Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2020), pp. 1–9 doi: 10.1111/pbi.13354

Trait associations in the pangenome of pigeon pea (*Cajanus cajan*)

Junliang Zhao[1,2], Philipp E. Bayer[3], Pradeep Ruperao[4,5], Rachit K. Saxena[5], Aamir W. Khan[3,5], Agnieszka A. Golicz[6], Henry T. Nguyen[7], Jacqueline Batley[3], David Edwards[3,*] and Rajeev K. Varshney[5,*]

[1]*Rice Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou, China*
[2]*Guangdong Key Laboratory of New Technology in Rice Breeding, Guangzhou, China*
[3]*School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth, WA, Australia*
[4]*National Institute of Agricultural Botany, Cambridge, UK*
[5]*International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India*
[6]*Plant Molecular Biology and Biotechnology Laboratory, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Melbourne, VIC, Australia*
[7]*Division of Plant Sciences, University of Missouri, Columbia, MO, USA*

Received 13 June 2019;
revised 11 December 2019;
accepted 14 January 2020.
*Correspondence (Tel +91 40 3071 3305; fax +91 40 30713074; email r.k.varshney@cgiar.org (RKV)) and (Tel +61 8 64882415; fax +61 8 6488 1380; email dave.edwards@uwa.edu.au (DE))
Keywords: pigeon pea, pangenome, GWAS, orphan crops, presence or absence variation.

Summary

Pigeon pea (*Cajanus cajan*) is an important orphan crop mainly grown by smallholder farmers in India and Africa. Here, we present the first pigeon pea pangenome based on 89 accessions mainly from India and the Philippines, showing that there is significant genetic diversity in Philippine individuals that is not present in Indian individuals. Annotation of variable genes suggests that they are associated with self-fertilization and response to disease. We identified 225 SNPs associated with nine agronomically important traits over three locations and two different time points, with SNPs associated with genes for transcription factors and kinases. These results will lead the way to an improved pigeon pea breeding programme.

Introduction

Pigeon pea, a member of tribe Phaseoleae, is a drought-tolerant crop grown in tropics and subtropical regions of the world. It is an important source of protein, vitamin B, carotene and ascorbic acid as well as an income generator, particularly in Tanzania, Malawi and Myanmar where it is produced as an export crop for India (Odeny, 2007; Saxena, 2008). Pigeon pea accessions possess significant variability for days to flowering (<50 to >160 days) and days to maturity (85–270 days; Vales *et al.*, 2012), as well as other important traits including disease resistance, abiotic stress tolerance, seed size, seed shape and number of seeds per pod (Saxena, 2008). The deep root system of pigeon pea fixes atmospheric nitrogen and improves the quality and structure of soil (Kumar Rao *et al.*, 2008). A draft genome sequence for pigeon pea was produced in 2012, and this is being used for the genetic improvement of this crop (Saxena *et al.*, 2017; Singh *et al.*, 2012; Varshney *et al.*, 2012). However, recent studies have shown that a single reference sequence cannot capture entire gene content of species because of significant gene presence or absence variation (PAV) (Golicz *et al.*, 2015a; Hurgobin *et al.*, 2018a). In order to understand pigeon pea gene diversity across the species and apply this for crop improvement, a pangenome is necessary.

The pangenome concept was first introduced to represent the full complement of genes within a bacterial species (Tettelin *et al.*, 2005) and comprises the core complement of genes common to all members of a species, and the variable or accessory genome represented by genes found in at least one but not all individuals. The variable genome contributes to species diversity and provides functions that are not essential, but which may provide a selective advantage under certain conditions, including resistance to biotic and abiotic stresses.

Emerging high-throughput DNA sequencing technologies led to the assembly of many crop genome reference assemblies and facilitated comparative genomic analysis between multiple individuals of same species revealing intraspecific diversity, leading to several published plant pangenomes. The first plant pangenome consisted of nine *Brassica oleracea* and one *Brassica macrocarpa* lines and identified 2154 additional genes (Golicz *et al.*, 2016), with PAV being predominantly associated with disease resistance genes (Bayer *et al.*, 2019). Shortly after, the first wheat pangenome was published consisting of 18 lines showing that up to 36% of wheat genes are variable (Montenegro *et al.*, 2017). Rapid advances in sequencing costs have led to larger studies, for example a pangenome of 725 tomato accessions revealed an additional 4873 genes not in the reference genome (Gao *et al.*, 2019). In legumes, pangenomes have been assembled using wild soya bean relatives (Li *et al.*, 2014) or by aligning ten legume genomes in order to find species-level differences (Wang *et al.*, 2017). To our knowledge, no pangenome has been assembled for orphan legumes.

In 2017, low coverage whole-genome sequencing was performed on 292 pigeon pea accessions, which encompass 95% of

Please cite this article as: Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., Khan, A. W., Golicz, A. A., Nguyen, H. T., Batley, J., Edwards, D.and Varshney, R. K. (2020) Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). *Plant Biotechnol. J.*, https://doi.org/10.1111/pbi.13354

© 2020 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

the total genetic diversity present in a larger composite collection of 1000 accessions, spanning the wide geographical distribution of pigeon pea (Varshney *et al.*, 2017). In the present study, sequencing data from the reference cultivar and 89 accessions with >9.5x coverage sequencing data were used to construct a pangenome for pigeon pea and for identification of the presence/absence of genes in these accessions. The pangenome constructed from these accessions comprises 55 512 genes, 13.41% of which demonstrate PAV in the accessions analysed, and 225 SNPs associated with nine phenotypes. Of these 225 SNPs, 21 replicated in different years and locations, indicating that there is a strong environmental component in pigeon pea's yield. Functional analysis suggests that variable genes are enriched for terms associated with self-fertilization and response to disease, and that associated SNPs are linked with transcription factors and kinases. The pigeon pea pangenome can serve as a valuable resource for harnessing the untapped genetic diversity within pigeon pea germplasm to enhance pigeon pea improvement.

Results

Pangenome assembly and annotation

Whole-genome resequencing data of 89 accessions with a minimum coverage of 9.5x were selected for building the pigeon pea pangenome from whole-genome resequencing data (Table S1). These 89 accessions include 70 from South Asia, 8 from sub-Saharan Africa, 7 from South-East Asia, 2 from Mesoamerica and 1 from Europe. Twenty-four accessions are breeding lines and 64 are landraces, while one individual is of unknown origin.

The pangenome was built using the iterative mapping and assembly approach (Golicz *et al.*, 2016) using the published genome assembly as the reference (Varshney *et al.*, 2012). The published reference genome assembly (C.cajan_V1.0) is 606 Mbp size with 48 680 predicted genes (Varshney *et al.*, 2012). However, reannotation in this study using 43 additional RNA-seq data as well as protein and EST sequences of pigeon pea as external evidence predicted a total of 53 612 genes in the draft reference. After pangenome construction and removal of contaminants, we assembled an additional 35 445 scaffolds, with a total length of 30 065 032 bp and containing 1900 additional genes, leading to a pangenome of 622 881 891 bp containing 55 512 genes.

Core and variable genes

The presence or absence of each gene was predicted for each accession based on the mapping of reads from each accession to the pangenome assembly using SGSGeneLoss (Golicz *et al.*, 2015b). The majority of genes were core (48 067, 86.6%), identified in all accessions, while 7445 (13.4%) of genes are variable, being absent from at least one individual. A total of 213 genes were only identified in a single pigeon pea accession, while the remaining variable genes were observed in more than one accession (Table S2). The size of the pangenome expanded with each additional line to 55 512 genes, and extrapolation leads to a predicted pangenome size of 57 768 $\pm$ 30 (Figure 1). The number of core genes decreased with each added accession to 48 067, with a predicted core gene content for the species of 47,742 $\pm$ 3 genes (Figure 1). Variable genes were shorter than core genes, with fewer exons per gene; however, the mean exon length is similar between core and variable genes (Table S3).

A dendrogram reconstructed using gene presence/absence variation suggests a close relationship between accessions (Figure 2). However, three landraces from South-East Asia (two from the Philippines and one from Indonesia) cluster away from the main group.

SNP discovery

Whole-genome sequence reads from the 89 pigeon pea accessions were mapped to the pangenome assembly for the identification of SNPs using GATK 3.8.0. A total of 8 475 005 SNPs were identified, of which, 181 393 were in the newly assembled pangenome scaffolds. The SNP frequency was higher in the new scaffolds (1/70) compared to the original reference (1/81). Only 455 014 SNPs (5.3%) had a low, moderate, or high impact on protein-coding sequences, of which 255 742 are predicted to lead to a mis-sense mutation. The frequency of mis-sense SNPs in the core gene set was (4.81 sites per kb) less than the variable gene set (6.34 sites per kb). After filtering for the GWAS based on SNP quality, minor allele frequency, and SNP depth, 3 713 723 SNPs remained.

Functional analysis of variable genes

Functional analysis of variable genes revealed enrichment of genes predicted to be involved in proteolysis, pollen tube reception, signal transduction, brassinosteroid mediated signalling pathway and defence response (Figure 3).

A total of 31 genes are classified with the GO term 'pollen tube reception', 19 of which are variable, with only two of these 19 present in the reference genome assembly. These 19 variable genes have high sequence identity with the FERONIA gene of Arabidopsis, which encodes a synergid-expressed, plasma membrane-localized receptor-like kinase, known to play an important role in reproductive isolation barriers (Escobar-Restrepo *et al.*, 2007).

Of the 278 genes which are annotated with the term 'plant-type hypersensitive response' (GO:0009626), 47 are variable genes, with 22 being assembled in the new pangenome contigs. Most of these genes contain the NBS-LRR motif, related to disease resistance. A dendrogram was generated using gene PAV of these 47 genes (Figure 4). The dendrogram groups differ from the dendrogram constructed using all genes (Figure 2), as ICP13579 clusters with the other Philippine individuals in a separate cluster in the hypersensitive response PAV dendrogram, while it is joined with the main cluster in the dendrogram with all genes.

The RGAugury pipeline was used for the identification and classification of R genes (Li *et al.*, 2016). In total, 909 R genes were identified and grouped into ten classes, namely NBS, CNL, TNL, TN, CN, NL, TX, RLK, RLP and 'Others' (Table 1). Among all classes, the majority of R genes were RLK (613 genes) followed by RLP (115) and CNL (61). Only five TNLs were identified, which is significantly less than the CNLs. Of the 909 R genes, 836 were core, and the remaining 73 were variable (Table 1).

SNP- and PAV-based genome-wide association study

The lines were assessed for 9 agronomically important traits ('Days to 50% flowering', 'Days to 75% maturity', 'Number of primary branches per plant', 'Number of secondary branches per plant', 'Plant height', 'Pods per plant', 'Seed weight', 'Seed yield per plant 'and 'Seeds per pod') for two years at three locations and correlation between these phenotypes was investigated (Table S4). Identical phenotypes for different years showed strong variability. Of the 2809 possible phenotype pairs, 21 pairs were

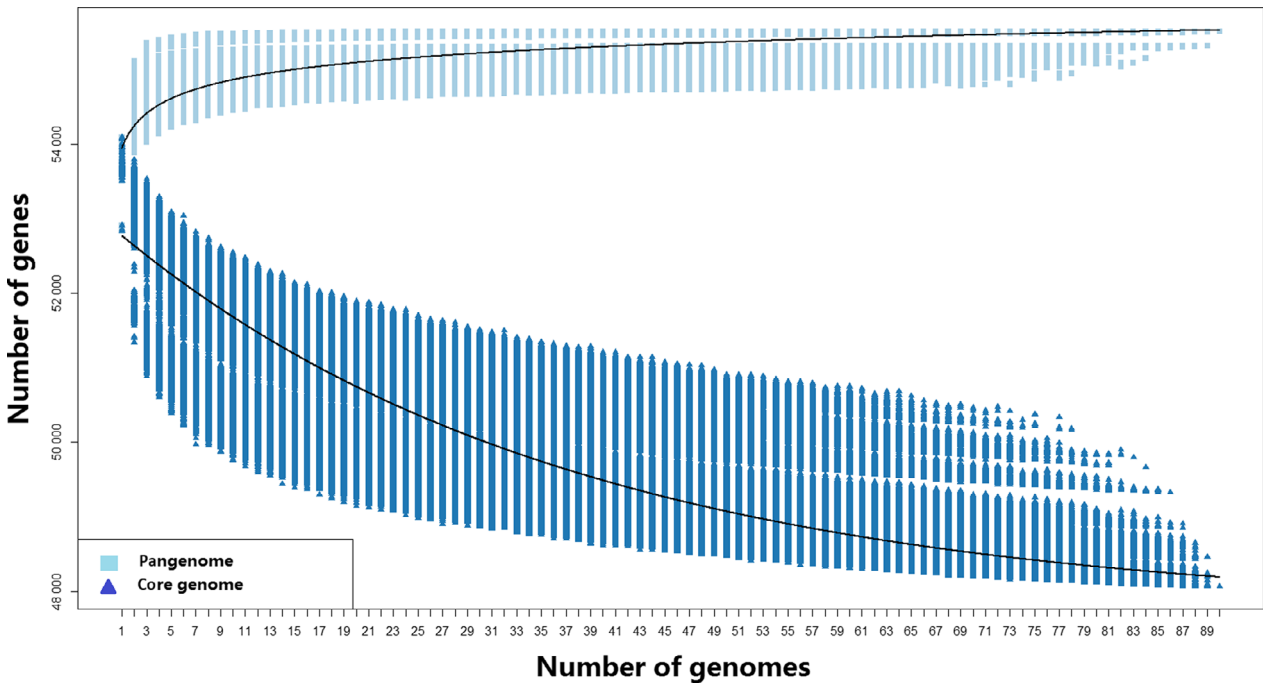

Figure 1 Growth curve for core and variable gene as modelled using nonlinear regression.

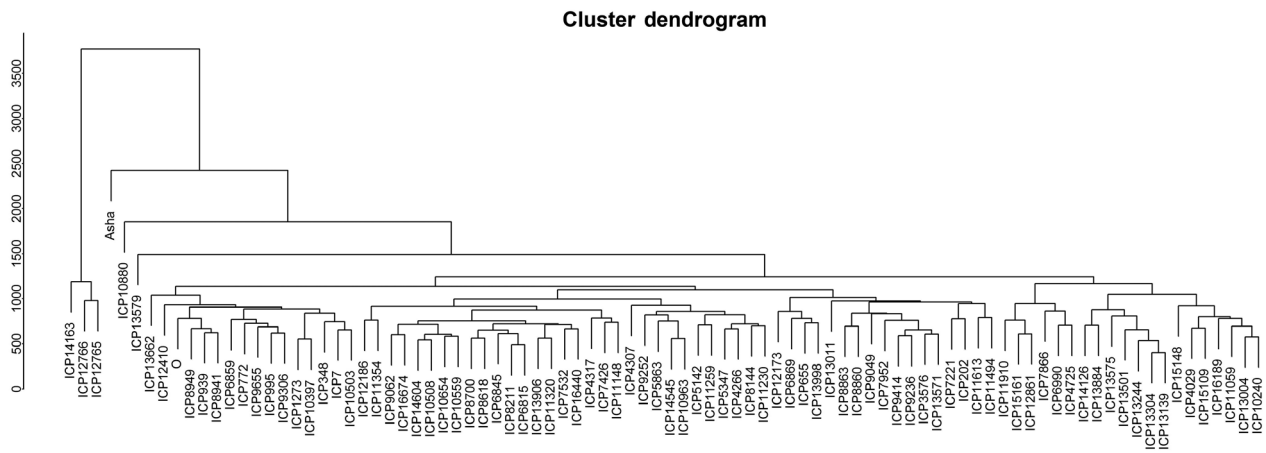

Figure 2 Dendrogram of all accessions based on presence–absence variation of all annotated genes in the pangenome.

highly correlated with an absolute value of r above 0.8 and with a p-value of approximately 0, while no phenotype pair was strongly negatively correlated.

Principal component analysis (PCA) revealed six outlier individuals from the Philippines and Indonesia (Figure S1), which were removed from further analysis. We found 225 SNPs associated with nine phenotypes over three locations and two different time points. Out of the 225 associated SNPs, 138 were located on unplaced contigs, and 15 were located on newly assembled pangenome contigs. Of the 225 associated SNPs, 21 SNPs (9.3%) were significantly correlated in two traits (Table S5).

Of the 225 candidate SNPs, 41 were located within gene sequences. For the remaining SNPs, we searched for the nearest upstream or downstream gene and identified candidate genes within the proximity of an associated SNP (Table S6). On average, only 14% (0%–30%) associated SNPs were shared between replicates of the same phenotype (Table S7). This may due to strong GxE interaction for pigeon pea which has been shown before (Upadhyaya *et al.*, 2012). In order to clarify this, we carried out one ANOVA per group of phenotypes and found significant ($P < 0.05$) GxE interactions for all phenotypes (Table S8).

Of the 41 SNPs within genes, 7 caused mis-sense changes, 2 cause synonymous changes, and 32 are intronic variants. Of the 43 genes with GO terms 16 had annotations containing the term GO:0005515 (protein binding), and 13 had annotations containing the terms GO:0006468 (protein phosphorylation) and GO:0004672 (protein kinase activity).

We performed GWAS with 4,635 variable genes as input and identified 3 variable genes associated with seed weight (Table 2), one of which (g01494) was on the additional pangenome contigs. The gene g01494 carries an endonuclease/exonuclease/

© 2020 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd.,

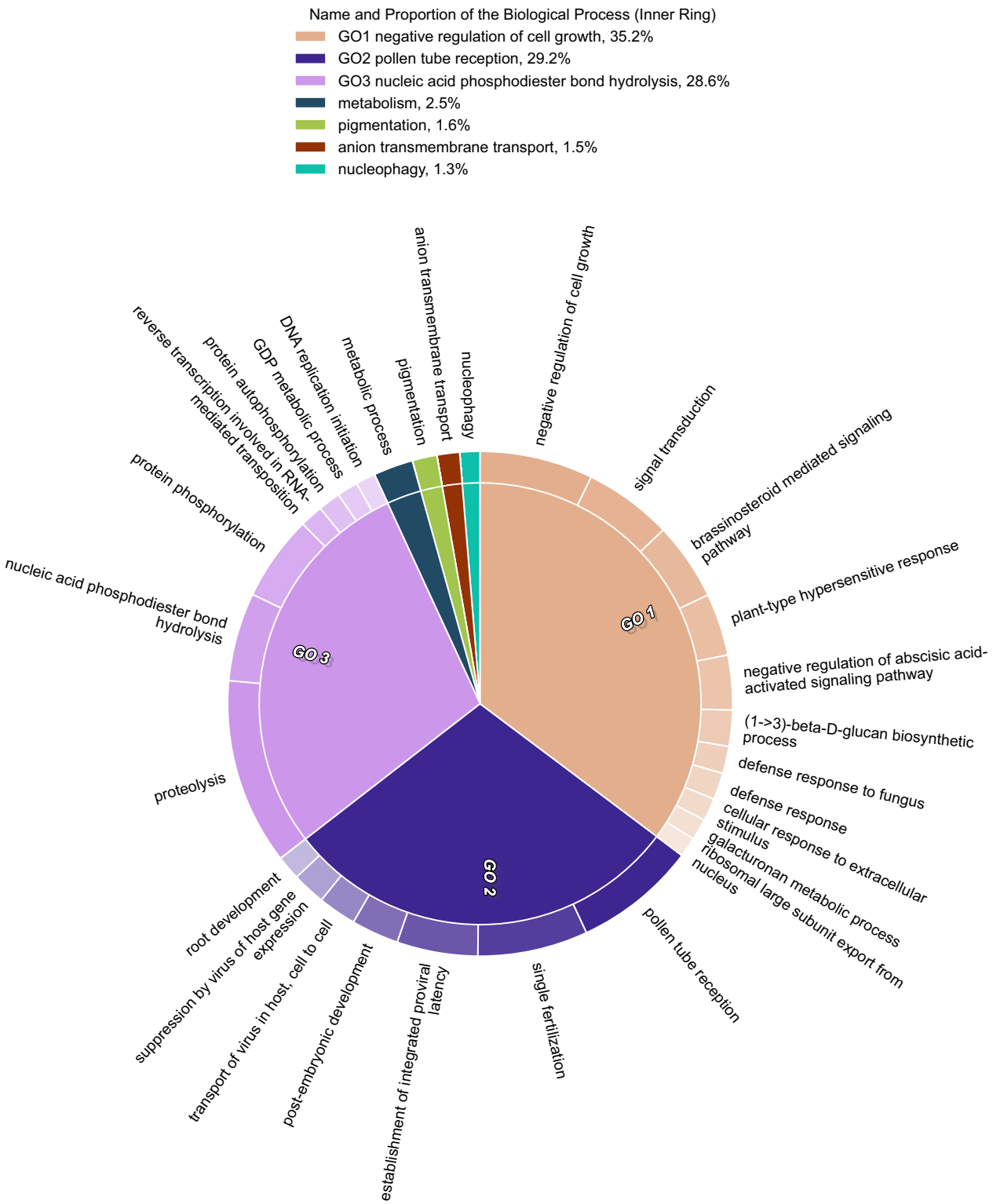

Figure 3 CirGO visualization of GO terms enriched in variable genes ($P < 0.05$). GO terms are mainly grouped into three groups: GO 1: negative regulation of cell growth, 35.2%. GO 2: pollen tube reception, 29.2%. GO 3: nucleic acid phosphodiester bond hydrolysis, 28.6%.

phosphatase domain (IPR036691) and is absent in ICP4725, ICP15161, ICP13004, ICP12861 and ICP10240, five individuals which have a seed weight at the upper end of the population's distribution (Table S9). None of the PAV-based candidate genes were co-located with the SNP-based candidate genes (Figure 5).

Discussion

The pangenome represents the entire gene set for a species and includes core genes, which are present in all individuals, and variable genes, which are absent in one or more individuals (Golicz *et al.*, 2016). High coverage DNA sequence data can be

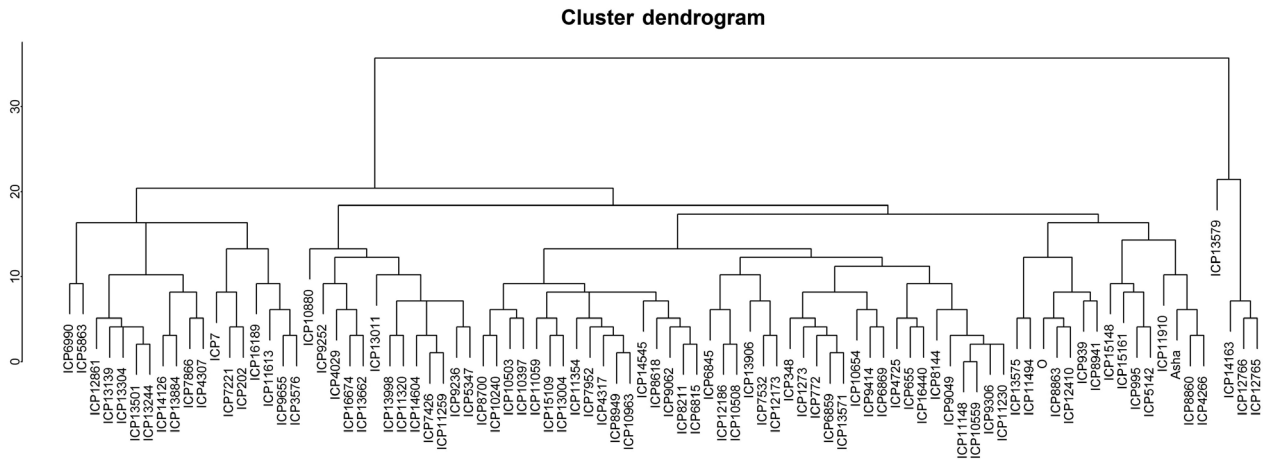

Figure 4 Dendrogram of all accessions based on presence–absence variation of genes belong to GO term 'plant-type hypersensitive response' (GO:000962) in the pangenome.

Table 1 Counts of NLR, RLK and RLP candidates in the pigeon pea reference genome and the pigeon pea pangenome extra contigs

Class	Pigeon pea reference	Pigeon pea pangenome extra contigs
CN	5	1
CNL	59	2
NBS	24	4
NL	47	7
OTHER	3	0
RLK (lrr)	233	1
RLK (lysm)	11	0
RLK (other receptor)	356	12
RLP (lrr)	107	7
RLP (lysm)	1	0
TN	1	0
TNL	5	0
TX	22	1

used for whole-genome assembly of multiple individuals followed by comparison to identify structural variations including gene presence/absence variation (Golicz *et al.*, 2016). However, this approach is expensive and maybe confounded by variations in assembly or annotation which are not representative of the genome (Bayer *et al.*, 2017). Alternatively, where relatively low sequence coverage is available, an iterative mapping and assembly approach can be used to construct a pangenome, followed by remapping of the sequence data to the pangenome to identify gene presence/absence variation. While this approach is not affected by false calls due to assembly or annotation variation, the majority of the newly assembled contigs are not physically placed within the pseudomolecules. We applied the iterative mapping and assembly approach to produce a pigeon pea pangenome assembly using 89 of the 292 sequenced accessions which have coverage of >9.5x. Accessions with lower coverage were not included in the pangenome analysis as they could not be reliably used to call presence/absence variation (Golicz *et al.*, 2016). However, this data set is representative of the diversity of the original data set of 292 sequenced accessions and contains the majority of continents and regions covered in the original data set.

The published reference assembly contains 48,680 predicted genes; however, reannotation predicted 53 612 genes, with the difference attributed to the increased amount of RNA-seq data used in the new annotation.

The assembled pangenome is 622 Mbp in length and contains 55 512 genes, 1900 more than the reference. The 30 Mbp of additional sequence represents a 5% increase compared to the reference assembly of 592 Mbp. The relative increase in pangenome assembly size is similar to that observed in rice (4%–6%) (Yao *et al.*, 2015), *Brachypodium distachyon* (5%) (Gordon *et al.*, 2017) and bread wheat (3.3%) (Montenegro *et al.*, 2017), but smaller than observed in *Brassica oleracea* (20%) (Golicz *et al.*, 2016). This small increase may be due to relatively low genetic diversity observed in pigeon pea (Varshney *et al.*, 2017).

Table 2 Variable genes linked with the phenotype '1314_100 Seed weight' in FarmCPU, their highest-scoring UniProt 100 BLAST hit and Pfam domains

Gene	Position	*P*-value	Effect in model	Highest-scoring UniProt 100 hit (identity %)	InterProScan domains
g49437	Chr 6 (14 552 342 bp)	1.5e-06	0.6	A0A151UDX0, Uncharacterized protein, *Cajanus cajan* (76.8%)	No IPR
g17019	Unplaced contigs	6e-08	−0.8	A0A151QP64, Uncharacterized protein, *Cajanus cajan* (97.2%)	No IPR
g01494	Unplaced contigs	2.5e-12	2	A0A151R3M6 Putative ribonuclease H protein At1g65750, *Cajanus cajan* (53.1%)	Endonuclease/exonuclease/ phosphatase superfamily (IPR036691)

© 2020 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–9

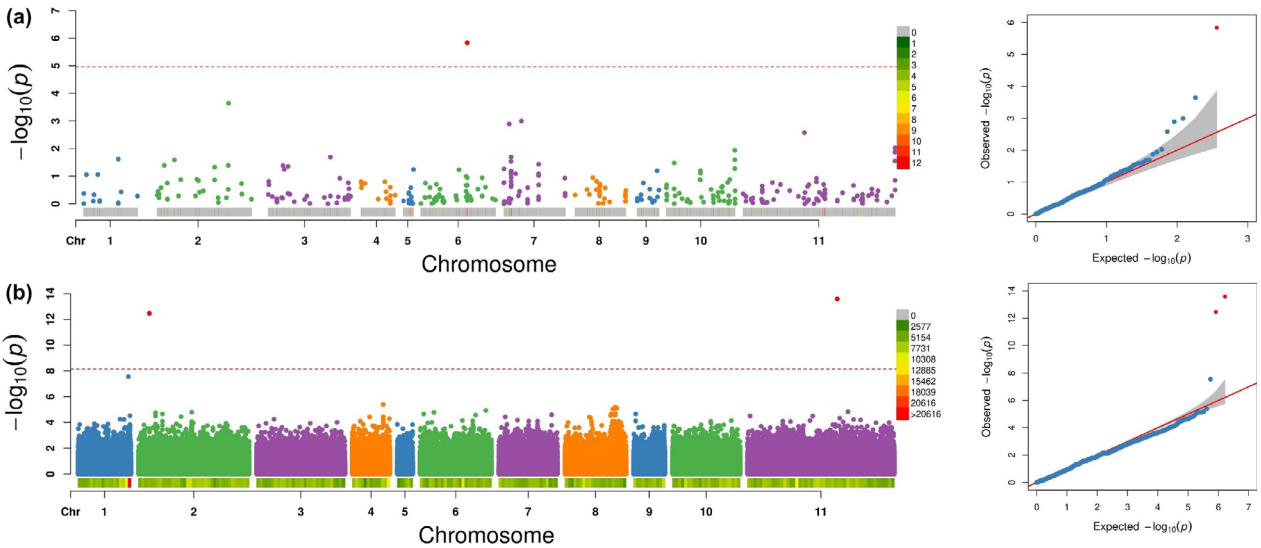

Figure 5 Manhattan and Q-Q plot for a) variable gene-based GWAS, phenotype 1314_100 Seed weight and b) SNP-based GWAS, phenotype Gul 1314_100 Seed weight. SNPs and genes on unplaced contigs and extra pangenome contigs not shown. Coloured bars below the chromosomes indicate gene – and SNP-density.

Functional analysis of variable genes demonstrated that they are enriched in GO terms associated with pollen tube reception and plant-type hypersensitive response. Self-fertilization is not obligatory in pigeon pea, and a considerable degree of natural out-crossing of around 20% has been documented (Saxena, 2008; Saxena *et al.*, 1990). We identified 17 novel candidate FERONIA gene orthologs in the pangenome in addition to the 14 FERONIA like genes annotated in the reference assembly. FERONIA has been shown to be central to many biological processes in plants, such as the control of seed size (Chalhoub *et al.*, 2014), the regulation of root hair development (Duan *et al.*, 2010) and plant immune signalling (Stegmann *et al.*, 2017). The different FERONIA orthologs here imply that the pigeon pea pangenome contains dispensable FERONIA alleles, which may be important for future breeding efforts.

Enrichment for plant-type hypersensitive response annotation in variable gene sets has been observed in other pangenome studies (Golicz *et al.*, 2016; Montenegro *et al.*, 2017). Only five TNLs were identified, which is significantly less than the CNLs; this observation is reported in other plants such as potato (Seo *et al.*, 2016). Of the 909 R genes, 836 were core, and the remaining 73 were variable (Table 1). A significant number of dispensable resistance genes were also observed in the *Brassica oleracea* (Golicz *et al.*, 2016) and *B. napus* pan genomes (Hurgobin *et al.*, 2018b).

Phylogenetic analysis of all 292 accessions using SNPs on the reference genome assembly demonstrated that five out of seven accessions (ICP 11230, ICP 11238, ICP 11015, ICP 11096 and ICP 11148), which are resistant to sterility mosaic disease cluster together, while the other two accessions (ICP 11059, ICP 7436) from the Philippines were distantly placed in another subgroup (Varshney *et al.*, 2017). Three of the seven accessions (ICP 11230, ICP 11148 and ICP 11059) are included in our analysis and these cluster based on the presence of variable plant-type hypersensitive response genes (Figure 4). As ICP 11059 does not join this cluster in the SNP analysis or the total PAV clustering, it suggests that the annotated gene-based PAV clustering may help to identify candidate genes associated with this important trait.

The pigeon pea individuals were assessed for 9 phenotypes, and correlations between these phenotypes were assessed to see whether pigeon pea breeding has led to phenotype linkages. Interestingly, some phenotypes one may expect to correlate do not correlate significantly. For example, the two phenotypes, number of seeds per pod and seed weight, may be expected to have an inverse correlation, as the plant has to choose whether to spend resources on making more seeds, or on making bigger seeds. These two phenotypes have a slightly positive correlation (0.5). However, the number of pods negatively correlated with seed weight with $r = -0.61$ and $r = -0.67$. This suggests that when bred for seed yield, the plants produce fewer pods with larger seeds while the number of seeds within the pods remains stable or increases.

Gene presence/absence variation (PAV) can lead to phenotypic diversity. We measured the association between all SNPs and variable genes and measured phenotypes. A previous GWAS in 292 pigeon pea individuals reported 241 associated regions (Varshney *et al.*, 2017).

The GWAS study performed in the present study is an advancement of the previous study reported in WGRS of 292 accessions (Varshney *et al.*, 2017) with additional trait phenotyping data from two locations, and additional pangenome contigs. A total of 90 SNPs for the cropping season 2013–14, 101 SNPs for the cropping season 2014–2015 and the remaining 34 SNPs for pooled data were found to be associated with target traits. We observed that many identified associated SNPs in one year or one location showed relatively weak or no associations in the other year or location, as observed in the previous pigeon pea GWAS.

Though a number of SNPs were detected for 'Seed weight', 'Days to 50% flowering', 'Days to 75% maturity' and 'Plant height' earlier using the earlier draft genome (Varshney *et al.*, 2017), the present study has identified additional SNPs for these traits. Varshney *et al.* (2017) found associations for 1 SNP for the phenotype 'Pods per plant', 1 SNP for 'Number of primary branches per plant' and 3 SNPs for 'Number of secondary branches per plant'.

This study found additional SNPs for the same phenotypes: 14 SNPs for 5 replicates of 'Pods per plant', 16 SNPs for 4 replicates of 'Number of primary branches per plant' and 5 SNPs for 1 replicate of 'Number of secondary branches per plant' (Table S6).

SNPs replicated rarely over different locations and phenotypes. As observed above, for the phenotype 'Number of primary branches' the majority of associated SNPs did not overlap between the three replicates. For example, for the pooled replicate there were three associated SNPs on pseudomolecule 1 at next to each other at position 11 Mbp, while for the replicate 'Gul_1314' and 'Gul_1415' no SNPs were associated on pseudomolecule 1, instead on the unplaced contigs, pseudomolecule 11 and pseudomolecule 2. This indicates that in pigeon pea, location and year have a strong impact on yield-related phenotypes. Pigeon pea is known to exhibit large phenotype variation due to strong GxE (Upadhyaya *et al.*, 2012), which we also observed in all sets of phenotype data of all locations used in this study (Table S8). Strong GxE interactions may explain the different results in different locations and years here.

Most of the associated SNPs linked to genes were linked to genes containing kinase or protein binding domains, indicating that these genes are transcription factors. Interestingly, none of the *Glycine max* orthologous reported here overlap with *Glycine max* GWAS candidate genes linked to seed weight reported previously (Jing *et al.*, 2018; Yu *et al.*, 2018), which may be due to selection for seed weight in *Cajanus cajan* acting on different pathways than in *Glycine max*.

We found an association for three genes with the seed weight phenotype. Proteins encoded by two of these three genes carry no predicted protein domains and have only hits with uncharacterized *C. cajan* proteins, so they require further analysis to determine their potential function. The third protein g01494 has an endonuclease/exonuclease/phosphatase superfamily domain, a domain common in proteins with diverse functions, though it also has similarity with a transposon TX1, suggesting that absence of this transposon may be linked to higher seed weight. There are many known examples of transposon insertion altering phenotype, including seed weight. In soya bean, the introduction of a 5.7kb transposon is associated with pink flowers and a 22% increase in seed weight (Zabala and Vodkin, 2005). Here, this gene was absent from five individuals on the high end of the seed weight distribution, suggesting that it could be negatively involved in seed weight regulation. This gene is variable and not found in the reference assembly demonstrating the value of using a pangenome reference for association studies.

Methods

Phenotyping

Pigeon pea accessions were phenotyped at three locations, in years 2013–14 and 2014–15 at Gulbarga and ICRISAT; 2014–15 at Tandur. The phenotyping data from ICRISAT used in the present study have been taken from (Varshney *et al.*, 2017). In order to generate reliable phenotyping data, all the accessions were planted in two replications in an alpha-lattice design. In each replication, data were recorded from three plants from each accession. All the accessions were phenotyped for nine yield and yield-related traits, including days to 50% flowering, days to 75% maturity, plant height, number of primary branches, number of secondary branches, pods per plant, number of seeds per pod, 100 seed wt and seed yield per plant. The phenotyping was conducted following standard procedures described in the GenBank manual (Upadhyaya and Gowda, 2009).

Pangenome assembly and annotation

Whole-genome sequence data of 89 pigeon pea accessions having more than 9.5x coverage from Varshney *et al.* (2017) were used for pangenome assembly using the iterative mapping and assembly approach. The majority of the diversity present in the original dataset is present in the pangenome dataset. The pangenome data contain 64 landraces, 24 breeding lines and 1 unknown, compared with 167 landraces, 117 breeding lines, 14 wild types and 2 unknowns in Varshney *et al.* (2017). All regions present in the original data are present in the pangenome, with a focus on India. Of the pangenome individuals, 67 come from India, 4 from the Philippines and 17 additional countries including Uganda (1 individual), Kenya (2 individuals) and Zaire (1 individual).

The pangenome was constructed by mapping of the sequence reads individually to the reference genome (Varshney *et al.*, 2012), using bowtie2 v2.3.0 (Langmead and Salzberg, 2012), followed by assembly of pooled unmapped using MaSuRCA v3.2.3 (Zimin *et al.*, 2013) to produce additional reference sequence. The assembled contig sequences were compared with the NCBI nt database (downloaded 2 May 2018) using BLAST v2.5.0. Contigs with best hits to non-green plants, chloroplast or mitochondrial sequences were removed. The remaining newly assembled contigs >1 Kb in length were annotated using MAKER2 (Holt and Yandell, 2011). De novo gene prediction was performed with SNAP (Schmid *et al.*, 2003) and Augustus (Stanke *et al.*, 2006). Publicly available ESTs (24 177), as well as 43 pigeon pea RNA-seq data sets (Table S10) and proteins (48 450) from NCBI, were used as evidence. The functional annotations were assigned by BLAST comparison with UniProt 90 of *A. thaliana*. Gene ontologies (GO terms) were assigned based according to GO terms of the best hit of each gene by home-made python scripts. GO enrichment was performed using Fisher's exact test as implemented in topGO package (Alexa *et al.*, 2006) with method 'elim' used to adjust for multiple comparisons. REVIGO (Supek *et al.*, 2011) was used to remove redundant GO categories from all GO terms enriched with a *P*-value below 0.05, and CirGO (Kuznetsova *et al.*, 2019) was used to visualize the results.

Gene presence/absence variation and pangenome modelling

Whole-genome sequence data for all 89 pigeon pea accessions were mapped to the reference genome using bowtie2 v2.2.5 (--end-to-end --sensitive -I 0 -X 1000; Langmead and Salzberg, 2012). SGSGeneLoss (Golicz *et al.*, 2015b) was used to determine whether a gene is present or absent. Curves describing pangenome size and core genome size were fitted in R (R Core Team, 2018) using the nls function (nonlinear least squares) from package stats, part of R. Points used in regression corresponded to all the possible combinations of genomes, similar to Hirsch *et al.* (2014).

SNP discovery and annotation

Whole-genome sequence reads were mapped to the pangenome using Bowtie2 v2.2.9 (-I 0 -X 1000) (Langmead and Salzberg, 2012). Parallel jobs were run using GNU parallel 20160622 (Tange, 2011). The resulting SAM files were converted to BAM format using samtools (Li *et al.*, 2009), followed by the removal of

© 2020 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 1–9

duplicate reads using picard tools v2.14 (http://broadinstitute.github.io/picard/). SNPs were called using UnifiedGenotyper in GATK 3.8.0 (McKenna *et al.*, 2010) and functionally annotated using SnpEff v4.3T (Cingolani *et al.*, 2012).

After removing SNPs with a minor allele frequency below 1%, a SNP quality score (QUAL) below 20 and a SNP depth (DP) below 10, we used BLINK v0.01 (standard settings; Huang *et al.*, 2019) for association analysis and rMVP to plot Manhattan and QQ-plots (https://github.com/XiaoleiLiuBio/rMVP). The *P*-value significance cut-off was set to 1.35e-08 (=0.05/3 713 723 SNPs).

The GWAS with the 4635 variable genes instead of SNPs using the SNP-based principal components as covariates was performed using FarmCPU as implemented in rMVP (standard settings) with a significance cut-off set to 1.08e-5 (0.05/4635). The presence/absence matrix was encoded as SNPs, where 'presence' was the minor and 'absence' was the major allele.

ANOVA was carried out using the R v3.5.1 function *aov* by grouping all locations and time points for each of the nine phenotypes using the formula *phenotype* ~ *location/year*. *P*-values were adjusted for multiple comparisons by using R's *p.adjust* method.

Genes upstream or downstream from candidate SNPs were mined using bedtools2 v2.27.1 *closest* (Quinlan and Hall, 2010). All candidate genes were aligned with the *Glycine max* Williams-82 Wm82.a2.v1 reference annotation (Schmutz *et al.*, 2010) using blastp (options: -evalue 1e-10; Camacho *et al.*, 2009) and annotations for these genes were extracted from SoyBase (Grant *et al.*, 2010).

Acknowledgments

This research was supported by the Australian Government through the Australian Research Council's *Linkage Projects* funding scheme (project LP140100537, LP160100030). PB acknowledges support of the Forrest Research Foundation.

Author contributions

JZ and PB carried out the bioinformatics analysis. RKS, AWK and NH grew and sequenced the pigeon pea individuals. JZ, PB, JB, DE and RKV jointly wrote the manuscript.

Competing interests

The authors declare no competing interests.

Data availability statement

The pigeon pea pangenome assembly, annotation and SNP data are available at https://doi.org/10.26182/5ca6cf7a482fe.

References

Alexa, A., Rahnenfuhrer, J. and Lengauer, T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Bayer, P.E., Hurgobin, B., Golicz, A.A., Chan, C.-K.K., Yuan, Y., Lee, H., Renton, M. *et al.* (2017) Assembly and comparison of two closely related Brassica napus genomes. *Plant Biotechnol. J.* **15**, 1602–1610.

Bayer, P.E., Golicz, A.A., Tirnaz, S., Chan, C.K., Edwards, D. and Batley, J. (2019) Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. *Plant Biotechnol. J.* **17**, 789–800.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I., Tang, H., Wang, X., Chiquet, J. *et al.* (2014) Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, **345**, 950–953.

Cingolani, P., Platts, A., le Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Duan, Q., Kita, D., Li, C., Cheung, A.Y. and Wu, H.-M. (2010) FERONIA receptor-like kinase regulates RHO GTPase signaling of root hair development. *Proc. Natl. Acad. Sci. USA*, **107**, 17821–17826.

Escobar-Restrepo, J.M., Huck, N., Kessler, S., Gagliardini, V., Gheyselinck, J., Yang, W.C. and Grossniklaus, U. (2007) The FERONIA receptor-like kinase mediates male-female interactions during pollen tube reception. *Science*, **317**, 656–660.

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

Golicz, A.A., Batley, J. and Edwards, D. (2015a) Towards plant pangenomics. *Plant Biotechnol. J.* **14**, 1099–1105.

Golicz, A.A., Martinez, P.A., Zander, M., Patel, D.A., Van De Wouw, A.P., Visendi, P., Fitzgerald, T.L. *et al.* (2015b) Gene loss in the fungal canola pathogen Leptosphaeria maculans. *Funct. Integrative Genom.* **15**, 189–196.

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A., Chan, C.K.K. *et al.* (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* **7**, 13390.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C. *et al.* (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184.

Grant, D., Nelson, R.T., Cannon, S.B. and Shoemaker, R.C. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* **38**, D843–846.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G. and Vaillancourt, B. (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, **26**, 121–135.

Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **12**, 491.

Huang, M., Liu, X., Zhou, Y., Summers, R.M. and Zhang, Z. (2019) BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience*, **8**, giy154.

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

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

Jing, Y., Zhao, X., Wang, J., Teng, W., Qiu, L., Han, Y. and Li, W. (2018) Identification of the genomic region underlying seed weight per plant in soybean (*Glycine max* L. Merr.) via high-throughput single-nucleotide polymorphisms and a genome-wide association study. *Front. Plant. Sci.* **9**, 1392.

Kumar Rao, J.V.D.K., Dart, P.J. and Sastry, P.V.S.S. (2008) Residual effect of pigeonpea (*Cajanus cajan*) on yield and nitrogen response of maize. *Exp. Agric.* **19**, 131–141.

Kuznetsova, I., Lugmayr, A., Siira, S.J., Rackham, O. and Filipovska, A. (2019) CirGO: an alternative circular way of visualising gene ontology terms. *BMC Bioinform.* **20**, 84.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–U354.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, Y.H., Zhou, G., Ma, J., Jiang, W., Jin, L.G., Zhang, Z., Guo, Y. *et al*. (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052.

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S. and You, F.M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genom.* **17**, 852.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chan, C.-K.K., Visendi, P. *et al*. (2017) The pangenome of hexaploid bread wheat. *Plant J.* **90**, 1007–1013.

Odeny, D.A. (2007) The potential of pigeonpea (*Cajanus cajan* (L.) Millsp.) in Africa. *Natural Resources Forum*, **31**, 297–305.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

R Core Team. (2018) *R: A Language and Environment for Statistical Computing*.

Saxena, K. (2008) Genetic improvement of pigeonpea–a review. *Trop. Plant. Biol.* **1**, 159–178.

Saxena, K.B., Singh, L. and Gupta, M.D. (1990) Variation for natural outcrossing in pigeonpea. *Euphytica*, **46**, 143–148.

Saxena, R.K., Kale, S.M., Kumar, V., Parupali, S., Joshi, S., Singh, V., Garg, V. *et al*. (2017) Genotyping-by-sequencing of three mapping populations for identification of candidate genomic regions for resistance to sterility mosaic disease in pigeonpea. *Sci. Rep.* **7**, 1813.

Schmid, K.J., Sorensen, T.R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. *Genome Res.* **13**, 1250–1257.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L. *et al*. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Seo, E., Kim, S., Yeom, S.-I. and Choi, D. (2016) Genome-wide comparative analyses reveal the dynamic evolution of nucleotide-binding leucine-rich repeat gene family among solanaceae plants. *Front. Plant Sci.* **7**, 1205.

Singh, N.K., Gupta, D.K., Jayaswal, P.K., Mahato, A.K., Dutta, S., Singh, S., Bhutani, S. *et al*. (2012) The first draft of the pigeonpea genome sequence. *J. Plant Biochem. Biotechnol.* **21**, 98–112.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–439.

Stegmann, M., Monaghan, J., Smakowska-Luzan, E., Rovenich, H., Lehner, A., Holton, N., Belkhadir, Y. *et al*. (2017) The receptor kinase FER is a RALF-regulated scaffold controlling plant immune signaling. *Science*, **355**, 287–289.

Supek, F., Bosnjak, M., Skunca, N. and Smuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, **6**, e21800.

Tange, O. (2011) Gnu parallel-the command-line power tool. *USENIX Magazine*, **36**, 42–47.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V. *et al*. (2005), Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA*, **102**, 13950–13955.

Upadhyaya, H.D. and Gowda, C.L. (2009). Characterization and Preliminary Evaluation. In: *Managing and Enhancing the Use of Germplasm-Strategies and Methodologies. Technical Manual no. 10* (pp 171-202). Int. Crops Res. Institute Semi-Arid Trop.

Upadhyaya, H.D., Kashiwagi, J., Varshney, R.K., Gaur, P.M., Saxena, K.B., Krishnamurthy, L., Gowda, C.L. *et al*. (2012) Phenotyping chickpeas and pigeonpeas for adaptation to drought. *Front. Physiol.* **3**, 179.

Varshney, R., Chen, W., Li, Y., Bharti, A., Saxena, R., Schlueter, J., Donoghue, M. *et al*.(2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89.

Vales, M.I., Srivastavab, R.K., Sultanab, R., Singhc, S., Singhc, I., Singhd, G., Patilb, S.B., *et al.* (2012) Breeding for Earliness in Pigeonpea: Development of New Determinate and Nondeterminate Lines. *Crop Sci.* **52**, 2507–2516.

Varshney, R., Saxena, R., Upadhyaya, H., Khan, A., Yu, Y., Kim, C., Rathore, A. *et al*. (2017) Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nat. Genet.* **49**, 1082–1088.

Wang, J., Sun, P., Li, Y., Liu, Y., Yu, J., Ma, X., Sun, S. *et al*. (2017) Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. *Plant Physiol.* **174**, 284–300.

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X. and Xie, W. (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* **16**, 1–20.

Yu, M., Liu, Z., Jiang, S., Xu, N., Chen, Q., Qi, Z. and Lv, W. (2018) QTL mapping and candidate gene mining for soybean seed weight per plant. *Biotechnol. Biotech. Equipment*, **32**, 1–7.

Zabala, G. and Vodkin, L.O. (2005) The mutation of glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. *Plant Cell*, **17**, 2619–2632.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L. and Yorke, J.A. (2013) The MaSuRCA genome assembler. *Bioinformatics*, **29**, 2669–2677.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Dotplot of PC1 vs PC2 of the pigeon pea population before removal of outliers.
Table S1 Individuals and sequencing coverage.
Table S2 Variable genes and number of accessions containing the genes out of 89 accessions and 1 reference cultivar.
Table S3. Gene and exon length compared between core and variable genes.
Table S4 Correlations between phenotypes studied here, correlation coefficient $r > 0.8$, $P < 0.05$.
Table S5 SNPs appearing in GWAS results more than once.
Table S6 Upstream and downstream genes of candidate SNPs.
Table S7 Shared SNPs between replicates of the same phenotype.
Table S8 Results of ANOVA analysis of GxE of 9 phenotypes in this study.
Table S9 Seed range measurements for all individuals.
Table S10 List of RNA-seq data used as evidence to annotated the pigeon pea pangenome.