the right of Fig. 2B missed the genes PHAVU_002G323800 g (TMVRN_NICGU, TMV resistance protein N), PHAVU_002G323400 g (TAO1, Disease resistance protein), PHAVU_002G323300 g (TMVRN_NICGU, TMV resistance protein N), and PHAVU_002G323200 g (RPP1, Probable disease resistance protein), whereas these genes in other samples were not missed. This suggests that the variant types in all the genes were different in all the samples. The mutational load of RGAs in the QTL intervals on chr8 and chr9 was less than that of the QTL interval on chr2 (Fig. 2C); however, the types of variation in RGAs in this interval were mainly missense variants, indicating positive selection in half of the samples. In the QTL interval on chr11, gene loss was the main variant type of PHAVU_011G014400 g (RPP13, Disease resistance protein), PHAVU_011G014500 g (RPP13, Disease resistance protein), and PHAVU_011G008100 g (MIK2, MDIS1-interacting receptor like kinase 2) in the population; however, the variant types of the other genes in this interval were mainly intron variants (Fig. 2D). The above results chiefly revealed the population variation of RGAs in the QTL interval for resistance in *Pseudomonas syringae pv. phaseolicola* and provided information for further screening of candidate genes.

Protease inhibitor gene clusters in the common bean genome

In plants, PIs are involved in defence response to phytophagous and pathogen infestation; therefore, in this study, we analysed the PI genes of the common bean. A total of 152 candidate PI genes were identified within the reference genome of common bean. Among them, the number of genes with serpin (PF00079), trypsin and PI (PF00197), Bowman-Birk serine PI family (PF00228), potato inhibitor I family (PF00280), peptidase inhibitor I9 (PF05922), inhibitor_I29 (PF08246), and aspartic acid proteinase inhibitor (PF16845) domains was 1, 26, 5, 3, 69, 37, and 11, respectively. By scanning the common bean genome with a sliding window size of 150 kb, we identified 10 genes containing the cathepsin propeptide inhibitor domain (PF08246) in the 29.4 Mb to 30.2 Mb interval of chr11.

Synteny analysis of the genomes of *V. unguiculata*, *G. max*, and *A. hypogaea* revealed that the cluster of PI genes present in the common bean genome was different in other legumes (Fig. 3A). The synteny region of the *V. unguiculata* genome was located at 26.81–27.41 Mb of chr11, and only four PI genes of the common bean were homologous to genes within this region in *V. unguiculata*. Tandem duplication of genes occurred in the homologues of PvCPi in *V. unguiculata*. Two PI genes in this genomic region of *V. unguiculata* were homologous to *G. max* genes. As shown in the synteny plot, tandem duplication of multiple PI genes occurred within 46.71–46.54 Mb of chr6 of *G. max*, with a total of seven PI genes. There was only one PI gene in the corresponding synteny region of *A. hypogaea*. These phenomena suggest that the gene clusters identified in common beans may arise after species divergence from other legumes. In the 683 common bean samples, the absence rate of PI genes was calculated in the gene cluster, and only PHAVU_011G131700 g (SAG39, Senescence-specific cysteine protease) and PHAVU_011G132000 g (SAG39, Senescence-specific cysteine protease) had high gene loss rates within this gene cluster (Fig. 3B). This suggests that the PI gene PAV was not under strong selection during domestication and breeding improvement.

Stress response atlas of common bean

The study of biotic and abiotic stresses is important for all crops. RNA-seq data for a wide range of common beans in response to biotic and abiotic stresses were collected from NCBI (Table S1). First, we selected transcriptome sequencing data that had three biological replicates and were all based on the pair-end sequencing strategy of the

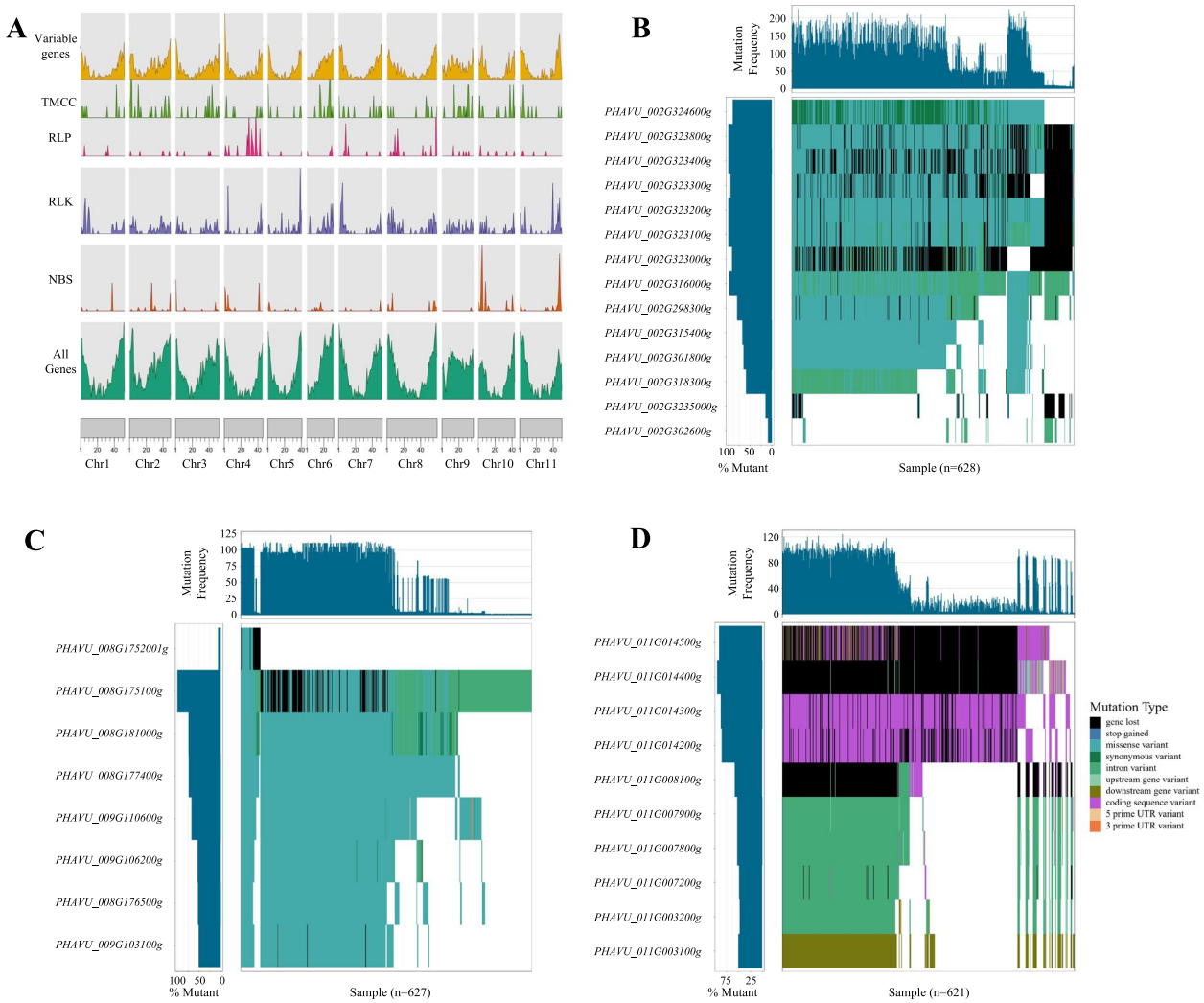

Fig. 2 Identification and analysis of RGA genes within the common bean pan-genome. **A** The distribution density of identified RGA genes on 11 chromosomes of common bean. **B**-**D**, Mutation analysis of RGA genes within the QTL interval for resistance to *Pseudomonas syringae pv. phaseolicola* located on chr2 (B), chr8 and chr9 (C), chr11 (D)

Illumina sequencing platform. These data included the biotic stresses of *M. incognita* infestation, *Xanthomonas* infestation, *arbuscular mycorrhizal* fungi infestation and *Xanthomonas axonopodis* infestation and the abiotic stresses of salt stress, low temperature, and high CO_2 concentration. After calculating the mean FPKM values for all the groups, the samples were classified using PCA. The first principal component (PC1) (explained 50.38% of variance) and second principal component (PC2) (explained 15.86% of variance) were able to distinguish between different tissues and stress types (Fig. 4B). This suggests that genes are specifically expressed in different tissues, growth periods, and growth environments. The heatmap of expression showed that different tissues had specific high and low gene expression (Fig. 4A). Among the abiotic stresses, 2685 response genes were common to salt stress, low-temperature stress, and drought stress, and a large number of response genes were stress-specific. For example, 3864 response genes were specific to low-temperature stress. There were only 38 response genes that were common to different sources of pathogen infestation, majority of which were specific (Fig. 4C). Hence, the gene response patterns provide a reference for further mining of broad-spectrum and specific resistance genes.

Based on PAV analysis, we found that the core, soft-core, shell, and cloud genes had different expression levels and responded to multiple stresses. The core genes had the highest expression level, the softcore gene had a similar expression level to that of the core

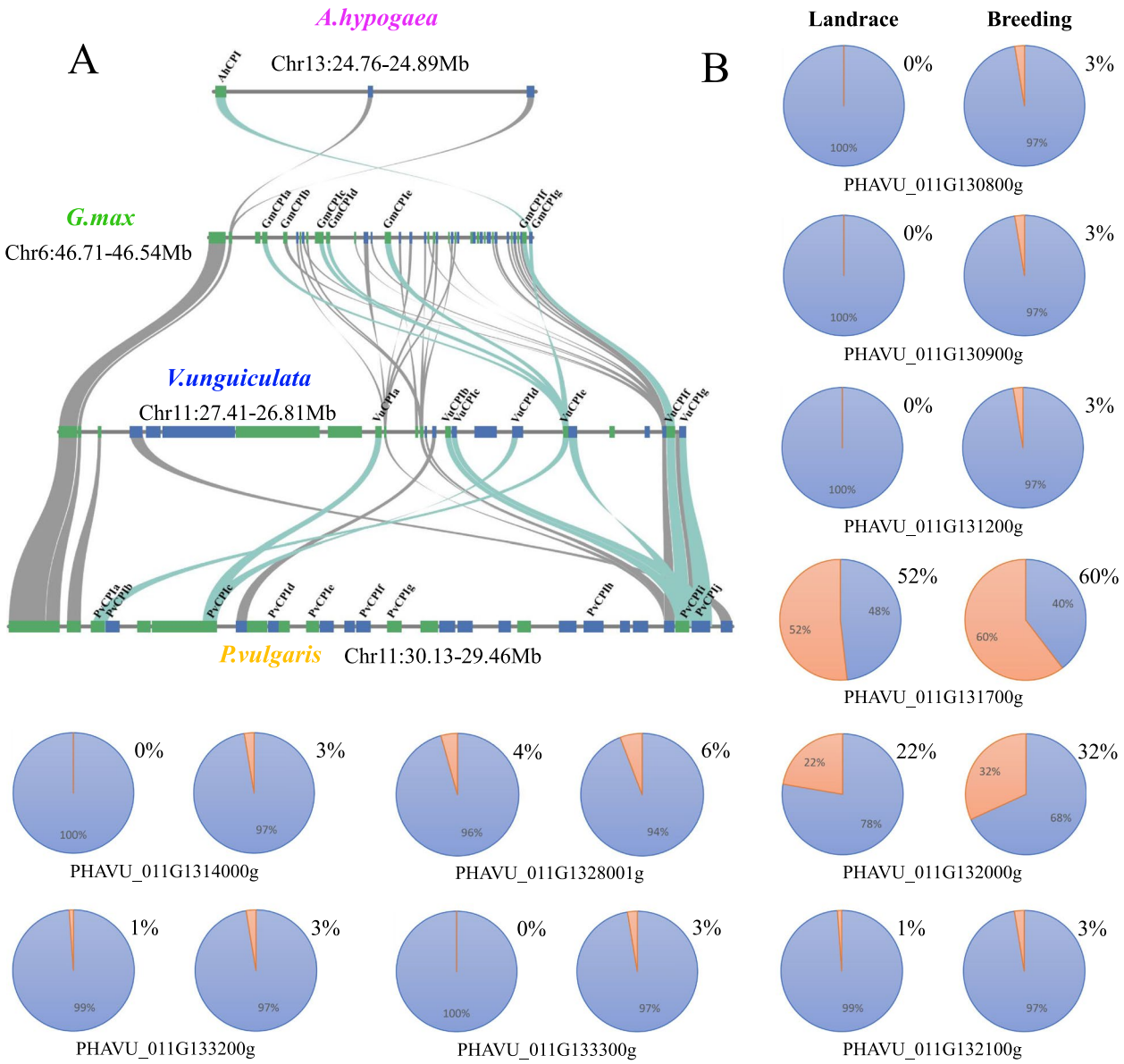

Fig. 3 Analysis of protease inhibitor gene clusters based on the reference genome of common bean. **A** Analysis of synteny between the protease inhibitor gene cluster on chromosome 11 of *Phaseolus vulgaris* L. and *V. unguiculata* as well as *G. max* and *A. hypogaea*. **B** PI genes within the protease inhibitor gene cluster on the common bean genome in the landrace and breeding lines

gene, the shell gene had a significantly lower expression level than that of the core and softcore genes, and the cloud gene had an extremely low expression level (Fig. 4F). These results suggest that core and softcore genes play an important role in maintaining the essential life activities of the common bean at a high level of expression. Core and softcore genes are not only expressed at high levels overall but also play a key role in biotic and abiotic stresses, accounting for the majority of stress-responsive genes (Fig. 4E). This suggests that most of the important functional genes are conserved, and only a few genes with PAV may be involved in shaping the phenotypic diversity of the different varieties. In the present study, analysis of the BHLH gene family in the common bean revealed that certain similar genes in one family were either core, softcore, or shell genes. As shown in Fig. 4D, these five gene family members were co-expressed, suggesting that they had similar expression patterns and functional complementarity.

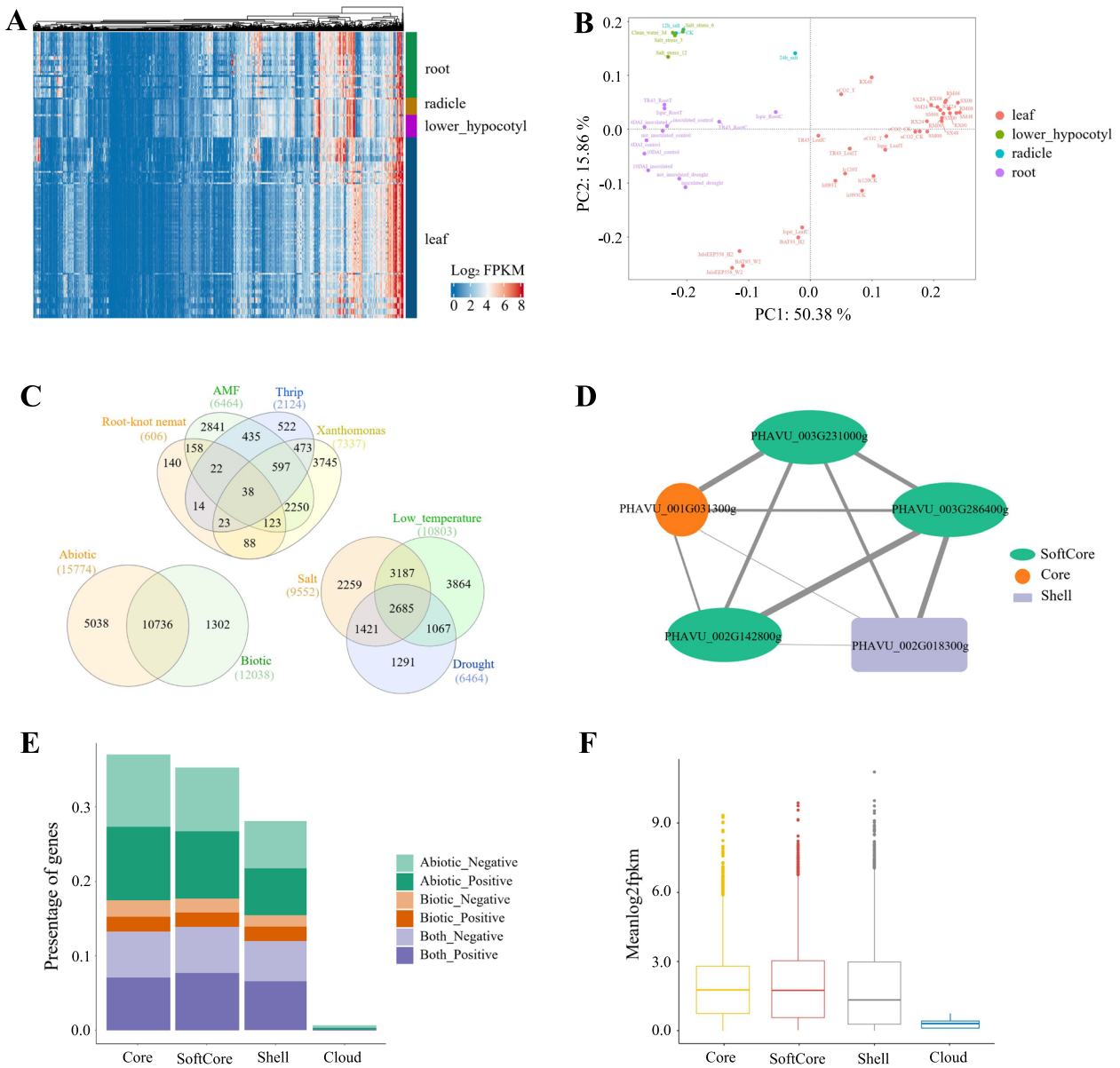

Fig. 4 Response atlas of biotic and abiotic stresses in common bean. **A** Expression heatmap of differentially expressed genes in each experimental design for stress response. **B** Results of PCA analysis based on average FPKM values of samples from each group. Colours represent different tissues of the common bean. **C** Venn plot of differentially expressed genes in the common bean under biotic and abiotic stresses. **D** Co-expression network of five genes in the bHLH gene family. The shape of the nodes represents the type of gene (core, softcore, and shell genes), and the thickness of the connecting lines represents the degree of correlation between the two genes. **E** Proportion of core, softcore, shell, and cloud genes response to different biotic and abiotic stresses. The types of response are classified as positive (upregulated in stress samples) and negative (downregulated in stress samples). **F** Box plot of the expression level of core, softcore, shell, and cloud genes in the common bean genome (FPKM). Each point represents the mean value of the expression of the gene in one group

The gene expression module of common bean is functional-, organ-, and PAV-specific

The DEGs in response to each stress were analysed using WGCNA, and a total of 14 co-expression modules were identified. Ten of the 14 modules were organ specific (Fig. 5). The blue module was specifically expressed in both leaves and radicles, and the genes in this module were enriched in photosynthesis (Q-value =9.6e-53), which is closely related to leaf function. The genes in this module were also enriched in shell genes, suggesting that these genes may have been selected during domestication. Genes in the red, magenta, tan, salmon, turquois,

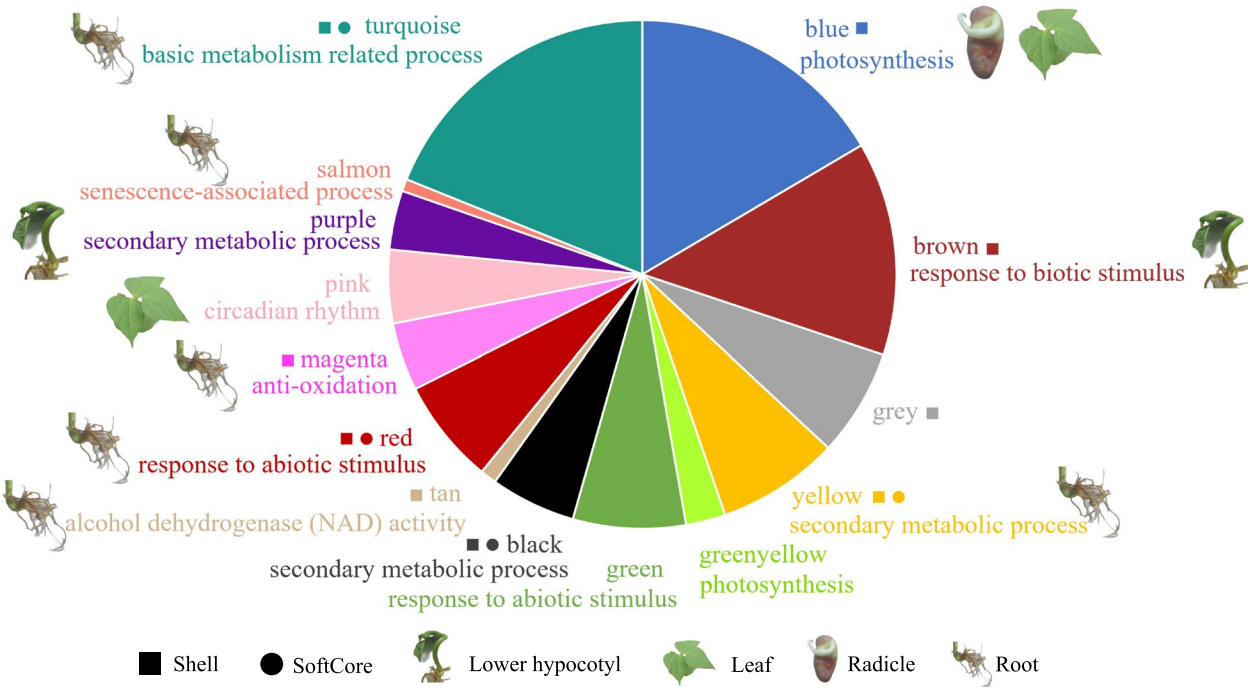

Fig. 5 Gene expression modules were obtained from co-expression analysis using WGCNA. Each colour represents an expression module. The figure of the organ near the module indicates that the genes within the module are specific to the expression of that organ. The squares and circles next to the modules indicate that the genes within the module are significantly enriched in shell genes or softcore genes. In addition, the functional descriptions next to the module represent the GO terms for the gene enrichment within the module

and yellow modules were expressed specifically in the root. In addition to the salmon module, the other modules were enriched in shell genes. GO enrichment analysis revealed that genes in the red, tan, salmon, and yellow modules were enriched in response to abiotic stimuli, alcohol dehydrogenase activity, senescence-associated vacuoles, and secondary metabolic processes, which are involved in stress responses.

A series of similar GO terms were enriched in magenta or turquoise modules; for example, many anti-oxidation-related GO terms were enriched in the magenta module, and some basic metabolism-related processes were enriched in turquoise. Genes in the purple and brown modules were specifically expressed in the lower hypocotyl, and genes in the brown module were enriched in response to biotic stimuli. The above results indicate that genes in different expression modules have tissue specificity as well as functional and conservation specificity.

PAV is prevalent in pathways

We performed pathway annotation based on the KEGG database for all pan-genome genes and assigned 29,462 genes to 410 pathways. Of the genes that could be assigned to pathways, 9579 were core genes, 7136 were softcore genes, 6951 were shell genes, and 5796 were cloud genes. Many varieties have lost some reference genes on some nodes of the pathway but have gained some new genes during domestication, leaving some nodes of the pathway with many homologous genes located on pan-genome additional contigs. There were 21 pathways with no variable gene expression, which indicates that these pathways are conservative and that the process of artificial selection does not exert selection pressure on these pathways. The peroxidase (K00430) in the phenylpropanoid biosynthesis (map00940) pathway has the most paralogous homologs (60) on pan-genome additional contigs, and this phenomenon of intra-species expansion of family members may be related to the selection pressure caused by the process of breeding for specific traits in common beans.

Eight selected pathways had high average gene absence rates, with genes only in the top 15 absence rates in each pathway (Fig. 6). Genes may have high absence rates owing to the fact that genomes of some samples gaining non-reference novel genes have very low gene frequencies in the population, resulting in a high average absence rate at one node of the pathway. The results showed that the PAV of the genes in each pathway and the expansion of the family had multiple patterns. For example, ALDO in glycolysis and gluconeogenesis; although the average gene absence rate of all homologues of this gene was 88.2% in 683 samples, 31 of its homologues were located

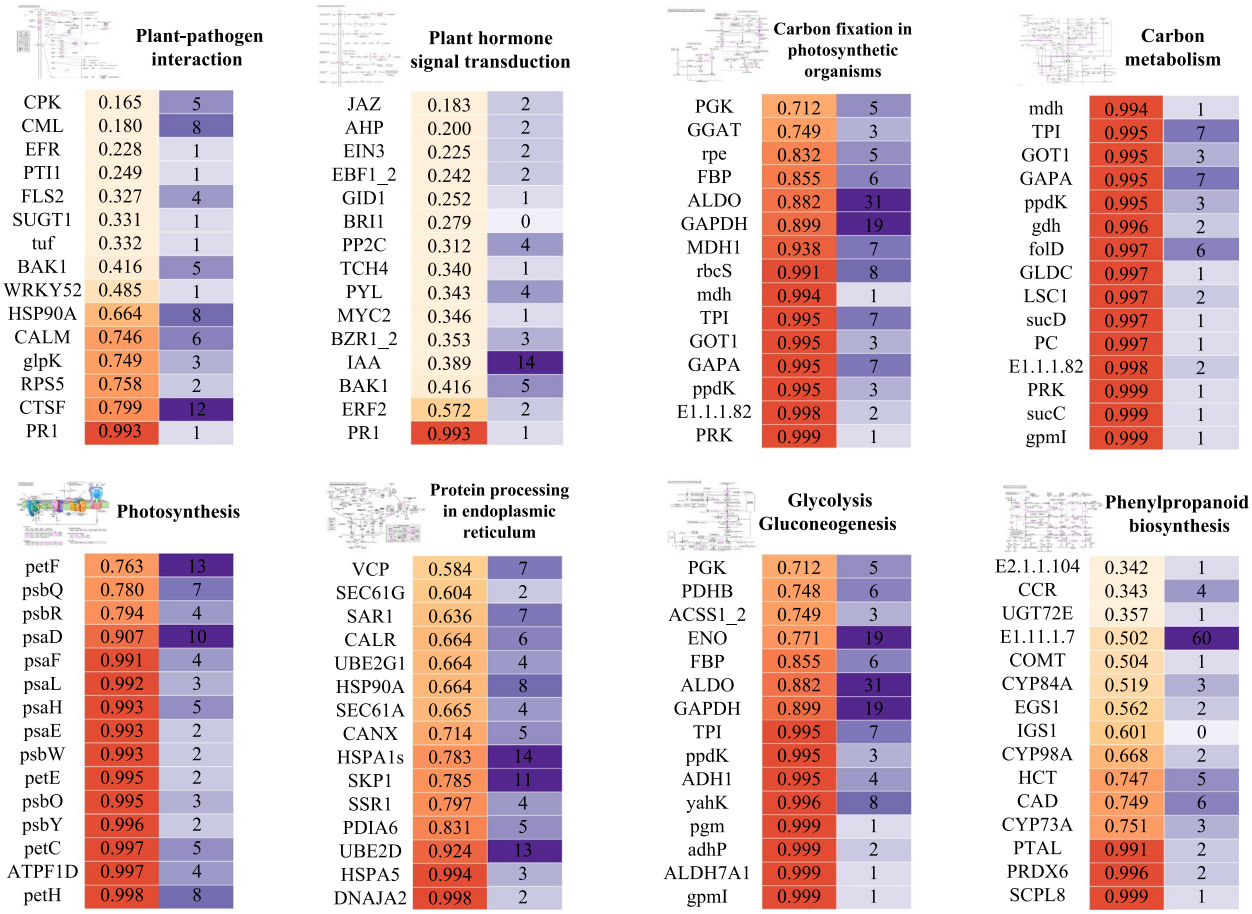

Fig. 6 Absence rates of genes on each KO node and the number of genes on non-reference in each of the eight pathways selected. The first column is the abbreviated name of the protein for that KO, the second column indicates the absence rate of the homologous gene encoding the protein in 683 individual beans, and the third column indicates the number of homologous genes encoding the protein located on non-reference contigs. The genes with the top 15 absence rates in each pathway are shown in the figure, and information on the other complete gene sets is illustrated in table S6. Abbreviations of all the proteins in this figure are defined in the attached table S7

in non-reference contigs. All 31 genes were shell genes, indicating that these expansions occur only in some varieties. The gene absence rate of the top 15 genes in both pathways, photosynthesis and carbon metabolism, was very high at over 76.3%. However, the top 15 genes in photosynthesis were all photosystem-related and may have some functional complementarity, whereas the top 15 genes in carbon metabolism belonged to a different gene family. Among the eight pathways in Fig. 6, the genes in two pathways, plant-pathogen interaction and plant hormone signal transduction, were relatively conserved, with an average gene absence rate of 16.5% and 18.3% for homologues of the 15 th gene.

Mutual exclusivity and co-occurrence of gene PAV within and between pathways

Based on the pan-genome constructed in this study, we were able to identify an abundance of gene PAV, and finding the type of relationships that exist between these variants is a research worthy of further study. We calculated the significance of mutual exclusivity and co-occurrence between gene pairs in each pairing pattern within and between all the pathways. Using a p-value of < 0.05 as the threshold, we identified a total of 8990 pairs of genes with mutual exclusivity and 30,272 pairs of genes with co-occurrence of PAVs within the pathway (Table S8, 9). A total of 136,887 and 314,888 pairs of genes were identified as mutually exclusive and co-occurring between pathways, respectively (Table S10, 11). There were both mutual exclusivity and co-occurrence gene PAVs within the same pathway; for example, 56 and 43 pairs of mutual exclusivity and co-occurrence gene PAVs in flavonoid biosynthesis, respectively (Fig. 7A). Some pathways have more mutually exclusive gene pairs than co-occurrence gene pairs, such as flavonoid biosynthesis, and some have more co-occurrence

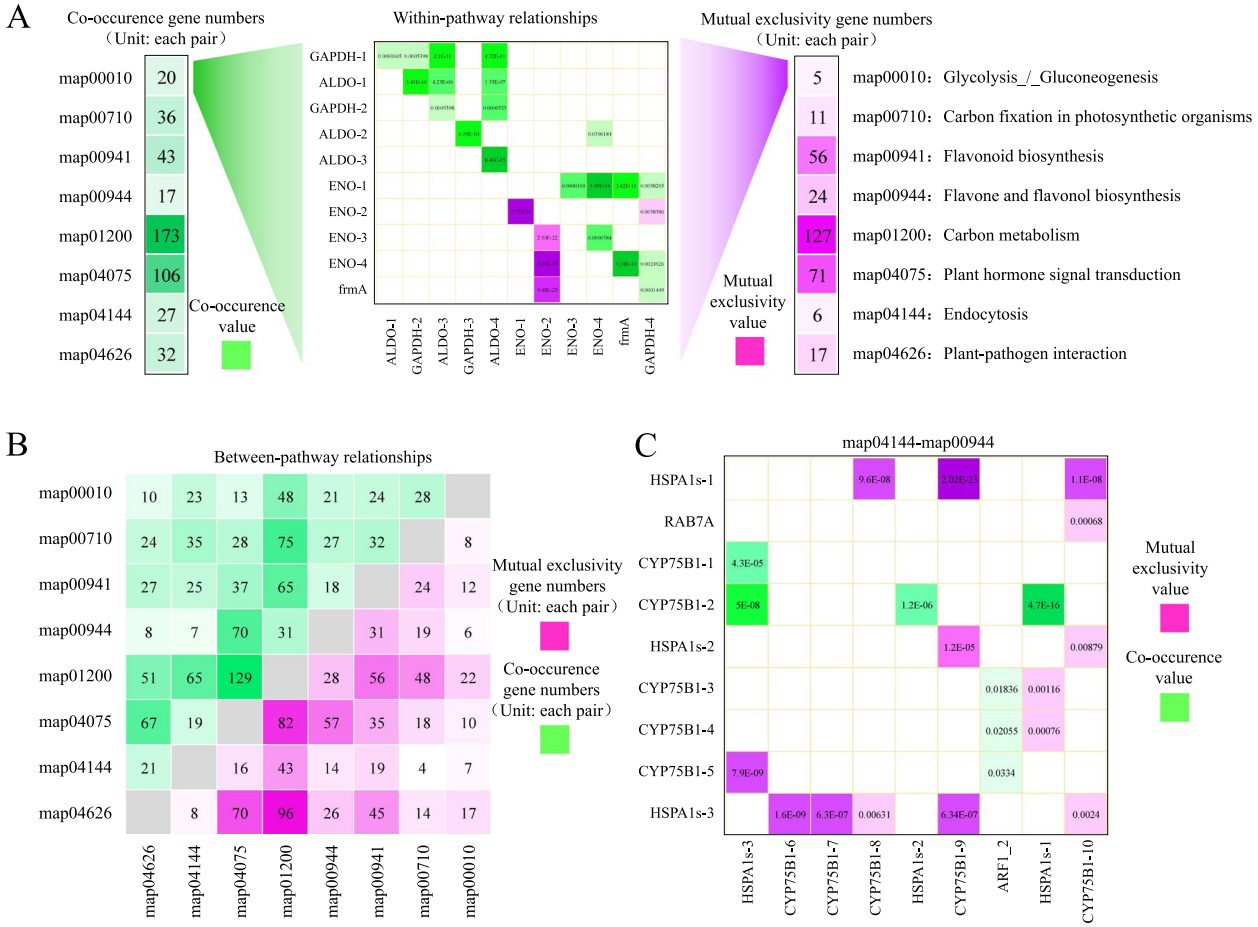

Fig. 7 Mutual exclusivity and co-occurrence of gene PAVs within and between pathways calculated based on PAV information in 683 samples of common bean. **A** Mutual exclusivity and co-occurrence of genes present within pathways. The numbers in the figure represent the number of gene pairs. The middle figure is the p-value of the mutual exclusivity and co-occurrence between genes in phenylalanine metabolism (map00360). **B** Mutual exclusivity and co-occurrence of genes between pathways. The numbers in the figure represent the number of gene pairs. **C** The p-value of the mutual exclusivity and co-occurrence of gene PAVs between Glycolysis/Gluconeogenesis (map00010) and flavonoid biosynthesis (map00941) pathways

gene pairs than mutual exclusive gene pairs, such as the plant-pathogen interaction, where there are 32 mutual exclusivity PAV pairs and 17 co-occurrence gene PAV pairs. Many gene pairs belong to the same gene family, such as the ALDO and ENO families of the glycolysis/gluconeogenesis pathway. These family members with mutual exclusivity and co-occurrence gene PAVs were close in the phylogenetic tree (Figure S1), suggesting that their functions were similar. This may reveal some functional gene complementation in common bean during domestication. Mutual exclusivity and co-occurrence gene PAVs were not only found within pathways but were also widespread between pathways (Fig. 7B); for example, between the endocytosis and flavone and flavonol biosynthesis pathways, seven co-occurrence and 14 pairs of mutual exclusivity gene PAVs were present (Fig. 7C), which reveals the complexity of biological activity. Furthermore, different pathways need to be linked to study the patterns of genetic mutations.

Discussion

The genome of a single strain does not represent the entire set of genes within a species; hence, reference genomes cannot be used to study the complete variation in a given population. Therefore, many studies have constructed pan-genomes using second or third generation population sequencing. The construction of a large number of linear and graph-based genomes has greatly facilitated the discovery of structural variation and gPAV for crop and animal researches [11, 17, 48]. Based on domestication and geographical isolation, the common bean has formed two centres of diversity for cultivated common beans in the Andes and Central America [49]. To

investigate the genetic diversity of these common bean lines, 683 common beans were sequenced by Wu et al. [3], which contained landraces and breeding lines from different localities. These samples contained abundant variation information, and the pan-genome constructed based on these samples contained 10,452 non-reference genes, which are important for candidate gene mining. Of the 38,577 genes in the common pan-genome, 11,024 and 6454 were shell and cloud genes, respectively, suggesting that there is a rich diversity of genes lost and obtained in the bean population owing to domestication or genetic drift. The Andean and Mesoamerican populations of common bean were separated in the phylogenetic tree based on the PAV information. This indicates that the PAV information identified in this study is accurate and can also be found in other pan-genomic studies such as *Brassica napus* and tomato [16, 22], which are important for further analyses.

Achieving greater resistance is a common goal in all crop breeding programs, and RGAs are important candidate genes for resistance breeding. Sufficient RGAs can be found in the pan-genome on further analysis using a robust RGA-identification pipeline [19, 20, 50]. The RGAs located on non-reference contigs identified in this study are an important addition to the resistance gene resources of the common bean. Of the 373 novel RGAs, 372 were variable genes (Table S12), suggesting that a large number of resistance genes were lost and gained during domestication and geographic isolation. During the breeding process that focuses on one trait, the genome is likely to lose genes associated with other traits [51] and may gain new genes owing to structural variation. Owing to the domestication of common beans in different regions with different breeding objectives, the genome of G19833 in the Andean pool is no longer available as a reference for all RGAs; additionally, 434 of the 1529 RGAs in the G19833 genome are variable genes. Regarding resistance, an important agronomic trait, numerous quantitative genetic studies have identified QTL for resistance, such as anthracnose resistance, bacterial blight resistance, and *Pseudomonas syringae pv. phaseolicola* resistance [40, 52, 53]. Many QTL studies have identified several genes within the interval, and resolving the variation in these genes can provide new insights into the selection of multiple candidate genes. For example, the 36 RGAs identified in the present study in *Pseudomonas syringae pv. phaseolicola* resistance QTL intervals had different variation patterns, including gene absence, which may help to understand the mutation load and impact of mutations in different RGAs. RGAs with higher mutation loads and gene loss, stop-gain, and missense variants may have been under higher selection pressure, and their variation may impact the resistance of different varieties.

Since plants are subject to infestation by pathogens, parasitic plants, and herbivores during growth, plant-derived PIs are promising defences for crop improvement and pest management [54]. Of the 152 candidate PI genes identified in this study, 10 cysteine proteinase inhibitor genes were clustered by tandem or segmental replication, which may play an important role in plant resistance [55, 56]. Comparative genomic analyses of common beans and other legumes suggest that the formation of the cysteine proteinase inhibitor gene cluster in common beans occurred after they diverged from a common ancestor. This may be related to the different biotic stresses encountered by different legumes, as different PIs may target different biotic stressors. Absence rates were very low for eight of the 10 genes in the gene cluster but were higher in all breeding lines than in landraces, suggesting that most breeding goals may have overlooked selection for PI gene-related traits. Two PI genes with high absence rates, PHAVU_011G131700 g (SAG39, Senescence-specific cysteine protease) and PHAVU_011G132000 g (SAG39, Senescence-specific cysteine protease), were also abundant in the breeding lines than in the landraces. These two genes are candidates for further study in insect and disease resistance breeding.

RGAs and PI genes are associated with biotic stresses; however, abiotic stresses are also important for plants that need to be addressed during growth. In this study, transcriptomes generated from nine independent biotic and abiotic stresses were analysed to obtain a more comprehensive stress response atlas for common beans. This part of the analysis revealed not only the organ and stress specificity of gene expression but also the relationship between gene PAVs and the specific expression of genes. For example, some expression modules showed multiple dimensions of enrichment, including gene function, organ, and gene PAV, which provides more information for further mining and exploitation of the genetic resources of common bean. Using the co-expression network, functional complementarity of genes in one gene family can be studied; for example, expression correlation between shell and core genes within the BHLH family exists, which indicates that they have similar expression patterns. When a shell gene is absent in some individuals, the core gene can compensate for its function. If both shell and core genes are present, the function of this gene may be enhanced, which is important for the developmental robustness of the plant and the alteration of traits owing to the enhanced functionality of a specific gene [57].

Pathways are an important avenue for biological research that integrates genes into a system. Variations in one node may affect the action of the entire pathway. To study gene mutations in pathways, we analysed the gene PAVs of all pathways and found inconsistencies in conservation between pathways. Gene family expansions are common in plants and can affect the synthesis of secondary metabolites or other pathways [58]. However, most gene family expansion studies are based on reference genome comparisons of each species, which can reveal species-specific gene family expansions [59]. Several new genes can be annotated in pan-genomes, providing a basis for studying the expansion of gene families within a species. For example, many RGAs not present in the reference genome have been identified in the pan-genomes of both *Brassica oleracea* and *Brassia napus*, providing additional resources for mining functional genes [19, 20]. Our analysis revealed a large number of non-reference novel genes in pathways associated with plant resistance and secondary metabolite synthesis. Majority of these new genes are variable genes, suggesting that their replication has only occurred in some varieties. For example, phenylpropanoid biosynthesis—a pathway important for growth and development, stress response, and nutritional value [60, 61]—has 60 novel peroxidase (E1.11.1.7) genes on non-reference contigs. The duplication of genes often leads to divergence in gene function [62], and a large number of gene duplications in peroxidase may result in unique traits in many varieties. Further analysis of these gene families could lead to more possibilities for common bean breeding.

In tumours, the mutual exclusivity and co-occurrence of somatic mutations reflect functional interactions between genes [63]. Several tumour driver mutations exhibit mutual exclusivity between them, possibly because of the redundancy of functions between these genes [64]. Co-occurring mutations represent the possibility that mutations in these genes may simultaneously activate different pathways and alter the phenotype, reflecting collaboration between gene functions [65]. However, the study of mutual exclusivity and co-occurrence should not be limited to the field of tumour research, as the large number of gene PAVs identified in the present study also requires the analysis of their relationships. By calculating the mutual exclusivity and co-occurrence between gene PAVs in the pathway, we found that a large number of mutually exclusive or co-occurrence relationships occurred between different members within a single gene family. In the stress response atlas section of this study, we found the co-expression of different members of the BHLH family with both core and shell genes. The combination of these two components suggests that members of the gene family have complex collaborative relationships with each other, as they participate in plant growth, development, and stress response. A study by Kwon et al. [66] on stem cell development in plant stem tips found that mutations in the stem cell regulator *CLV3* caused an excessive proliferation of floral organs in many plants, and a paralogous homolog of the *CLV3* gene in tomato was able to partially suppress this abnormal phenotype. This suggests that many genes function more robustly in the presence of a 'spare gene'. However, these mutually exclusive genes may cause significant phenotypic changes if both the genes are absent; thus, the mutually exclusive PAV gene family members may contain such 'spare genes'. This provides a new perspective on the expansion of genes in a pan-genomic context and the redundancy of functions between these genes as well as may aid in certain researches, such as gene editing. Furthermore, if editing of target genes is performed to study their function, it is best to consider these genes with mutually exclusive PAV relationships. In contrast, co-occurrence of the gene PAVs provides alternative perspective on the results, where different genes or pathways have to be altered simultaneously to undergo a phenotypic change when the plant is adapted to different environments or under different selection pressures. This relationship between mutations has been further studied in tumours and, to a lesser extent, in plants; for example in *Amaranthus palmeri,* where G399 A was found to be most likely to co-occur with other ppo2 mutations in the same allele [67]. This suggests that there are many complex links between mutations in gene pairs and that the functions of genes may be multifaceted and may be involved in different pathways. When a candidate gene is identified in a population or an individual strain that forms a great phenotype, analysis of the genes and pathways that have a co-occurring PAV relationship is a great reference for a deeper understanding of gene function and better breeding work. In conclusion, this study is the first to construct a common bean pan-genome, providing a richer genetic resource for the study of common bean. It also provides a basis for further understanding of the gene PAV in common beans during domestication and breeding by examining various aspects of the relationship between insect resistance, stress resistance, and gene PAVs in the pathway.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11662-2.

Supplementary Material 1. Figure S1 Phylogenetic tree of ALDO (A) and ENO (B) gene families of common bean.

Supplementary Material 2. Figure S2 (A) Gene frequency in Andean and Mesoamerican. (B) KEGG enrichment analysis of the genes significantly different frequencies in Andean and Mesoamerican.

Supplementary Material 3. Table S1 List of the SRA number of RNA-seq data used for stress response atlas analysis of common bean.

Supplementary Material 4. Table S2 List of the SRA number of RNA-seq data used for gene structure annotation.

Supplementary Material 5. Table S3 Quality information of the 683 resequencing data used in this study.

Supplementary Material 6. Table S4 The PAV information of all genes in the pan-genome.

Supplementary Material 7. Table S5 Genes with significantly different frequencies in the population.

Supplementary Material 8. Table S6 Gene absence rates of all the genes in the eight selected pathways.

Supplementary Material 9. Table S7 Abbreviations of all the proteins in the eight selected pathways.

Supplementary Material 10. Table S8 Genes with mutual exclusivity PAVs within the pathways.

Supplementary Material 11. Table S9 Genes with co-occurrence of PAVs within the pathways.

Supplementary Material 12. Table S10 Genes with mutual exclusivity PAVs between the pathways.

Supplementary Material 13. Table S11 Genes with co-occurrence of PAVs between the pathways.

Supplementary Material 14. Table S12 The RGA candidates identified in the pan-genome.

Authors' contributions
XW, YS and XDK conceived and designed the experiments. XW, SSC, YS and XDK contributed to paper writing. FL, MY, QNW conducted the experiment. QQZ, BJ and YXH contributed to the data analysis. All authors contributed to the article and approved the submitted version.

Funding
This work was supported by the National Natural Science Foundation of China (Grant No.32100352, 32100355, 31871964), the Natural Science Fund of Anhui Province (Grant No.1908085QC93), the Natural Science Foundation of Universities of Anhui Province (Grant No.KJ2020 A0094), the Major Science and Technology Projects in Anhui Province (Grant No.202003a06020009) and the National Science & Technology Fundamental Resources Investigation Program of China (Grant No.2019 FY101800).

Data availability
The Common bean pangenome assembly is available at National Genomics Data Center (BioProject: PRJCA038428; https://ngdc.cncb.ac.cn/gwh/Assembly/92566/show). The Common bean pangenome assembly and annotation are also available at Figshare database (https://figshare.com/articles/dataset/Common_bean_pangenome/21154846). If you are interested in the relevant data generated by this study, you may also request it from the corresponding author.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 January 2024 Accepted: 1 May 2025
Published online: 16 May 2025

References

1. Broughton WJ, Hernández G, Blair M, Beebe S, Gepts P, Vanderleyden J. Beans (Phaseolus spp.) – model food legumes. Plant Soil. 2003;252:55–128.
2. Joshi PK, Rao PP. Global pulses scenario: status and outlook. Ann N Y Acad Sci. 2017;1392:6–17.
3. Wu J, Wang L, Fu J, Chen J, Wei S, Zhang S, et al. Resequencing of 683 common bean genotypes identifies yield component trait associations across a north–south cline. Nat Genet. 2020;52:118–25.
4. Xie D, Xu Y, Wang J, Liu W, Zhou Q, Luo S, et al. The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. Nat Commun. 2019;10:5158.
5. Kang L, Qian L, Zheng M, Chen L, Chen H, Yang L, et al. Genomic insights into the origin, domestication and diversification of *Brassica juncea*. Nat Genet. 2021;53:1392–402.
6. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46:707–13.
7. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019;51:30–5.
8. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim HR, Martinez PA, et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun. 2016;7:1–8.
9. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet. 2018;50:278–84.
10. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-Genome of wild and cultivated soybeans. Cell. 2020;182:162–176.e13.
11. Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell. 2021;184:3542-3558.e16.
12. Hu Z, Sun C, Lu K, Chu X, Zhao Y, Lu J, et al. EUPAN enables pan-genome studies of a large number of eukaryotic genomes. Bioinformatics. 2017;33:2408–9.
13. Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, et al. HUPAN: a pan-genome analysis pipeline for human genomes. Genome Biol. 2019;20:149.
14. Gui S, Wei W, Jiang C, Luo J, Chen L, Wu S, et al. A pan-Zea genome map for enhancing maize improvement. Genome Biol. 2022;23:178.
15. Liu C, Wang Y, Peng J, Fan B, Xu D, Wu J, et al. High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. Plant Commun. 2022;3:100352.
16. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet. 2019;51:1044–51.
17. Wang K, Hu H, Tian Y, Li J, Scheben A, Zhang C, et al. The Chicken pan-genome reveals gene content variation and a promoter region deletion in IGF2BP1 affecting body size. Mol Biol Evol. 2021;38:5066–81.
18. Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing heritability and empowers tomato breeding. Nature. 2022;606:527–34.
19. Bayer PE, Golicz AA, Tirnaz S, Chan CKK, Edwards D, Batley J. Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. Plant Biotechnol J. 2019;17:789–800.
20. Dolatabadian A, Bayer PE, Tirnaz S, Hurgobin B, Edwards D, Batley J. Characterization of disease resistance genes in the *Brassica napus* pangenome reveals significant structural variation. 2020;18:969–82.
21. Ou L, Li D, Lv J, Chen W, Zhang Z, Li X, et al. Pan-genome of cultivated pepper (Capsicum) and its use in gene presence–absence variation analyses. New Phytol. 2018;220:360–3.
22. Song J-M, Liu D-X, Xie W-Z, Yang Z, Guo L, Liu K, et al. BnPIR: *Brassica napus* pan-genome information resource for 1689 accessions. Plant Biotechnol J. 2021;19:412–4.
23. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
24. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.
25. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:1–14.

26. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
27. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7.
28. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinforma. 2004;Chapter 4:1–14.
29. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
30. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
31. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: A b initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34 WEB. SERV. ISS.:435–9.
32. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
34. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
35. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
36. Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, et al. HUPAN: A pan-genome analysis pipeline for human genomes. Genome Biol. 2019;20:1–11.
37. Li P, Quan X, Jia G, Xiao J, Cloutier S, You FM. RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. BMC Genomics. 2016;17:852.
38. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.
39. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
40. González AM, Yuste-Lisbona FJ, Godoy L, Fernández-Lozano A, Rodiño AP, De Ron AM, et al. Exploring the quantitative resistance to Pseudomonas syringae pv. phaseolicola in common bean (*Phaseolus vulgaris* L.). Mol Breed. 2016;36:166.
41. Skidmore ZL, Wagner AH, Lesurf R, Campbell KM, Kunisaki J, Griffith OL, et al. GenVisR: Genomic Visualizations in R. Bioinformatics. 2016;32:3012–4.
42. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:1–14.
43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
44. Putri GH, Anders S, Pyl PT, Pimanda JE, Zanini F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics. 2022;38:2943–5.
45. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin P, O'Hara B, et al. Package "vegan" - Community Ecology Package. R News. 2015;8:48–50.
46. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
47. Ferrer-Bonsoms JA, Jareno L, Rubio A. Rediscover: an R package to identify mutually exclusive mutations. Bioinformatics. 2021.
48. Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat Plants. 2019;5:54–62.
49. Bitocchi E, Rau D, Bellucci E, Rodriguez M, Murgia ML, Gioia T, et al. Beans (*Phaseolus ssp.*) as a model for understanding crop evolution. Front Plant Sci. 2017;8.
50. Sun Y, Wang J, Li Y, Jiang B, Wang X, Xu W-H, et al. Pan-genome analysis reveals the abundant gene presence/absence variations among different varieties of melon and their influence on traits. Front Plant Sci. 2022;13:420.
51. Tieman D, Zhu G, Resende MFRJ, Lin T, Nguyen C, Bies D, et al. A chemical genetic roadmap to improved tomato flavor. Science. 2017;355:391–4.
52. Chen M, Wu J, Wang L, Mantri N, Zhang X, Zhu Z, et al. Mapping and genetic structure analysis of the anthracnose resistance locus Co-1HY in the common bean (Phaseolus vulgaris L.). PLoS One. 2017;12:1–18.
53. Miklas PN, Smith JR, Hang AN, Grafton KF, Kelly JD. Release of navy and black bean germplasm lines USNA-CBB-1, USNA-CBB-2, USNA-CBB-3, USNA-CBB-4 and USBK-CBB-5 with resistance to common bacterial blight. Annu Rep Bean Improv Coop Bean Improv Coop. 2001;44:181–2.
54. Sujata Singh Archana Singh SKPMIKS. Protease inhibitors: recent advancement in its usage as a potential biocontrol agent for insect pest management. Insect Sci. 2020;27:186–201.
55. Meng F, Li Y, Liu Z, Wang X, Feng Y, Zhang W, et al. Potential molecular mimicry proteins responsive to α-pinene in bursaphelenchus xylophilus. Int J Mol Sci. 2020;21:982.
56. Saify Nabiabad H, Amini M, Kianersi F. Ipomoea batatas: papain propeptide inhibits cysteine protease in main plant parasites and enhances resistance of transgenic tomato to parasites. Physiol Mol Biol Plants. 2019;25:933–43.
57. Huang X, Xiao N, Zou Y, Xie Y, Tang L, Zhang Y, et al. Heterotypic transcriptional condensates formed by prion-like paralogous proteins canalize flowering transition in tomato. Genome Biol. 2022;23:78.
58. Xu Z, Pu X, Gao R, Demurtas OC, Fleck SJ, Richter M, et al. Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. BMC Biol. 2020;18:63.
59. Zhang J, Fu X-X, Li R-Q, Zhao X, Liu Y, Li M-H, et al. The hornwort genome and early land plant evolution. Nat Plants. 2020;6:107–18.
60. Zhang Q, Li M, Xia CY, Zhang WJ, Yin ZG, Zhang YL, et al. Transcriptome-based analysis of salt-related genes during the sprout stage of common bean (Phaseolus vulgaris) under salt stress conditions. Biotechnol Biotechnol Equip. 2021;35:1086–98.
61. Valdés-López1 O, Hernández G. Phenylpropanoids as master regulators: state of the art and perspectives in common bean (Phaseolus vulgaris). Front Plant Sci. 2014;5:336.
62. Qiao X, Yin H, Li L, Wang R, Wu J, Wu J, et al. Different modes of gene duplication show divergent evolutionary patterns and contribute differently to the expansion of gene families involved in important fruit traits in pear (*Pyrus bretschneideri*). Front Plant Sci. 2018;9:161.
63. Canisius S, Martens JWM, Wessels LFA. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. Genome Biol. 2016;17:1–17.
64. Remy E, Rebouissou S, Chaouiya C, Zinovyev A, Radvanyi F, Calzone L. A modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. Cancer Res. 2015;75:4042–52.
65. El Tekle G, Bernasocchi T, Unni AM, Bertoni F, Rossi D, Rubin MA, et al. Co-occurrence and mutual exclusivity: what cross-cancer mutation patterns can tell us. Trends in Cancer. 2021;7:823–36.
66. Kwon CT, Tang L, Wang X, Gentile I, Hendelman A, Robitaille G, et al. Dynamic evolution of small signalling peptide compensation in plant stem cell control. Nat Plants. 2022;8:346–55.
67. Noguera MM, Rangani G, Heiser J, Bararpour T, Steckel LE, Betz M, et al. Functional PPO2 mutations: co-occurrence in one plant or the same ppo2 allele of herbicide-resistant Amaranthus palmeri in the US mid-south. Pest Manag Sci. 2021;77:1001–12.

Publisher's Note