The Plant Genome

ORIGINAL ARTICLE

Special Section: Grain Quality and Nutritional Genomics for Breeding Next Generation Crops

Whole-genome sequencing based discovery of candidate genes and diagnostic markers for seed weight in groundnut

Sunil S. Gangurde[1,2,#] | **Aamir W. Khan**[1,#] | **Pasupuleti Janila**[1] | **Murali T. Variath**[1] | **Surendra S. Manohar**[1] | **Prashant Singam**[2] | **Annapurna Chitikineni**[1] | **Rajeev K. Varshney**[1,3] | **Manish K. Pandey**[1]

[1]Center of Excellence in Genomics & Systems Biology (CEGSB), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad502324, India

[2]Dep. of Genetics, Osmania Univ., Hyderabad 500007, India

[3]State Agricultural Biotechnology Centre, Crop Research Innovation Centre, Food Futures Institute, Murdoch Univ., Murdoch, Western Australia 6150, Australia

Correspondence:
Manish K. Pandey and Rajeev K. Varshney, Center of Excellence in Genomics & Systems Biology (CEGSB), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502324, India.
Email: m.pandey@cgiar.org; rajeev.varshney@murdoch.edu.au

#Sunil S. Gangurde and Aamir W. Khan contributed equally.

Abstract

Seed weight in groundnut (*Arachis hypogaea* L.) has direct impact on yield as well as market price because of preference for bold seeds by consumers and industry, thereby making seed-size improvement as one of the most important objectives of groundnut breeding programs globally. Marker-based early generation selection can accelerate the process of breeding for developing large-seeded varieties. In this context, we deployed the quantitative trait locus-sequencing (QTL-seq) approach on a biparental mapping population (Chico × ICGV 02251) to identify candidate genes and develop markers for seed weight in groundnut. A total of 289.4–389.4 million reads sequencing data were generated from three libraries (ICGV 02251 and two extreme bulks) achieving 93.9–95.1% genome coverage and 8.34–9.29× average read depth. The analysis of sequencing data using QTL-seq pipeline identified five genomic regions (three on chromosome B06 and one each on chromosomes B08 and B09) for seed weight. Detailed analysis of above associated genomic regions detected 182 single-nucleotide polymorphisms (SNPs) in genic and intergenic regions, and 11 of these SNPs were nonsynonymous in the genomic regions of 10 candidate genes including *Ulp proteases* and *BIG SEED locus* genes. Kompetitive allele specific polymerase chain reaction (KASP) markers for 14 SNPs were developed, and four of these markers (snpAH0031, snpAH0033, snpAH0037, and snpAH0038) were successfully validated for deployment in breeding for large-seeded groundnut varieties.

1 | INTRODUCTION

Groundnut, or peanut (*Arachis hypogaea* L.), is an important oilseed and food crop cultivated in semi-arid regions of the world. This crop is cultivated on over 31.6 million ha with an annual production of 53.6 Tg (http://www.fao.org/faostat/en). Groundnut is a nutritious legume crop that is known for its healthy oil, protein, and fiber content. Although the combination of big seed size and high oil content ensures higher pod yield and oil yield, the combination of big seed size and low oil content ensures industry and consumer satisfaction. Bold-seeded groundnut is generally preferred in confections, being an important physical quality that attracts the immediate attention of consumers (Venuprasad et al., 2011). Groundnut with both big seed size and high oil content have great potential to benefit groundnut farmers, industry, and, most importantly, consumers.

The size of the pod and seed in groundnut are positively correlated, whereas the relationship between pod size and shelling percentage is not always positive, which limits success in developing varieties bearing large pods with high shelling percentage (de Godoy & Norden, 1981). Seed size has been one of the important domesticated traits in breeding groundnut varieties with different growth habits (Runner and Spanish). In general, Runner type varieties, having large seeds, are suitable for food purposes, whereas Spanish type, with small seed size, are mostly grown for oil extraction. Because of multiple preparations being available in the market for food usage, there is an increased demand of groundnut with more market-preferred features such as big seed size and high oleic acid. This trend is growing more in Asia and Africa where the major proportion of produce is used for oil extraction. It is very crucial to study the genetic basis of regulation of seed weight in groundnut to discover candidate genes and to develop diagnostic markers that can be used in marker-based early generation selection (MEGS) for improving the seed weight of elite groundnut cultivars.

Genomics-assisted breeding has been playing an important role in groundnut, and there are many examples of successful efforts in developing improved varieties with disease resistant and high oleic acid content using diagnostic markers (see Pandey et al., 2020; Varshney et al., 2021). The availability of linked markers for seed weight will provide opportunity to perform MEGS in breeding bold-seeded groundnut varieties. It is worth mentioning here that phenotypic selection for seed size is very difficult in groundnut, as the pods are inside the soil and the breeder needs to dig out pods in order to have an idea of pod and seed size. Under such scenario, the MEGS for seed size will provide great support to make decisions on other important traits in further generations. The previous genetic mapping efforts in groundnut for seed weight identified three significant quantitative trait loci (QTLs) located in a 2.7-Mb region at the end of chromosome A05 (Luo et al.,

---

Core ideas

- QTL-seq analysis identified three genomic regions associated with groundnut seed weight were on chromosomes B06, B08, and B09.
- Eleven non-synonymous SNPs identified in the regions of 10 important candidate genes associated with seed weight in groundnut.
- The *BIG SEEDs* locus associated with groundnut seed weight was identified in genomic region on chromosome B09.
- Four diagnostic KASP markers were developed and validated for deploying genomic-assisted breeding to improve seed weight in groundnut.

2017) as well as pod weight on chromosome A05 (Luo et al., 2018) and a major QTL on chromosome A05 for seed number per pod (Chen et al., 2019a). The meta-QTL analysis using consensus map narrowed down the genomic region for seed weight from 2.7 to 0.7 cM on chromosome A05(Lu et al., 2018). Further, a recent nested-association mapping approach in groundnut identified associated genomic loci on chromosomes A05, A06, A08, A09, B05, B06, B08, and B09 for seed weight (Gangurde et al., 2020). None of these studies, however, reported development of diagnostic markers for seed size to accelerate breeding for large-seeded groundnut varieties.

Availability of reference genomes coupled with the next-generation sequencing technologies have accelerated candidate gene discovery and marker development followed by their deployment in breeding including in groundnut (see Pandey et al., 2016; Varshney et al., 2019; Bertioli et al., 2019; Zhuang et al., 2019). Among these approaches, a next-generation sequencing-based bulk segregant analysis approach, also known as QTL sequencing (QTL-seq), has been widely used for discovery of genomic regions and candidate genes as well as diagnostic markers. This approach provided successful results for resistance to late leaf spot and rust (Pandey et al., 2017; Clevenger et al., 2018) and bacterial wilt (Luo et al., 2019a), shelling percentage (Luo et al., 2019b), testa color (Zhao et al., 2020), and fresh-seed dormancy (Kumar et al., 2020) in groundnut. The QTL-seq approach uses the advantage of sequencing and generates thousands of genome-wide single-nucleotide polymorphism (SNP) variants for bulks by using reference-guided assembly of either parent of a recombinant inbred line (RIL) population (Takagi et al., 2013). In the present study, we used QTL-seq approach to study the molecular basis of seed weight in groundnut by using a RIL population and identified associated genomic regions and candidate genes for seed weight in

addition to development and validation of diagnostic markers for MEGS.

2 | MATERIALS AND METHODS

2.1 | Plant materials and construction of bulks with extreme phenotypes

A RIL population (Chico × ICGV 02251) comprising of 385 lines (RILs) was developed by crossing groundnut cultivars Chico, a small-seeded cultivar, and ICGV 02251 a large-seeded genotype (Figure 1a,b). Extensive phenotyping data for 100-seed weight (100SW) was generated on a RIL population in the F_9 generation and was phenotyped at ICRISAT, Hyderabad, India, during three seasons (rainy 2014, postrainy 2014–2015, and rainy 2019). The extreme pools were prepared for 100SW based on phenotyping data obtained for the year postrainy 2014–2015. For developing the extreme bulks for trait 100SW, 25 RILs with high mean seed weight and 25 RILs with low mean seed weight were selected. The equimolar concentration of DNA from 25 RILs with high mean seed weight were pooled together to construct the high-seed-weight (HS) bulk; similarly, DNA from RILs with low mean seed weight was pooled together to construct the low-seed-weight (LS) bulk. Thus, two extreme bulks were prepared for sequencing. High-seed-weight parent ICGV 02251 was used as a reference parent to calculate SNP indices for LS and HS bulks (Figure 1d).

2.2 | Construction of libraries and sequencing for bulks and parent

The whole-genome resequencing data were generated for three samples, namely ICGV 02251 (bold-seeded parent), HS bulk, and LS bulk, and were prepared and used for sequencing on Illumina HiSeq 2500 at the Center of Excellence in Genomics and Systems Biology (CEGSB; cegsb.icrisat.org), ICRISAT, Hyderabad, as described in Pandey et al. (2017). Sequencing data generated have been deposited at National Center for Biotechnology Information (NCBI) Sequence Read Archive database with BioProject ID PRJNA752462. In brief, a single Illumina library for each sample was made using TruSeq DNA Sample Prep kit. Two micrograms of DNA from each of these three samples was first sheared using diagenode Bioruptor NGS and then was subjected to end repairing and adapter ligation. Two percent agarose gel was used for selection of libraries on the basis of size. Further, the libraries were enriched using adaptor-compatible polymerase chain reaction (PCR) primers. A chip assay was performed for size distribution of amplified DNA libraries and was checked on an Agilent Technologies 2100 Bioanalyzer. The DNA libraries were sequenced on Illumina HiSeq 2500 (Illumina Inc.) to generate 250 bp paired-end reads.

2.3 | Development of reference guided assembly and QTL-seq analysis

The quality assessment and data estimation before and after quality control was performed using Raspberry of NGS-QC box (Katta et al., 2015). The cleaned reads of parent ICGV 02251 were first aligned to the combined assembly of groundnut progenitors (A- and B-diploid subgenomes) (Bertioli et al., 2016) using BWA aligner (Li and Durbin, 2009). Postprocessing and filtering of the alignment files was performed using Coval (Kosugi et al., 2013). The variants called after aligning the reads were then used to develop reference-based assembly of the parent (ICGV 02251) by substituting the bases with high confidence variants in the genome. The reads from HS and LS bulks for 100SW were then aligned to ICGV 02251 assembly and variants were called for HS and LS bulks by using the QTL-seq pipeline developed by the Iwate Biotechnology Research Center, Japan (Takagi et al., 2013).

2.4 | Calculation of SNP index to identify associated SNPs

The SNP index for each SNP at its respective genomic position was calculated for HS and LS bulks as per the formula given in Abe et al. (2012), Takagi et al. (2013), Pandey et al. (2017), and Luo et al. (2019a, 2019b). The SNPs at positions with high read depth of <10 in both the bulks and SNP index of <0.3 in either of the bulks were filtered out for calculation of ΔSNP index. Only SNP positions with ΔSNP index = −1 were considered as the causal SNPs and associated with 100SW. The possible effects of the identified SNPs were inferred using SnpEff v3.0 open-source program (Cingolani et al., 2012). Further, the annotations of SNPs associated with 100SW were searched on diploid and tetraploid genomes (www.peanutbase.org), and corresponding candidate genes affected by the causal SNPs were discovered.

2.5 | KASP marker development and validation

The SNPs with ΔSNP index = −1 affecting the function of important candidate genes were used for marker development. The SNPs were selected from the regions near to candidate genes in the four genomic regions on three different chromosomes. The developed Kompetitive allele specific PCR (KASP) markers were validated in contrasting germplasm lines from ICRISAT. In this context, the SNPs were converted into KASP markers by using 50-bp upstream and 50-bp
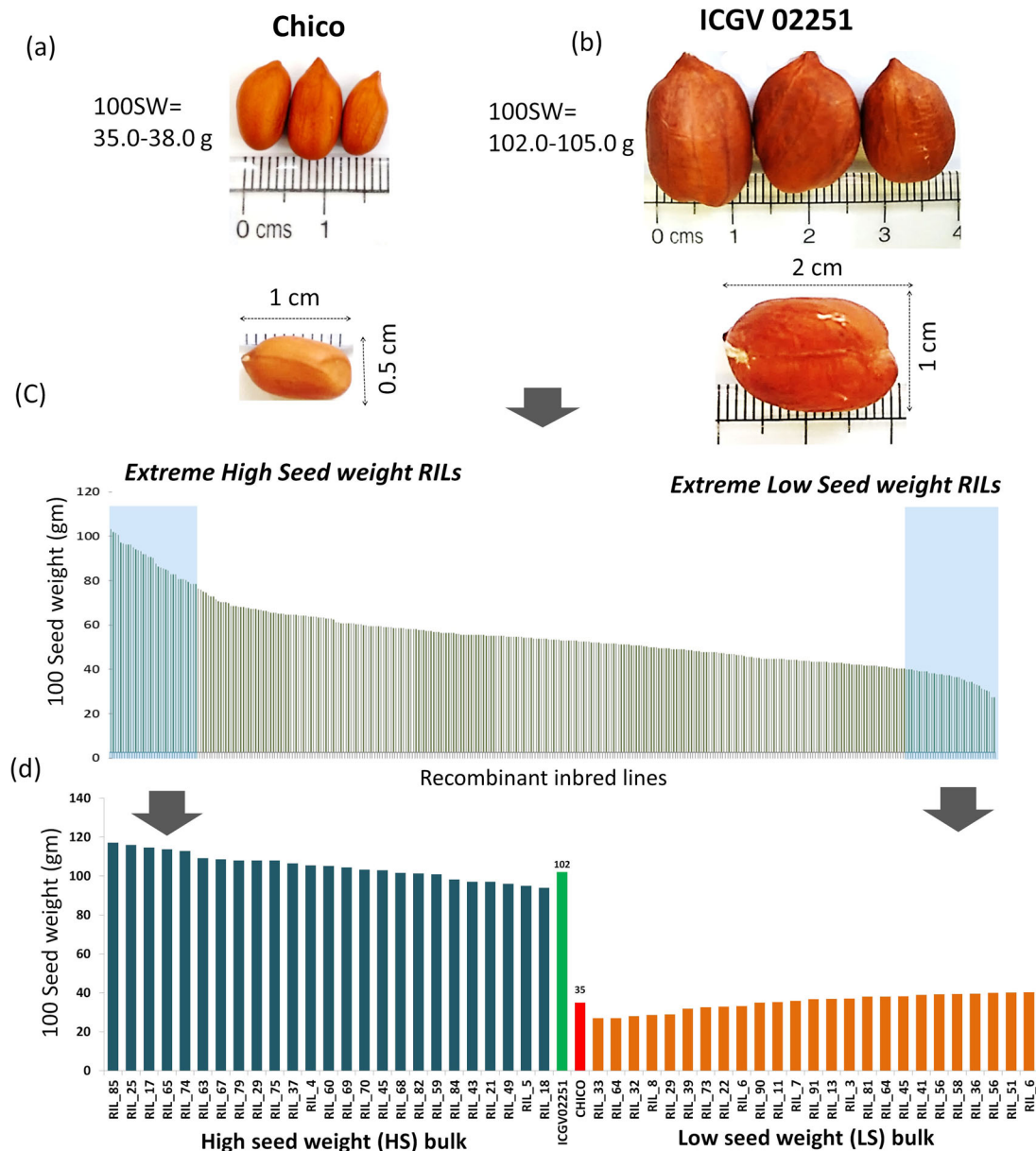
FIGURE 1 Construction of bulks with extreme phenotypes for 100-seed weight (100SW). (a) Chico, low-100SW parent for 100SW. (b) ICGV 02251, high 100SW parent for 100SW. (c) Frequency distribution for mean 100SW values of recombinant inbred line (RIL) population. These mean values were calculated based on phenotyping data generated in 2014 at ICRISAT. (d) Phenotypic variability among the RILs selected for development of extreme bulks for 100SW. Based on the phenotyping of RIL population, 25 high 100SW RILs and 25 low 100SW RILs were used to constitute low-seed weight and high-seed weight bulks

downstream sequences for the development of user-friendly and cost-effective markers (He et al., 2014). While designing the KASP assays for each SNP marker, two allele-specific forward primers and one common reverse primer was designed (Intertek Pvt. Ltd.). All the KASP primers used in this study are listed in Supplemental Table S1. The developed KASP markers were validated on a validation panel of 39 genotypes comprised of 18 bold-seeded, 16 small-seeded, and five F_1 generation lines derived from small- and bold-seeded genotypes (Supplemental Table S2).

3 | RESULTS

3.1 | Phenotyping and construction of extreme HS and LS bulks

A RIL population (Chico × ICGV 02251) comprising 350 RILs was phenotyped for 100SW. The 100SW of large-seeded parent ICGV 02251 was 98.2 ± 0.62 g, whereas that of small-seeded Chico was 35.3 ± 0.42 g (Figure 1a,b). The population showed high phenotypic variability for 100SW (Supplemental

TABLE 1 Summary of Illumina sequencing of parental line and bulks for 100-seed weight in groundnut

Sample ID	Total reads (raw)	Total reads after quality control	Data generated	Genome coverage	Average mapping depth (×)	Alignment percentage
			Gb	%		%
ICGV 02251	289,451,760	201,662,379	36.18	85.89	8.34×	93.93
HS bulk	329,428,944	233,019,623	41.18	85.90	9.27×	95.00
LS bulk	319,222,984	222,958,479	39.91	86.16	9.29×	95.12

Note. HS bulk, high-seed-weight bulk; LS bulk, low-seed-weight bulk.

Table S3), ranging from 28.7 to 117.2 g with normal distribution and partial transgressive segregation over both parents (Figure 1c). Based on the phenotyping data, 25 extreme RILs with high seed weight (90.3–117.2 g) and 25 RILs with low seed weight (28.7–41.5 g) were selected to construct HS and LS bulks (Figure 1d), respectively. The average 100SW of HS bulk was 102.1 g, while that of LS bulk was 37.6 g.

3.2 | Whole-genome sequencing of bulks and SNP discovery

Both bulks (HS and LS), along with bold-seeded parent ICGV 02251, were used for construction of Illumina sequencing libraries and sequenced using Illumina HiSeq 2500. A total of 289.4 million reads (36.2 Gb) were generated for bold-seeded parent ICGV 02251, 329.4 million reads (41.2 Gb) for HS bulk and 319.2 million reads (39.9 Gb) for the LS bulk (Table 1). After quality control analysis of high-quality reads, a total of 210.6 million reads for ICGV 02251, 233.0 million reads for HS bulk, and 222.9 million reads for LS bulk, they were aligned on diploid reference genomes (https://peanutbase.org/peanut_genome) (Supplemental Table S4). A reference-guided assembly for parent ICGV 02251 was developed, which achieved 93.9% alignment and 85.9% genome coverage on reference genome with 8.34× of average read depth. Quality reads of HS bulk were mapped on a reference-guided assembly and resulted in 95.0% alignment and 85.9% genome coverage with average depth of 9.27×. Similarly, the mapping of LS bulk achieved 95.1% alignment and 86.2% genome coverage with average depth of 9.29× (Table 1; Supplemental Table S4). The comprehensive sequence analysis for detection of sequence variants between HS and LS bulks identified a total of 2,22,423 genome-wide SNPs of which 106,850 SNPs (48% of total SNPs) were homozygous SNPs. The SNP index was calculated for all the SNPs detected between the two bulks and plotted against SNP position in the genome and represented graphically. Further, these SNP indices from the HS and LS bulks were used for calculation of the ΔSNP index (Supplemental Table S5, Supplemental Figures S1–S3).

3.3 | Candidate genomic regions for seed weight

The SNP index was calculated for HS and LS bulks to discover the genomic regions responsible for 100SW. The deviation of SNP index from 0.5 to 0 indicates the proportion of more alleles contributing from one parent, and 0.5 to 1.0 indicates the proportion of alleles contributing more from another parent. Three significant genomic regions linked to seed weight were identified on chromosomes B06 and one genomic region each on chromosome B08 and B09 using the criteria of a sliding window of 2.0-Mb intervals with 50-kb region increasing on physical map and ΔSNP index with a statistical confidence of $P < .01$ between the HS and LS bulks (Supplemental Figures S1–S3). Furthermore, after analyzing the sequences of LS and HS bulks with ICGV 02251 assembly as a reference, 104 SNPs in three genomic regions were identified on chromosome B06. Of these regions, the first genomic region of 8.9 Mb (31.57–40.52 Mb) harbored 55 SNPs, the second region of 1.1 Mb (53.56–54.66 Mb) had 21 SNPs, while the third genomic region of 1.64 Mb (81.17 –82.81 Mb) contained 28 SNPs. Similarly, one 2.64 Mb (10.95–13.59 Mb) region containing 48 SNPs was identified on chromosome B08, while a 930-Kb (145,797,595–146,728,008 bp) region harboring 30 nonsynonymous SNPs was identified on chromosome B09. The negative ΔSNP index indicates presence of alleles in the bulks from parental genome of ICGV 02251 (Figure 2). We also confirmed the presence of the associated genomic regions identified in the present study on chromosome B06 (logarithm of the odds = 4.9; percentage variance explained = 22.3%) and B09 (logarithm of the odds = 6.6; percentage variance explained = 10.6%) by performing genetic mapping and QTL discovery using the same RIL population (Supplemental Table S6).

3.4 | Discovery of highly significant SNPs and putative candidate genes for seed weight

The SNP index analysis using original genome assembly and ICGV 02251 assembly reference with combination of HS and
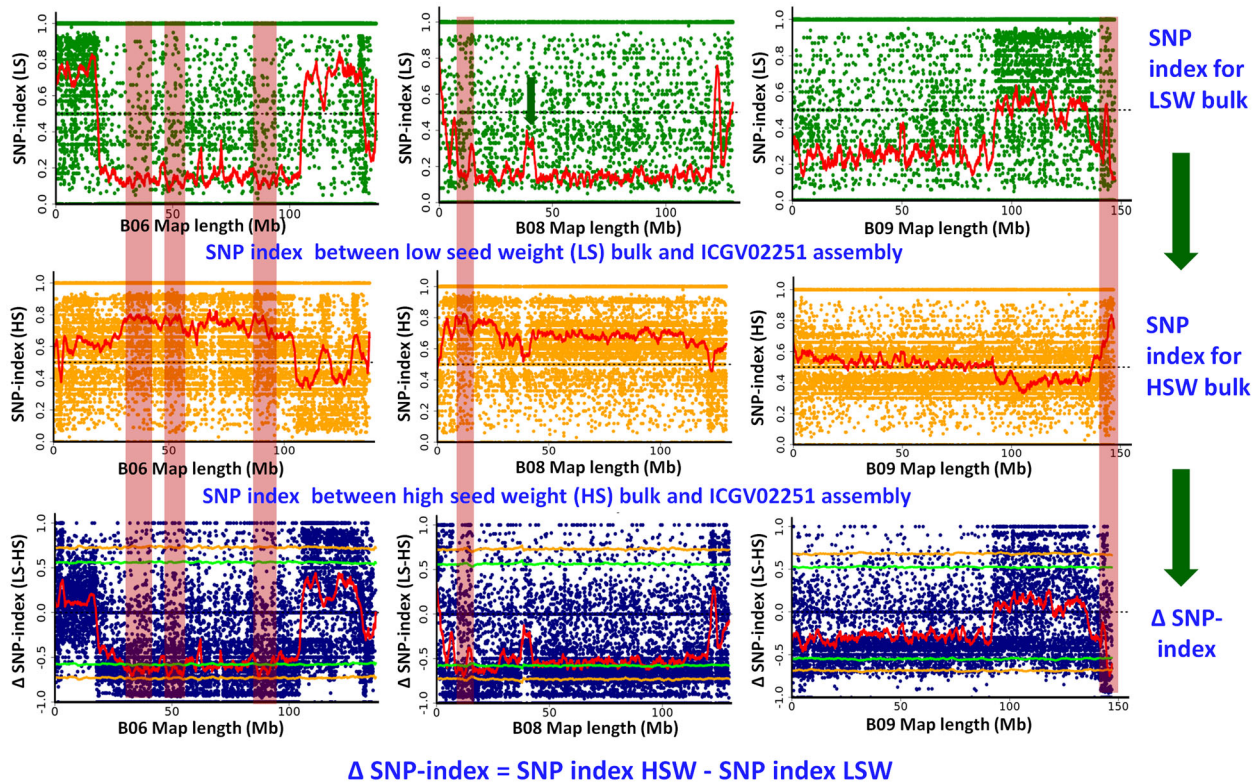
FIGURE 2 Quantitative trail locus-sequencing approach for mapping genomic regions controlling seed weight using the ICGV 02251 as reference parent. Single-nucleotide polymorphism (SNP) index plot between low bulk and ICGV 02251 assembly (green color), high bulk and ICGV02251 assembly (yellow color), and ΔSNP index plot (blue color) on chromosomes B06, B08, and B09 with statistical confidence interval under the null hypothesis of no QTL (orange, $P < .01$ and green $P < .05$). The associated genomic region is shaded in pink color on all three chromosomes

LS bulks detected three promising genomic regions on chromosome B06 associated with seed weight. These genomic regions harbored 104 effective SNPs with read depth ≥ 7; the SNP index significantly deviated from 0.5 and ΔSNP index -1, which is higher than the statistical confidence of $P < .01$ down to the line of 99% confidence interval at upper and lower side (Figure 2; Supplemental Table S7). The candidate genes identified based on diploid reference genomes were also found in corresponding genomic regions of the tetraploid reference genome (Supplemental Table S8). Of the 104 SNPs, 102 SNPs were located in intergenic region and two SNPs in the genic region (one each in intronic and exonic region). The SNP B06_54033720 was located in intronic region of a candidate gene *Ulp1 protease*, while the nonsynonymous SNP (B06_32002355) was located in the coding region of *unknown protein* (*Araip.K8UVH*) (Figure 3) according to the annotations of diploid as well as tetraploid genomes (www. peanutbase.com).

The genomic region of 2.6 Mb on chromosome B08 harbored 48 SNPs, and five of these SNPs were located in intronic regions of four different candidate genes. The first SNP (B08_10998096) was located in the genic region of candidate gene *Serine/threonine-protein kinase* (*Araip.UL7VH*),

while two SNPs, namely B08_11007755 and B08_11007758, were present in the single genic region *basic leucine zipper domain* (*Araip.IA4XE*). The fourth SNP was located at B08_11594032 in the candidate gene *uncharacterized protein* (*Araip.P0NXK*), while the fifth SNP, B08_11696813, was in genic region of a candidate gene *Araip.WJ0B5* on chromosome B08 coding for unknown protein (Figure 4).

A region of 930 Kb detected on chromosome B09 housed 30 SNPs, and six of these SNPs were found affecting the function of six candidate genes. The first SNP at B09_145870014 affected the function of a candidate gene *methyltransferase* (*Araip.0U7×2*), while second SNP located at B09_145954466 affected candidate gene *EPIDERMAL PATTERNING FACTOR* (*Araip.9SE5V*). A third SNP at position B09_146042834 affected candidate gene *Aluminum induced protein* with *YGL* and *LRDR* (*Araip.5A463*), while a fourth SNP (B09_146042834) affected the candidate gene *MADS-box transcription factor* (*Araip.K558L*). A fifth SNP (B09_146448322) affected the function of *TraB family protein* (*Araip.MI2NR*), while a sixth SNP (B09_146719055) affected the candidate gene *tripeptidyl peptidase* (*Araip.L3SN7*) (Figures 4 and 5).

FIGURE 3 Important genes identified in the genomic region on the chromosome B06. Magnified view of regions from peanut base harboring nonsynonymous single-nucleotide polymorphisms (SNPs) in various genic and intergenic regions. Nonsynonymous SNPs are indicated in vertical red dotted lines with identified SNP change. Seed and pod development related genes such as *Ulp1 proteases* (a SUMO conjugate), *auxin responsive factor*, *acetyl coenzyme A*, *growth regulating factor 4*, and some *unknown proteins* were located in the three genomic regions on chromosome B06

The expression of genes identified in the present study was studied using the recently developed groundnut gene expression atlas for the subspecies *fastigiata* (Sinha et al., 2020) and *hypogaea* (Clevenger et al., 2016) (Supplemental Tables S8 and S9). To check the specificity of gene expression in seed, we checked the expression levels of 49 candidate genes identified from three associated genomic regions located on chromosomes B06, B08, and B09 in the gene expression atlas. Of these 49 genes, 21 genes were highly expressed in flower, pod, and seed tissues. Three genes, namely *SIGNAL PEPTIDE PETIDASE*, *EPIDERMAL PATTERNING FACTOR*, and *WRKY* transcription factor, showed high expression in flower tissue. Four genes, including seed biotin containing protein, two isoforms of disease resistance *NBS-LRR proteins*, and *vicilin 47-kD protein TNP1*, were highly expressed in matured seed stage. Genes related to *TIFY family proteins*, *peroxidase super family proteins*, *pentatricopeptide repeats*, *glutamate synthase 1*, and *proteins kinases* showed high expression in developing seed. *Choline transporters* and *serine threonine ATP kinases* were highly expressed in embryo, while *ATP sulfurylase 1* was expressed in the matured pod wall (Figure 6).

3.5 | Development and validation of KASP markers for seed weight

Kompetitive allele specific PCR markers epitomized cost-effective and less laborious assays of genotyping in order to perform early generation selection in breeding lines for indirect selection of a particular phenotype (Bohar et al., 2020). To develop and validate diagnostic markers for seed weight, 14 SNPs (five SNPs on B06, five SNPs on B08, and four SNPs on B09) were targeted for development of KASP markers. Primers were successfully developed for 14 SNPs and validated on a validation panel with contrasting seed weight (Supplemental Table S1). The validation panel was comprised of bold-seeded lines with 100SW ranging from 75 to 80 g and small-seeded lines with 100SW of 25–30 g. Five F_1 populations derived from different crosses were also included
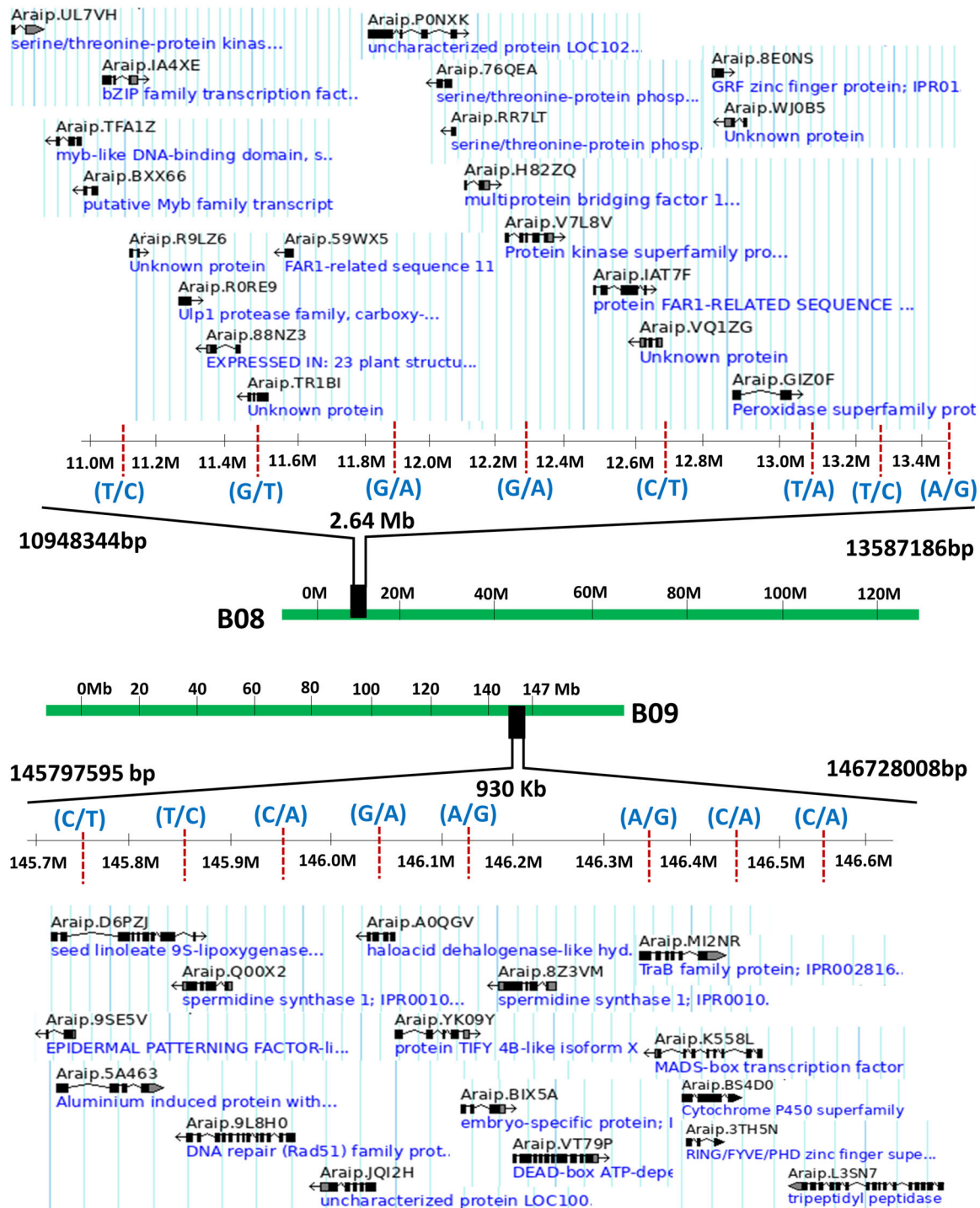
FIGURE 4 Important genes identified in the genomic region on the chromosomes B08 and B09. Magnified view from peanut base of a region harboring nonsynonymous single-nucleotide polymorphisms (SNPs) in various intergenic regions. Nonsynonymous SNPs are indicated in vertical red dotted lines with identified SNP change. Seed and pod development related genes such as *serine/threonine protein phosphatases* and *serine/threonine protein kinases* were identified on chromosome B08 in the 2.64-Mb genomic region. Whereas, spermidine synthase and seed linoleate *MADS box* and *DEAD box* genes were located in the 930-Kb region on chromosome B09

in the panel to study the robustness of markers to detect heterozygosity (Supplemental Table S2). Of the 14 KASP markers validated, four KASPs (snpAH0031, snpAH0033, snpAH0037, and snpAH0038) showed expected polymorphism in the germplasm (Table 3). Interestingly, all these

four polymorphic KASP markers differentiating genotypes with small- and bold-seeded genotypes belonged to genomic region detected on chromosome B06 (Figure 7). The developed KASP markers were highly polymorphic between bold- and small-seeded groundnut genotypes and can be used as
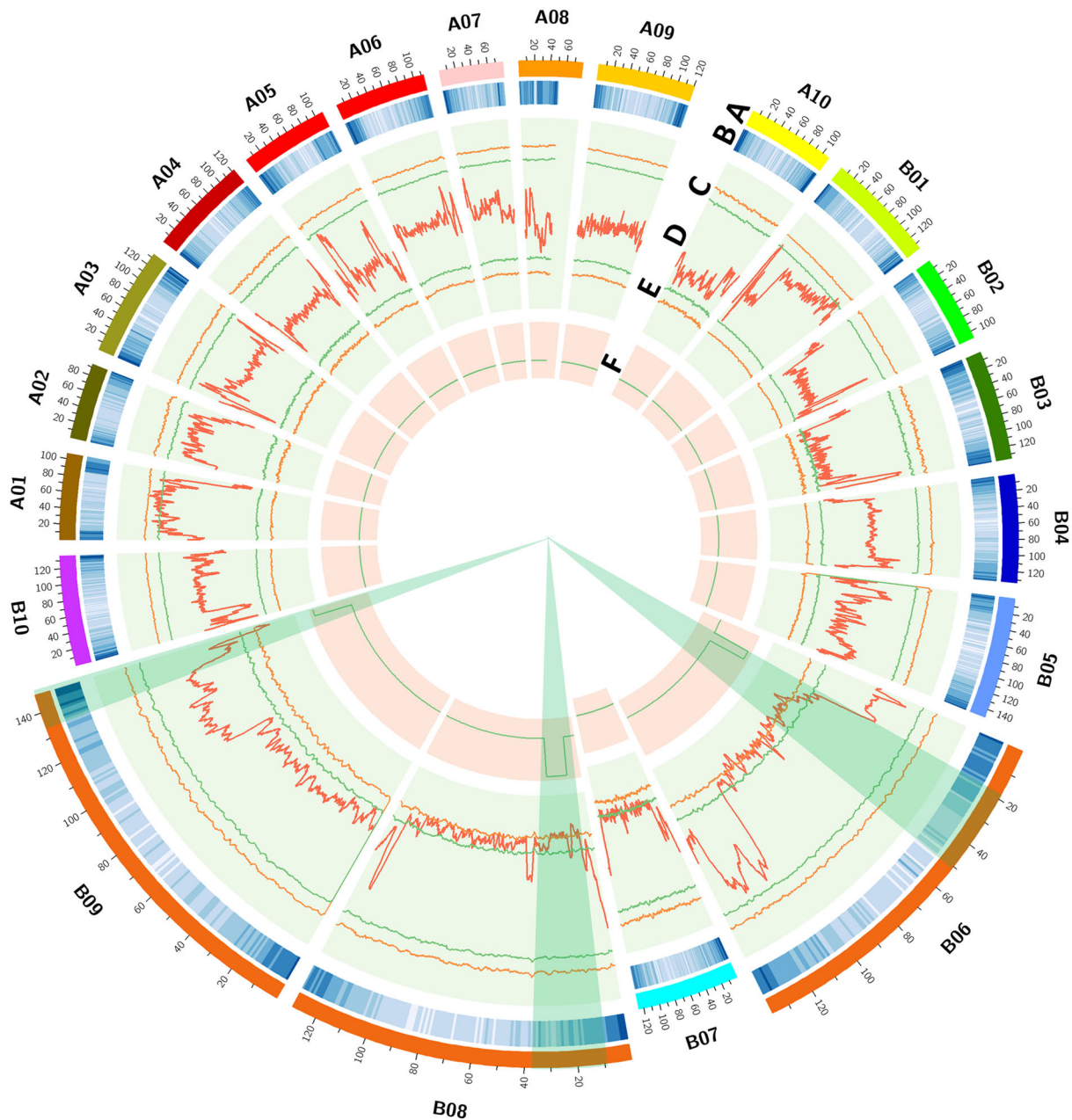
FIGURE 5 Circos plot represents genome-wide ΔSNP index on 20 chromosomes of groundnut colocalized in previously reported quantitative trait loci (QTL) from traditional genetic mapping. In the illustration, the tracks from outside to inside indicate (a) chromosomes of reference genome *Arachis duranensis* and *A. ipaensis*; (b) genome-wide density of annotated genes on groundnut genomes; the density of genes is higher in telomere than centromeric region of each chromosome; (c) upper probability values at 99% confidence and upper probability values at 95% confidence; (d) ΔSNP index plot using the ICGV 02251 assembly as reference; (e) lower probability values at 95% confidence and lower probability values at 99% confidence; and (f) physical position of earlier mapped QTL for groundnut seed weight through nested association mapping based genome-wide association study and genetic mapping (Gangurde et al., 2020)

potential diagnostic markers for selecting bold-seeded segregating breeding material at the very early stage of varietal development process. In addition, the KASPs snpAH0039, snpAH0041, and snpAH0042 have moderately distinguished the genotypes on validation panel and showed specificity to certain genetic backgrounds to be used as diagnostic markers for seed weight. All validated KASP markers successfully

confirmed F_1 population lines derived from crosses of various small- and bold-seeded lines.

4 | DISCUSSION

Next-generation sequencing technologies have revolutionized trait dissection, candidate gene discovery and have helped
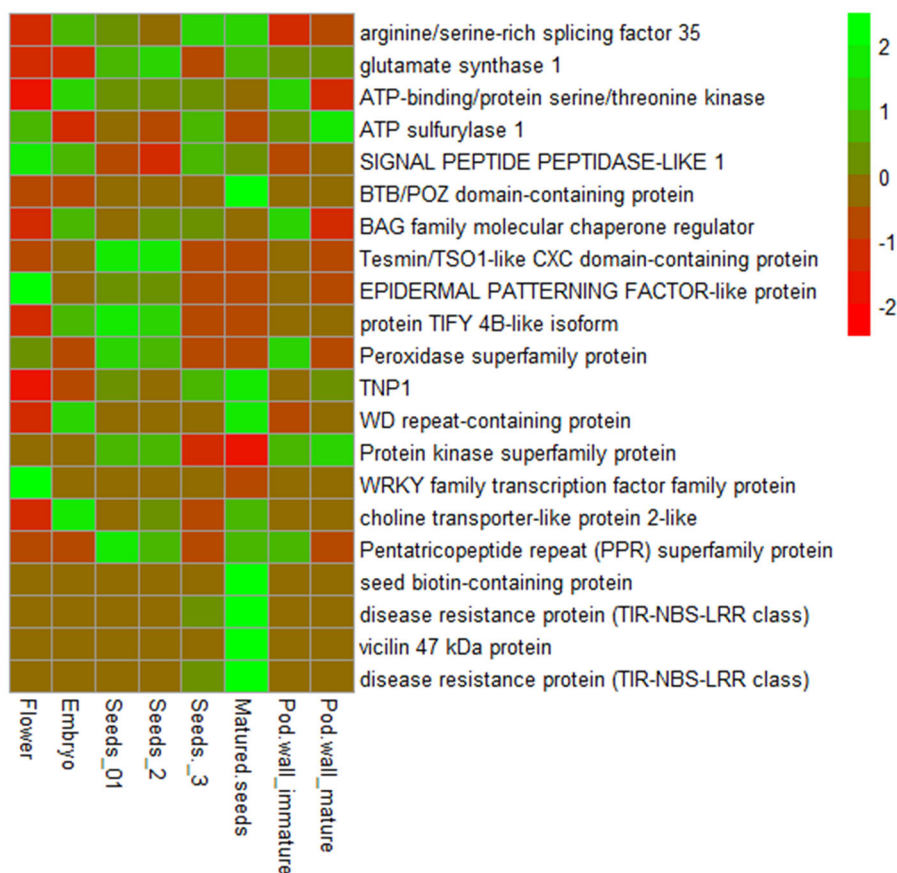
FIGURE 6 Expression of candidate genes especially expressed in flower, embryo, seed, and pod developmental stages based on gene expression atlas. Twenty-one genes were expressed especially in seed, pod, embryo, and flower. The gene expression data was extracted from groundnut subspecies *fastigiata* gene expression (AhGEA) atlas (Sinha et al., 2020)

in developing cost-effective genomic tools. Groundnut has entered in postgenome sequencing era with huge genomic resources (see Pandey et al., 2012, 2020). Availability of a high-quality reference genome for both the progenitor diploid species (*A. duranensis* and *A. ipaensis*) (Bertioli et al., 2016; Chen et al., 2016) and subspecies of cultivated tetraploid groundnut (*A. hypogaea* subspp. *hypogaea* by Bertioli et al., 2019 and *A. hypogaea* subspp. *fastigiata* by Chen et al., 2019b and Zhuang et al., 2019). These genomic resources provided opportunity for developing genotyping assays (see Pandey et al., 2012; Varshney et al., 2013) in addition to sequencing-based gene discovery (Pandey et al., 2016; Varshney et al., 2019). In addition to high-quality reference genomes, global gene expression atlas has also become available for both the subspecies namely subspp. *fastigiata* (Sinha et al., 2020) and subspp. *hypogaea* (Clevenger et al., 2016).

Sequencing-based trait mapping approaches, such as QTL-seq, provides several advantages over traditional genetic mapping approaches as it leads to candidate gene discovery as well as diagnostic marker development. Following the principles of bulked segregant analysis (Michelmore et al., 1991),

whole genome sequencing of pooled samples of extreme phenotypes in the QTL-seq approach have been deployed earlier in groundnut, which facilitated faster discovery of candidate genes and markers for several important traits in groundnut including resistance to foliar fungal disease (Pandey et al., 2017; Clevenger et al., 2018) and bacterial wilt (Luo et al., 2019a), shelling percentage (Luo et al., 2019b), fresh seed dormancy (Kumar et al., 2020), and purple testa color (Zhao et al., 2020). Taking benefit of these developments and availability of reference genome sequences, we successfully used the QTL-seq approach and identified genomic regions, candidate genes, and, more importantly, developed four diagnostic KASP markers associated with groundnut seed weight.

For performing the QTL-seq experiment, we obtained 8.34–9.29× average read depth of sequencing data realizing the large genome size (2.7 GB) of cultivated tetraploid groundnut. Previous QTL-seq experiments in groundnut used 10.3–11.6× average read depth for resistance to rust and late leaf spot (Pandey et al., 2017), ~6.5× average depth for fresh seed dormancy (Kumar et al., 2020), while 15.6× for shelling percentage (Luo et al., 2019a) and 14.3× for bacterial wilt resistance (Luo et al., 2019b). The QTL-seq experiments

(a)

DNA \ Assay	SD_WT		snpAH0031	snpAH0033	snpAH0035	snpAH0037	snpAH0038	snpAH0039	snpAH0040	snpAH0041	snpAH0042	snpAH0043	snpAH0047	snpAH0048	snpAH0049	snpAH0050
ICGV 07368	75-80	Bold seeded lines	G	T	G	C	T	A	T	Y	G	T	T	A	T	T
ICGV 10200	75-80		G	T	G	C	T	A	Y	C	G	T	C	A	T	T
ICGV 10207	75-80		G	T	G	C	T	A	T	Y	G	T	C	A	T	T
ICGV 10208	75-80		G	T	G	C	T	A	T	Y	G	T	C	A	T	T
ICGV 10209	75-80		G	T	G	C	T	A	T	C	A	?	T			
ICGV 10213	75-80		G	T	G	C	T	A	T	C	G	T	C	G	C	T
ICGV 10216	75-80		G	T	G	C	T	A	T	C	G	T	C	G	C	T
ICGV 10217	75-80		G	T	G	C	T	A	T	Y	G	T	C	G	C	T
ICGV 10218	75-80		G	T	G	C	T	A	T	C	G	T	C	G	C	T
ICGV 10220	75-80		G	T	G	C	T	A	T	G	T	C	A	Y	T	
ICGV 11310	75-80		G	T	G	C	T	A	T	C	G	T	C	A	Y	T
ICGV 11311	75-80		G	T	G	C	T	A	Y	T	G	T	-	A	T	T
ICGV 11314	75-80		G	T	G	C	T	A	T	C	G	T	C	A	T	T
ICGV 16314	75-80		G	T	G	C	T	A	T	C	G	T	C	A	T	T
ICGV 16316	75-80		G	T	G	C	T	A	T	C	G	T	C	A	T	T
ICGV 16318	75-80		G	T	G	C	T	A	T	C	G	T	C	A	T	T
ICGV 00440	75-80		G	T	G	C	T	A	T	C	G	T	C	A	T	T
ICGV 98412	75-80		G	T	G	C	T	A	T	C	G	T	C	A	T	T
ICGV 10001 X ICGV 10200	F1	F1s	R	Y	K	M	Y	W	Y	?	K	Y	-	A	T	T
ICGV 10007 X ICGV 10209	F1		R	Y	K	M	Y	A	Y	C	G	T	C	A	Y	T
ICGV 10008 X ICGV 10213	F1		R	Y	K	M	Y	W	T	K	Y	C	-	Y	T	
ICGV 10016 X ICGV 10216	F1		R	Y	K	M	Y	W	T	T	K	T	G	C	W	
ICGV 13027 X ICGV 10218	F1		R	Y	K	M	Y	W	T	K	Y	C	G	C	W	
ICGV 16017	30	Small seeded lines	G	T	G	C	C	A	T	T	G	C	A	Y	T	
ICGV 16668	30		A	C	K	A	C	A	Y	C	G	T	C	A	?	T
ICGV 16669	30		A	C	K	A	C	A	T	C	G	T	C	A	T	T
ICGV 16673	30		A	C	K	A	C	A	T	C	G	T	C	A	T	T
ICGV 16687	30		A	C	K	A	C	A	T	C	G	T	C	A	T	T
ICGV 16700	30		G	T	G	C	C	A	Y	Y	G	T	C	A	T	T
ICGV 15404	30		A	C	K	A	C	T	T	T	T	T	C	G	C	T
ICGV 15416	30		A	C	K	A	C	T	T	T	T	T	C	G	C	T
ICGV 15417	30		A	C	K	A	C	T	T	T	T	T	C	G	C	T
ICGV 15423	30		A	C	K	A	C	T	T	T	T	T	C	G	C	T
ICGV 15424	30		A	C	K	A	C	T	T	T	T	T	C	G	C	T
ICGV 15425	30		A	C	K	A	C	T	T	T	T	T	C	G	C	T
ICGV 15426	30		A	C	K	A	C	T	T	T	T	T	C	G	C	T
ICGV 00350	35		A	C	K	A	C	-	C	T	T	T	C	G	C	T
ICGV 00351	35		A	C	K	A	C	T	T	T	T	T	C	G	C	T
ICGV 02266	38		A	C	K	A	C	T	T	T	T	T	C	G	C	T

(b) Genomics SNP 4.0.0.0 — Project number 8.053, KSNP number 008-0018, BLANK — snpAH0031
1.71
A:A, A:G, G:G, NTC, Missing, Bad, Short, Dupe, Over, Uncallable

(d) Genomics SNP 4.0.0.0 — Project number 8.052, KSNP number 008-0015, BLANK — snpAH0037
2.17
C:C, C:A, A:A, NTC, ?, Missing, Bad, Short, Dupe, Over, Uncallable

(c) Genomics SNP 4.0.0.0 — Project number 8.053, KSNP number 008-0010, BLANK — snpAH0033
1.84
T:T, T:C, C:C, NTC, ?, Missing, Bad, Short, Dupe, Over, Uncallable

(e) Genomics SNP 4.0.0.0 — Project number 8.056, KSNP number 008-0011, BLANK — snpAH0038
1.53
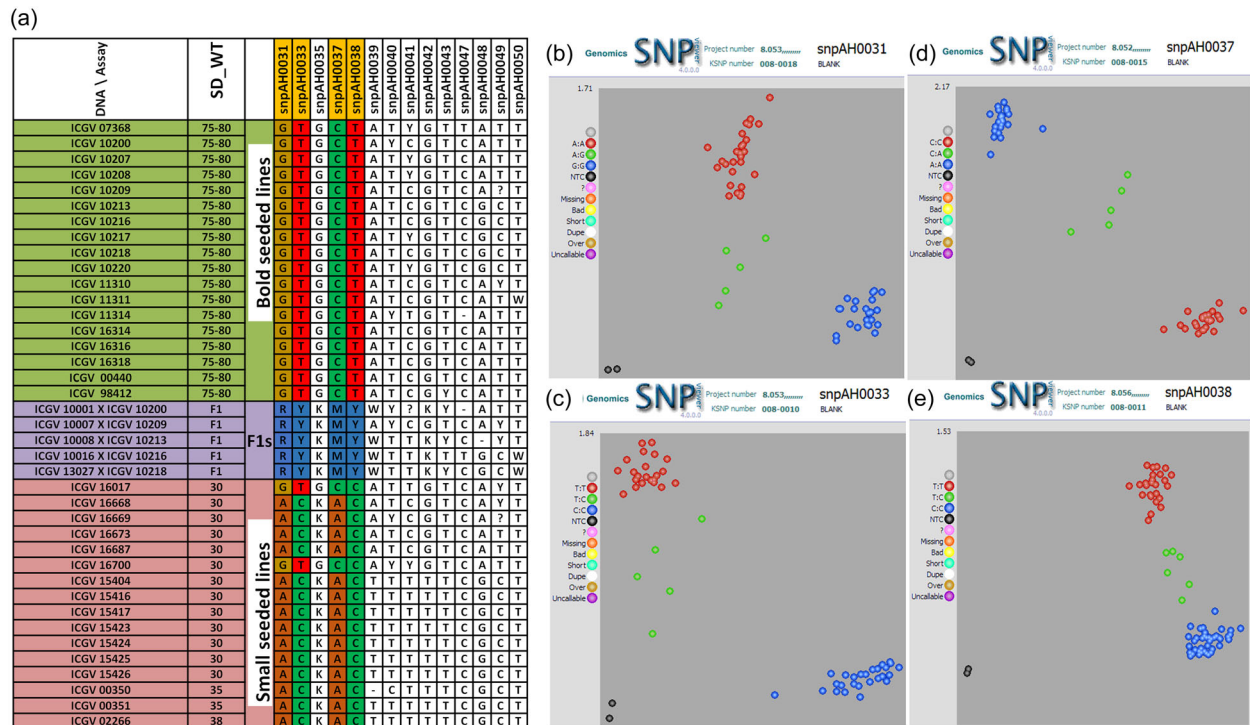T:T, T:C, C:C, NTC, ?, Missing, Bad, Short, Dupe, Over, Uncallable

FIGURE 7 Development and validation of Kompetitive allele specific polymerase chain reaction (KASP) markers from potential candidate genes identified for seed weight. (a) Validation panel includes 18 bold-seeded groundnut varieties (75–80 g 100-seed weight), 16 small-seeded groundnut varieties (25–30 g 100-seed weight), and five F_1 hybrids derived from different combinations of bold- and small-seeded groundnut varieties. Four single-nucleotide polymorphisms (SNPs) show clear homozygous clusters and one heterozygous cluster for heterozygotes for (b) snpAH0031 for gene *Ulp1 protease* (cysteine-type peptidase activity) (*Araip.5J71F*), (c) snpAH0033 (intergenic SNP between *Araip.A3T8C* and *Araip.A5MIH*) for nearest gene ATP binding; *nucleoside-triphosphatase*, (d) snpAH0037 (intergenic SNP between *Araip.VA5I5* and *Araip.9M8Z9*) for nearest gene protein kinase superfamily protein, (e) snpAH0038 (intergenic SNP between *Araip.9M8Z9* and *Araip.FJY6W*) for nearest gene protein kinase superfamily protein

performed in diploid crop species, such as chickpea (*Cicer arietinum* L.), cucumber (*Cucumis sativus* L.), and pigeonpea (*Cajanus cajan* L. Huth) with small to medium genome size, were conducted based on just 3.7–6.52× average depth (Das et al., 2015; Lu et al., 2014; Singh et al., 2016). Therefore, the results in the present study reconfirm that 9–10× coverage is sufficient for performing QTL-seq analysis in groundnut.

To understand the superiority of the QTL-seq approach over genetic mapping, the identified genomic regions for seed weight were compared with previous genetic mapping studies conducted to identify the genomic regions associated with seed weight. In the present study, we identified genomic regions on chromosome B06, B08, and B09 associated with seed weight. Recently, a nested-association mapping approach identified three and two SNPs on chromosome B06 and B09, respectively, associated with pod weight and seed weight (Gangurde et al., 2020). However, in the present study, we identified a total of 104 SNPs in three genomic regions on B06 and 30 SNPs in the genomic region on chromosome B09. Several studies reported consistent major-effect QTL for seed weight on chromosome B06 (Zhang et al., 2019; Wang et al., 2018; Huang et al., 2015). Here, we discovered three genomic

regions of sizes 8.9 (31.5–40.52 Mb), 1.1 (53.56–54.66 Mb), and 1.64 Mb (81.17–82.81 Mb) on chromosome B06 with a total of 104 SNPs, indicating a plausible role of chromosome B06 in regulation of seed weight. Shelling percentage and seed weight are interdependent traits; genomic regions and candidate genes on chromosome A09 and B02 were identified for shelling percentage using the QTL-seq approach (Luo et al., 2019b). More importantly, biparental genetic mapping in the same population could identify only one major and consistent QTL on chromosome A09 for shelling percentage (Luo et al., 2017). A recent study using nested-association mapping for seed and pod weights identified associated genomic regions on chromosomes A05 and B05, A06 and B06, A08 and B08, and B09 (Gangurde et al., 2020), which may be due to contribution from multiple parents. Nevertheless, in the present study, a very narrow genomic region of 930 Kb was identified on chromosome B09 while analyzing reference genome with HS and LS pools. Another genetic mapping study reported major-effect and consistent QTL on chromosome B07 and B08 for seed weight (Mondal et al., 2015), which matches with the genomic region 2.63 Mb (10.94–13.58 Mb) identified in the present study on chromosome

TABLE 2 Single-nucleotide polymorphism (SNPs) with ΔSNP index = −1 on chromosome B06, B08, and B09 identified using high-seed-weight (HS) and low-seed-weight (LS) bulk and ICGV02251 assembly with corresponding genes

Chromosome	Physical position	ICGV 02251 base	LS base	HS base	Gene interval for intergenic SNPs	Feature of nearest gene
	bp					
Araip.B06	32,044,422	T	C	T	Araip.0FI9Y–Araip.NVZ99	MYB transcription factor MYB118
	32,152,814	C	T	C	Araip.KLU9X–Araip.2412K	Adenylate kinase family protein
	32,349,226	G	A	G	Araip.2412K–Araip.WIZ5S	WRKY family transcription factor family protein
	32,489,453	C	T	C	Araip.M1NAC–Araip.7L5D1	Unknown protein
	33,164,677	G	T	G	Araip.U0213–Araip.TE1P2	CDPK-related kinase 3
	33,884,746	A	G	A	Araip.U9R7I–Araip.K48V4	Uncharacterized protein
	53,557,434	G	T	G	Araip.L9809–Araip.V79U1	WD repeat-containing protein
	54,290,325	A	T	A	Araip.116MM–Araip.AEB2X	Triosephosphate isomerase
	54,401,527	G	T	G	Araip.AEB2X–Araip.97NYT	TNP1
	81,630,940	C	G	C	Araip.8DG38–Araip.7PD4P	Growth-regulating factor 4
	81,889,671	A	T	A	Araip.RVJ8S–Araip.VA5I5	BTB/POZ domain-containing protein
	82,026,094	T	C	T	Araip.9M8Z9–Araip.FJY6W	Protein kinase superfamily protein
Araip.B08	11,256,389	G	A	G	Araip.ZL8H3–Araip.AC1PS	Fe(II)-dependent oxygenase superfamily protein
	11,815,376	T	G	T	Araip.DK4JW–Araip.D4EPK	Seed biotin-containing protein
	11,840,276	G	A	G	Araip.D4EPK–Araip.LJQ0L	Protein YLS7
	12,062,456	G	A	G	Araip.I0QD1–Araip.FM3FX	Exocyst complex component sec10
	12,188,146	C	T	C	Araip.C3D0C–Araip.CMJ6A	ATP-binding/protein serine/threonine kinase
	12,787,501	G	A	G	Araip.GIZ0F–Araip.ATS2M	Peroxidase superfamily protein
	13,084,370	G	A	G	Araip.01CTA–Araip.PIP5Q	Disease resistance protein (TIR-NBS-LRR class)
	13,243,168	G	A	G	Araip.6EA1G–Araip.AH9TM	Pentatricopeptide repeat (PPR) superfamily protein
	13,403,557	G	C	G	Araip.EF3LF–Araip.Z2IAU	Glutamate synthase 1
Araip.B09	145,797,595	C	T	C	Araip.YWU3B–Araip.M2BYP	Uncharacterized protein
	145,810,717	A	T	A	Araip.T82B5–Araip.BY68W	Vicilin 47 kDa protein
	145,852,122	T	C	T	Araip.BY68W–Araip.0U7×2	Homocysteine S-methyltransferase 3
	145,880,280	G	A	G	Araip.0Y9HM–Araip.45TAK	Tesmin/TSO1-like CXC domain-containing protein
	145,924,830	C	T	C	Araip.Q0MMZ–Araip.ZG2VU	Choline transporter-like protein 2-like
	145,959,066	C	T	C	Araip.9SE5V–Araip.7IG36	EPIDERMAL PATTERNING FACTOR-like protein

(Continues)

TABLE 2 (Continued)

Chromosome	Physical position	ICGV 02251 base	LS base	HS base	Gene interval for intergenic SNPs	Feature of nearest gene
	145,969,360	T	C	T	*Araip.7IG36–Araip.YK09Y*	Protein TIFY 4B-like isoform
	145,995,740	G	A	G	*Araip.UP18N–Araip.24AK5*	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase
	146,102,002	A	G	A	*Araip.I8ABE–Araip.J63ZS*	Argonaute family protein
	146,161,582	C	T	C	*Araip.21M98–Araip.A561Y*	Alpha-L-fucosidase 1
	146,430,068	G	A	G	*Araip.G4AXZ–Araip.MI2NR*	Zinc finger CCCH domain-containing protein
	146,460,214	C	T	C	*Araip.ZB3J2–Araip.B0R6B*	Arginine/serine-rich splicing factor 35
	146,651,299	G	A	G	*Araip.N2U4Q–Araip.C19HK*	Actin depolymerizing factor 7
	146,728,008	G	A	G	*Araip.L3SN7–Araip.U4GR2*	Tripeptidyl peptidase

B08. Of the two genetic mapping studies reported so far in groundnut, one study using biparental population identified nine minor effect QTLs for seed and pod size located on the chromosomes A05, A06, A09, B10, B04, A03, B05, and B08 (Chu et al., 2020). Recently, with tetraploid genome assembly, the regions associated with groundnut seed weight were identified on chromosome A07 and B02 (Zhuang et al., 2019).

Keeping in mind the importance of information on candidate genes and significantly associated SNPs for marker development, we targeted associated regions located on chromosomes B06, B08, and B09. The 13 SNPs discovered from these regions affected the function of nine known candidate genes, namely *Ulp protease*, *serine/threonine-protein kinase*, *basic leucine-zipper domain*, *methyltransferase*, *EPIDERMAL PATTERNING FACTOR*, *aluminium induced protein with YGL and LRDR*, *MADS-box transcription factor*, *TraB family protein*, and *tripeptidyl peptidase*. The *Ulp* (ubiquitin like proteases) *proteases (Araip.5J71F)* gene identified on chromosome B06 are *SUMO (small ubiquitin-like modifier) specific cysteine (Cys) proteases* with *SUMO peptidase* activity, which are made by posttranslational modifications and results in dynamic changes to protein function (Augustine and Vierstra, 2018; van den Burg et al., 2010; Gareau and Lima, 2010). It is reported that 40 SUMO proteins showed higher expression of *SUMO1/2* and *SCE1 (SUMO Conjugating Enzyme 1)* genes during pod development in groundnut. In addition, structural modifications among SUMO conjugates were also observed in response to abiotic stresses (Liu et al., 2019). On chromosome B09, one SNP was identified at position B09_145954466 located in genic region of *EPIDERMAL PATTERNING FACTOR* (EPF) (*Araip.9SE5V*). *TaEPFL1* (*Triticum aestivum* epidermal patterning factor 1) gene was highly expressed in abnormal pistillody sta-

mens as compared with pistil and stamens in *HTS1* mutant of wheat (*Triticum aestivum* L.) (Sun et al., 2019). An isoform of EPF, *BnEPFL6* (*Brassica napus epidermal patterning factor 6*) gene was found essential for the elongation of filaments in rape (*Brassica napus L.*) (Huang et al., 2014). We identified one SNP (B08_10998096) in genic region of serine/threonine protein kinase (*Araip.UL7VH*) on chromosome B08 (B08_10998096), which is known to be involved in the plant growth and development in *Arabidopsis thaliana* (L.) Heynh. (Uhlken et al., 2014). A SNP (B06_81932178) was identified in the intergenic region near to protein kinases (*Araip.9M8Z9*). It is well accepted that receptor-like protein kinases, such as *HAESA*, are involved in floral organ development (Jinn et al., 2000; Taylor et al., 2016). A receptor-like protein kinase coding brassinosteroid gene *BRI1* regulates cell growth in *Arabidopsis* (Li and Chory, 1997). *CLAVATA1* genes are receptor kinase coding genes involved in meristem development, cell proliferation, and differentiation (Clark et al., 1997). On chromosome B08, two SNPs (B08_11007755 and B08_11007758) were identified in the genic region of basic leucine zipper domain (bZIP) transcription factor (*Araip.IA4XE*), which are associated with essential processes like seed development, energy balance, and abiotic or biotic stresses resistance (Noman et al., 2017). Another SNP (B09_146312888) on chromosome B09 was identified in genic region of *MADS box-domain* transcription factors (*Araip.K558L*), which are basic members of regulatory networks underlying multiple developmental pathways in plants, animals, and fungi (Nakashima & Yamaguchi-Shinozaki, 2013; Janiak et al., 2016; Wils & Kaufmann, 2017; Cho, 2018). In *Arabidopsis*, *MADS-box* genes are involved in developmental processes such as meristem specification, flowering transition, and seed, root, and

TABLE 3 Summary of Kompetitive allele specific polymerase chain reaction (KASP) markers developed from significant single-nucleotide polymorphisms (SNPs) in the genomic regions of important candidate genes for seed weight

KASP marker	Chr	Position (bp)	'Chico', low bulk base	ICGV02251, high bulk base	Gene ID	Feature	Genic or intergenic	Position in tetraploid genome Chr	Start (bp)	End (bp)
snpAH0031	B06	32,002,355	A	G	Araip.5J71F	Ulp1 protease (cysteine-type peptidase activity)	Genic	Ahy16	37,070,658	37,070,214
snpAH0033	B06	33,002,322	C	T	Araip.A3T8C–Araip.A5MIH	ATP binding; nucleoside-triphosphatase	Intergenic	Ahy16	37,848,691	37,849,457
snpAH0035	B06	54,033,720	T	G	Araip.K8UVH	Unknown Protein	Genic	Ahy16	61,321,826	61,322,309
snpAH0037	B06	81,932,178	A	C	Araip.VA5I5–Araip.9M8Z9	Protein kinase superfamily protein	Intergenic	Ahy16	94,695,418	94,694,858
snpAH0038	B06	82,423,709	C	T	Araip.9M8Z9–Araip.FJY6W	Protein kinase superfamily protein	Intergenic	Ahy16	94,739,497	94,737,483
snpAH0039	B08	10,998,096	T	A	Araip.UL7VH	Serine/threonine-protein kinase	Genic	Ahy18	12,153,016	12,154,671
snpAH0040	B08	11,007,755	C	T	Araip.IA4XE	Basic-leucine zipper domain	Genic	Ahy18	12,162,664	12,163,477
snpAH0041	B08	11,007,758	T	C	Araip.IA4XE	Basic-leucine zipper domain	Genic	Ahy18	12,162,664	12,163,477
snpAH0042	B08	11,594,032	T	G	Araip.P0NXK	Uncharacterized protein	Genic	Ahy18	12,745,717	12,746,277
snpAH0043	B08	11,696,813	C	T	Araip.WI0B5	Unknown Protein	Genic	Ahy18	12,842,397	12,841,709
snpAH0047	B09	145,954,466	T	C	Araip.9SE5V	EPIDERMAL PATTERNING FACTOR	Genic	Ahy19	157,520,950	157,520,709
snpAH0048	B09	146,312,888	G	A	Araip.K558L	MADS-box transcription factor	Genic	Ahy19	157,884,580	157,884,310
snpAH0049	B09	146,448,322	C	T	Araip.MI2NR	TraB family protein	Genic	Ahy19	158,029,721	158,030,576
snpAH0050	B09	146,719,055	A	T	Araip.L3SN7	Tripeptidyl peptidase	Genic	Ahy19	158,283,152	158,282,550

flower development (Smaczniak et al., 2012). Three SNPs (B09_145852122, B09_145861466, and B09_145867485) were identified in intergenic region of *homocysteine S-methyltransferase*, known to be involved in seed development. Reciprocal cross between the wild type and met1-6 mutant carrying a mutation in *DNA METHYLTRANSFERASE 1 (MET1)* gene results in hypomethylated maternal genomes cause significantly larger, while paternal genomes cause smaller F_1 seeds (Xiao et al., 2006). A recessive mutation in another gene that dramatically reduces DNA methylation, *DECREASE IN DNA METHYLATION1*, also causes parent-of-origin effects on F_1 seed size (Table 2). Based on the possible function of the associated candidate genes with seed weight, an epigenetic change in the genome seems to play important role in defining the seed size in groundnut.

After identifying the candidate genes, we have checked the tissue specificity of these genes based on the expression information available through gene expression atlas for subspp. *fastigiata* (Sinha et al., 2020) and *hypogaea* (Clevenger et al., 2016). Among 182 SNPs, 169 were identified adjacent (intergenic) to 48 candidate genes involved in various physiological processes during plant development. Upon comparing the expression profile of 49 genes in various seed and pod developmental stages using groundnut gene expression atlas (Sinha et al., 2020) (Supplemental Figure S4; Table 2; Supplemental Tables S6 and S7), 21 genes were found to be differentially expressed during embryo, seed, and pod development (Figure 6). Seed biotin containing protein, disease resistance *NBS-LRR protein*, *vicillin*, *TNP1*, and *WD repeat* containing protein genes were expressed in mature seeds. Whereas transcription factors *WRKY*, *EPIDERMAL PATTERNING FACTOR*, and signal peptide peptidase were highly expressed in flower tissue. *Pentatricopeptide repeats*, *TIFY family proteins*, *Tesmin/TSO1 domains*, *glutamate synthase*, and *protein kinases* are specifically expressed in seeds. Embryo is also an important tissue in the developing seed. *Choline transporters*-like proteins are specifically expressed in developing embryo tissue. Very few genes were expressed in pod walls with moderate expression values. *Protein kinases*, *ATP sulfurylase*, and *peroxidase superfamily* proteins were found to be expressed in pod walls. In the present study, we identified a sharp peak on chromosome B09 that housed *TIFY family proteins*. The homologues of *BIG SEED* locus belong to *TIFY family proteins* on chromosome B09. BIG SEED 1 (*BS1*) and BIG SEED 2 (*BS2*) loci encodes *TIFY family proteins* reported in barrel clover (*Medicago truncatula* Gaertn.) as negative regulators of seed weight (Ge et al., 2016).

Diagnostic markers accelerate the process of developing improved varieties through facilitating selection using marker-assisted early generation selection based on single-seed chipping and rapid generation advancement (Parmar et al., 2021). Deployment of diagnostic markers, including use of KASP markers, for resistance to nematode, leaf rust, and late leaf spot, in addition to high oleic acid content is now been routinely used for improving foliar disease resistance and oil quality using marker-assisted selection (Chu et al., 2011; Varshney et al., 2014; Janila et al., 2016; Kolekar et al., 2017; Bera et al., 2018; Bera et al., 2019; Deshmukh et al., 2020; Shasidhar et al., 2020; Jadhav et al., 2021). Realizing the importance, we developed and validated four KASP markers (snpAH00031 and snpAH00033 from 8.9-Mb region and snpAH00037 and snpAH00038 from 1.64-Mb region of chromosome B06) for seed weight, targeting potential candidate genes to deploy marker-assisted selection. Therefore, one SNP from each associated region can be used together for selection of large-seeded plants. More importantly, the snpAH0031 should be given high priority, as it is present in the genic region of *Ulp-proteases* (SUMO Proteins), which is known for involvement in the regulation of seed size in groundnut. The four KASP markers successfully distinguished between small- and large-seeded groundnut genotypes. These KASP markers also showed heterozygous calls for F_1 populations derived from the crosses between small- and large-seeded groundnut genotypes, which indicated the plausible use of the developed KASP markers for F_1 population confirmation and identification of heterozygotes in segregating breeding population during marker-assisted selection.

5 | CONCLUSION

The present study demonstrated next-generation sequencing-based trait mapping approach as one of the most efficient and fast-track techniques leading to genomic regions and discovery of putative candidate genes and SNPs associated with 100SW in groundnut. It is reasonable to target the genes from the B genome, such as *Ulp proteases* and *BIG SEED* locus candidate genes, for introgression of bold-seeded trait in groundnut. However, more evidence is required to functionally validate these genes; plausibly silencing of *BIG SEED* locus using RNAi would be best to validate the function of *BIG SEED* locus. To characterize the impact of candidate gene *Ulp protease* or *SUMO proteins* identified in the present study, overexpression studies as well as use of CRISPR/Cas9 may provide more insights on precise function. More importantly, the validated KASP markers are available through this research to deploy in routine breeding programs for breeding bold-seeded varieties to meet the preferences of industry and consumers.

DATA AVAILABILITY STATEMENT
Sequencing data generated in this study have been deposited at National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database with BioProject ID: PRJNA752462.

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

Sunil S. Gangurde: Data curation; Formal analysis; Software; Methodology; Writing – review & editing. Aamir W. Khan: Data curation; Formal analysis; Software. Pasupuleti Janila: Investigation; Methodology; Resources. Murali T. Variath: Resources. Surendra S. Manohar: Resources. Prashant Singam: Data curation; Investigation; Methodology. Annapurna Chitikineni: Investigation; Methodology; Resources. Rajeev K. Varshney: Conceptualization; Funding acquisition; Resources; Writing – review & editing. Manish K. Pandey: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Writing – review & editing.

CONFLICT OF INTEREST

The authors declare there is no conflict of interest.

ORCID

Sunil S. Gangurde https://orcid.org/0000-0001-7119-8649
Rajeev K. Varshney https://orcid.org/0000-0002-4562-9131
Manish K. Pandey https://orcid.org/0000-0002-4101-6530

REFERENCES

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., Matsumura, H., Yoshida, K., Mitsuoka, C., Tamiru, M., Innan, H., Cano, L., Kamoun, S., & Terauchi, R. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature Biotechnology*, *30*, 174–178. https://doi.org/10.1038/nbt.2095

Augustine, R. C., & Vierstra, R. D. (2018). SUMOylation: Re-wiring the plant nucleus during stress and development. *Current Opinion in Plant Biology*, *45*, 143–54. https://doi.org/10.1016/j.pbi.2018.06.006

Bera, S. K., Kamdar, J. H., Kasundra, S. V., Dash, P., Maurya, A. K., Jasani, M. D., Chandrashekar, A. B., Manivannan, N., Vasanthi, R. P., Dobariya, K. L., & Pandey, M. K. (2018). Improving oil quality by altering levels of fatty acids through marker-assisted selection of *ahfad2* alleles in peanut (*Arachis hypogaea* L.). *Euphytica*, *214*, 162. https://doi.org/10.1007/s10681-018-2241-0

Bera, S. K., Kamdar, J. H., Kasundra, S. V., Patel, S. V., Jasani, M. D., Maurya, A. K., Dash, P., Chandrashekar, A. B., Rani, K., Manivannan, N., & Janila, P. (2019). Steady expression of high oleic acid in peanut bred by marker-assisted backcrossing for fatty acid desaturase mutant alleles and its effect on seed germination along with other seedling traits. *PLoS One*, *14*, e0226252. https://doi.org/10.1371/journal.pone.0226252

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K., Liu, X., Gao, D., Clevenger, J., Dash, S., Ren, L., Moretzsohn, M. C., Shirasawa, K., Huang, W., Vidigal, B., Abernathy, B., Chu, Y., Niederhuth, C. E., Umale, P., & Ozias-Akins, P. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics*, *48*, 438–446. https://doi.org/10.1038/ng.3517

Bertioli, D. J., Jenkins, J., Clevenger, J., Gao, D., Dudchenko, O., Seijo, G., Leal-Bertioli, S., Ren, L., Farmer, A., Pandey, M. K., Samoluk, S., Abernathy, B., Agarwal, G., Ballen, C., Cameron, C., Campbell, L., Chavarro, C., Chitikineni, A., Chu, Y., & Schmutz, J. (2019). The genome sequence of peanut (*Arachis hypogaea*), a segmental allotetraploid. *Nature Genetics*, *51*, 877–884. https://doi.org/10.1038/s41588-019-0405-z

Bohar, R., Chitkineni, A., & Varshney, R. K. (2020). Genetic molecular markers to accelerate genetic gains in crops. *Biotechniques*, *69*, 158–160. https://doi.org/10.2144/btn-2020-0066

Chen, X., Li, H., Pandey, M. K., Yang, Q., Wang, X., Garg, V., Li, H., Chi, X., Doddamani, D., Hong, Y., Upadhyaya, H., Guo, H., Khan, A. W., Zhu, F., Zhang, X., Pan, L., Pierce, G. J., Zhou, G., Krishnamohan, K. A., & Yu, S. (2016). Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis, and allergens. *Proceedings of the National Academy of Sciences*, *113*, 6785–6790. https://doi.org/10.1073/pnas.1600899113

Chen, X., Lu, Q., Liu, H., Zhang, J., Hong, Y., Lan, H., Li, H., Wang, J., Liu, H., Li, S., Pandey, M. K., Zhang, Z., Zhou, G., Yu, J., Zhang, G., Yuan, J., Li, X., Wen, S., Meng, F., & Liang, X. (2019b). Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. *Molecular Plant*, *12*, 920–934. https://doi.org/10.1016/j.molp.2019.03.005

Chen, Y., Wang, Z., Ren, X., Huang, L., Guo, J., Zhao, J., Zhou, X., Yan, L., Luo, H., Liu, N., & Chen, W. (2019a). Identification of major QTL for seed number per pod on chromosome A05 of tetraploid peanut (*Arachis hypogaea* L.). *The Crop Journal*, *7*, 238–248. https://doi.org/10.1016/j.cj.2018.09.002

Cho, J. (2018). Transposon-derived non-coding RNAs and their function in plants. *Frontiers in Plant Sciences*, *9*, 600. https://doi.org/10.3389/fpls.2018.00600

Chu, Y., Chee, P., Isleib, T. G., Holbrook, C. C., & Ozias-Akins, P. (2020). Major seed size QTL on chromosome A05 of peanut (*Arachis hypogaea*) is conserved in the US mini core germplasm collection. *Molecular Breeding*, *40*, 6. https://doi.org/10.1007/s11032-019-1082-4

Chu, Y., Wu, C. L., Holbrook, C. C., Tillman, B. L., Person, G., & Ozias-Akins, P. (2011). Marker-assisted selection to pyramid nematode resistance and the high oleic trait in peanut. *The Plant Genome*, *4*. https://doi.org/10.3835/plantgenome2011.01.0001

Cingolani, P., Platts, A., Wangle, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly*, *6*, 80–92. https://doi.org/10.4161/fly.19695

Clark, S. E., Williams, R. W., & Meyerowitz, E. M. (1997). The *CLAVATA1* gene encodes a putative receptor kinase that controls shoot and floral meristem size in Arabidopsis. *Cell*, *89*, 575–585. https://doi.org/10.1016/S0092-8674(00)80239-1

Clevenger, J., Chu, Y., Chavarro, C., Botton, S., Culbreath, A., Isleib, T. G., Holbrook, C. C., & Ozias-Akins, P. (2018). Mapping late leaf spot resistance in peanut (*Arachis hypogaea*) using QTL-seq reveals

markers for marker-assisted selection. *Frontiers in Plant Sciences*, 9, 83. https://doi.org/10.3389/fpls.2018.00083

Clevenger, J., Chu, Y., Scheffler, B., & Ozias-Akins, P. (2016). A developmental transcriptome map for allotetraploid *Arachis hypogaea*. *Frontiers in Plant Sciences*, 7, 1446. https://doi.org/10.3389/fpls.2016.01446

Das, S., Upadhyaya, H. D., Bajaj, D., Kujur, A., Badoni, S., Kumar, V., Tripathi, S., Gowda, C. L., Sharma, S., Singh, S., Tyagi, A. K., & Parida, S. K. (2015). Deploying QTL-seq for rapid delineation of a potential candidate gene underlying major trait-associated QTL in chickpea. *DNA Research*, 22, 193–203. https://doi.org/10.1093/dnares/dsv004

de Godoy, I. J., & Norden, A. J. (1981). Shell and seed size relationships in peanuts. *Peanut Science*, 8, 21–24. https://doi.org/10.3146/i0095-3679-8-1-6

Deshmukh, D., Marathi, B., Sudini, H. K., Variath, M. T., Chaudhari, S., Manohar, S. S., Rani, C. V. D., Pandey, M. K., & Pasupuleti, J. (2020). Combining high oleic acid trait and resistance to late leaf spot and rust diseases in groundnut (*Arachis hypogaea* L.). *Frontiers in Genetics*, 11, 514. https://doi.org/10.3389/fgene.2020.00514

Gangurde, S. S., Wang, H., Yaduru, S., Pandey, M. K., Fountain, J. C., Chu, Y., Isleib, T., Holbrook, C. C., Xavier, A., Culbreath, A. K., Ozias-Akins, P., Varshney, R. K., & Guo, B. (2020). Nested-association mapping (NAM) based genetic dissection uncovers candidate genes for seed and pod weights in peanut (*Arachis hypogaea*). *Plant Biotechnology Journal*, 18, 1457–1471. https://doi.org/10.1111/pbi.13311

Gareau, J. R., & Lima, C. D. (2010). The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nature Reviews Molecular Cell Biology*, 11, 861–871. https://doi.org/10.1038/nrm3011

Ge, L., Yu, J., Wang, H., Luth, D., Bai, G., Wang, K., & Chen, R. (2016). Increasing seed size and quality by manipulating *BIG SEEDS1* in legume species. *Proceedings of the National Academy of Sciences*, 113, 12414–12419. https://doi.org/10.1073/pnas.1611763113

He, C., Holme, J., & Anthony, J. (2014). SNP genotyping: the KASP assay. *Crop Breeding*, 1145, 75–86. https://doi.org/10.1007/978-1-4939-0446-4_7

Huang, L., He, H., Chen, W., Ren, X., Chen, Y., Zhou, X., Xia, Y., Wang, X., Jiang, X., Liao, B., & Jiang, H. (2015). Quantitative trait locus analysis of agronomic and quality-related traits in cultivated peanut (*Arachis hypogaea*). *Theoretical and Applied Genetics*, 128, 1103–1115. https://doi.org/10.1007/s00122-015-2493-1

Huang, Y., Tao, Z. S., Liu, Q., Wang, X. F., Yu, J. Y., Liu, G. H., & Wang, H. Z. (2014). *BnEPFL6*, an EPIDERMAL PATTERNING FACTOR-LIKE (EPFL) secreted peptide gene, is required for filament elongation in *Brassica napus*. *Plant Molecular Biology*, 85, 505–517. https://doi.org/10.1007/s11103-014-0200-2

Jadhav, M. P., Patil, M. D., Hampannavar, M., Dattatreya, P., Shirasawa, K., Pasupuleti, J., Pandey, M. K., Varshney, R. K., & Bhat, R. S. (2021). Enhancing oleic acid content in two commercially released peanut varieties through marker-assisted backcross breeding. *Crop Science*, 61, 2435–2443. https://doi.org/10.1002/csc2.20512

Janiak, A., Kwaśniewski, M., & Szarejko, I. (2016). Gene expression regulation in roots under drought. *Journal of Experimental Botany*, 67, 1003–1014. https://doi.org/10.1093/jxb/erv512

Janila, P., Pandey, M. K., Shasidhar, Y., Variath, M. T., Sriswathi, M., Khera, P., Manohar, S. S., Nagesh, P., Vishwakarma, M. K., Mishra, G. P., & Radhakrishnan, T. (2016). Molecular breeding for introgression of fatty acid desaturase mutant alleles (*ahFAD2A* and *ahFAD2B*) enhances oil quality in high and low oil containing peanut genotypes. *Plant Science*, 242, 203–213. https://doi.org/10.1016/j.plantsci.2015.08.013

Jinn, T. L., Stone, J. M., & Walker, J. C. (2000). HAESA, an *Arabidopsis* leucine-rich repeat receptor kinase, controls floral organ abscission. *Genes & Development*, 14, 108–117.

Katta, M. A. V. S. K., Khan, A. W., Doddamani, D., Thudi, M., & Varshney, R. K. (2015). NGS-QCbox and raspberry for parallel, automated and rapid quality control analysis of large-scale next generation sequencing (Illumina) data. *PLoS One*, 10, e0139868. https://doi.org/10.1371/journal.pone.0139868

Kolekar, R. M., Sukruth, M., Shirasawa, K., Nadaf, H. L., Motagi, B. N., Lingaraju, S., Patil, P. V., & Bhat, R. S. (2017). Marker-assisted backcrossing to develop foliar disease-resistant genotypes in TMV 2 variety of peanut (*Arachis hypogaea* L.). *Plant Breeding*, 136, 948–953. https://doi.org/10.1111/pbr.12549

Kosugi, S., Natsume, S., Yoshida, K., MacLean, D., Cano, L., Kamoun, S., & Terauchi, R. (2013). Coval: Improving alignment quality and variant calling accuracy for next-generation sequencing data. *PLoS One*, 8, e75402. https://doi.org/10.1371/journal.pone.0075402

Kumar, R., Pasupuleti, J., Vishwakarma, M. K., Khan, A., Manohar, S. S., Gangurde, S. S., Murali, T. V., Shasidhar, Y., Pandey, M. K., & Varshney, R. K. (2020). Whole genome re-sequencing-based QTL-seq identified candidate genes and molecular markers for fresh seed dormancy in groundnut. *Plant Biotechnology Journal*, 18, 992–1003. https://doi.org/10.1111/pbi.13266

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, J., & Chory, J. (1997). A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. *Cell*, 90, 929–938. https://doi.org/10.1016/S0092-8674(00)80357-8

Liu, Y., Zhu, J., Sun, S., Cui, F., Han, Y., Peng, Z., Zhang, X., Wan, S., & Li, G. (2019). Defining the function of SUMO system in pod development and abiotic stresses in peanut. *BMC Plant Biology*, 19, 593. https://doi.org/10.1186/s12870-019-2136-9

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., Sun, J., Zhang, Z., Weng, Y., & Huang, S. (2014). QTL-seq identifies an early flowering QTL located near *Flowering Locus T* in cucumber. *Theoretical and Applied Genetics*, 127, 1491–1499. https://doi.org/10.1007/s00122-014-2313-z

Lu, Q., Liu, H., Hong, Y., Li, H., Liu, H., Li, X., Wen, S., Zhou, G., Li, S., Chen, X., & Liang, X. (2018). Consensus map integration and QTL meta-analysis narrowed a locus for yield traits to 0.7 cM and refined a region for late leaf spot resistance traits to 0.38 cM on linkage group A05 in peanut (*Arachis hypogaea* L.). *BMC Genomics*, 19, 887. https://doi.org/10.1186/s12864-018-5288-3

Luo, H., Guo, J., Ren, X., Chen, W., Huang, L., Zhou, X., Chen, Y., Liu, N., Xiong, F., Lei, Y., Liao, B., & Jinag, H. (2018). Chromosomes A07 and A05 associated with stable and major QTLs for pod weight and size in cultivated peanut (*Arachis hypogaea* L.). *Theoretical and Applied Genetics*, 131, 267–282. https://doi.org/10.1007/s00122-017-3000-7

Luo, H., Pandey, M. K., Khan, A. W., Guo, J., Wu, B., Cai, Y., Huang, L., Zhou, X., Chen, Y., Chen, W., Liu, N., Lei, Y., Liao, B., Varshney, R. K., & Jiang, H. (2019b). Discovery of genomic regions and candidate genes controlling shelling percentage using QTL-seq approach in

cultivated peanut (*Arachis hypogaea* L.). *Plant Biotechnology Journal*, *17*, 1248–1260. https://doi.org/10.1111/pbi.13050

Luo, H., Pandey, M. K., Khan, A. W., Wu, B., Guo, J., Ren, X., Zhou, X., Chen, Y., Chen, W., Huang, L., Liu, N., Lei, Y., Liao, B., Varshney, R. K., & Jiang, H. (2019a). Next-generation sequencing identified genomic region and diagnostic markers for resistance to bacterial wilt on chromosome B02 in peanut (*Arachis hypogaea* L.). *Plant Biotechnology Journal*, *17*, 2356–2369. https://doi.org/10.1111/pbi.13153

Luo, H., Xu, Z., Li, Z., Li, X., Lv, J., Ren, X., Huang, L., Zhou, X., Chen, Y., Yu, J., Chen, W., Lei, Y., Liao, B., & Jiang, H. (2017). Development of SSR markers and identification of major quantitative trait loci controlling shelling percentage in cultivated peanut (*Arachis hypogaea* L.). *Theoretical and Applied Genetics*, *130*, 1635–1648. https://doi.org/10.1007/s00122-017-2915-3

Michelmore, R. W., Paran, I., & Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences*, *88*, 9828–9832. https://doi.org/10.1073/pnas.88.21.9828

Mondal, S., Phadke, R. R., & Badigannavar, A. M. (2015). Genetic variability for total phenolics, flavonoids and antioxidant activity of testaless seeds of a peanut recombinant inbred line population and identification of their controlling QTLs. *Euphytica*, *204*, 311–321. https://doi.org/10.1007/s10681-014-1324-9

Nakashima, K., & Yamaguchi-Shinozaki, K. (2013). ABA signaling in stress-response and seed development. *Plant Cell Reports*, *32*, 959–970. https://doi.org/10.1007/s00299-013-1418-1

Noman, A., Liu, Z., Aqeel, M., Zainab, M., Khan, M. I., Hussain, A., Ashraf, M. F., Li, X., Weng, Y., & He, S. (2017). Basic leucine zipper domain transcription factors: the vanguards in plant immunity. *Biotechnology Letters*, *39*, 1779–1791. https://doi.org/10.1007/s10529-017-2431-1

Pandey, M. K., Khan, A. W., Singh, V. K., Vishwakarma, M. K., Shasidhar, Y., Kumar, V., Garg, V., Bhat, R. S., Chitikineni, A., Janila, P., Guo, B., & Varshney, R. K. (2017). QTL-seq approach identified genomic regions and diagnostic markers for rust and late leaf spot resistance in groundnut (*Arachis hypogaea* L.). *Plant Biotechnology Journal*, *15*, 927–941. https://doi.org/10.1111/pbi.12686

Pandey, M. K., Monyo, E., Ozias-Akins, P., Liang, X., Guimarães, P., Nigam, S. N., Upadhyaya, H. D., Janila, P., Zhang, X., Guo, B., & Cook, D. R. (2012). Advances in Arachis genomics for peanut improvement. *Biotechnology Advances*, *30*, 639–651. https://doi.org/10.1016/j.biotechadv.2011.11.001

Pandey, M. K., Pandey, A. K., Kumar, R., Nwosu, C. V., Guo, B., Wright, G. C., Bhat, R. S., Chen, X., Bera, S. K., Yuan, M., & Jiang, H. (2020). Translational genomics for achieving higher genetic gains in groundnut. *Theoretical and Applied Genetics*, *133*, 1679–1702. https://doi.org/10.1007/s00122-020-03592-2

Pandey, M. K., Roorkiwal, M., Singh, V. K., Ramalingam, A., Kudapa, H., Thudi, M., Chitikineni, A., Rathore, A., & Varshney, R. K. (2016). Emerging genomic tools for legume breeding: current status and future prospects. *Frontiers in Plant Sciences*, *7*, 45510.

Parmar, S., Deshmukh, D. B., Kumar, R., Manohar, S. S., Joshi, P., Sharma, V., Chaudhari, S., Variath, M. T., Gangurde, S. S., Bohar, R., & Singam, P. (2021). Single seed-based high-throughput genotyping and rapid generation advancement for accelerated groundnut genetics and breeding research. *Agronomy*, *11*, 1226. https://doi.org/10.3390/agronomy11061226

Shasidhar, Y., Variath, M. T., Vishwakarma, M. K., Manohar, S. S., Gangurde, S. S., Sriswathi, M., Sudini, H. K., Dobariya, K. L., Bera, S. K., Radhakrishnan, T., & Pandey, M. K. (2020). Improvement of three popular Indian groundnut varieties for foliar disease resistance and high oleic acid using SSR markers and SNP array in marker-assisted backcrossing. *The Crop Journal*, *8*, 1–15. https://doi.org/10.1016/j.cj.2019.07.001

Singh, V. K., Khan, A. W., Saxena, R. K., Kumar, V., Kale, S. M., Sinha, P., Chitikineni, A., Pazhamala, L. T., Garg, V., Sharma, M., Sameer Kumar, C. V., Parupalli, S., Vechapalu, S., Patil, S., Muniswamy, S., Ghanta, A., Yamini, K. N., Dharmaraj, P. S., & Varshney, R. K. (2016). Next-generation sequencing for identification of candidate genes for *Fusarium* wilt and sterility mosaic disease in pigeonpea (*Cajanus cajan*). *Plant Biotechnology Journal*, *14*, 1183–1194. https://doi.org/10.1111/pbi.12470

Sinha, P., Bajaj, P., Pazhamala, L. T., Nayak, S. N., Pandey, M. K., Chitikineni, A., Huai, D., Khan, A. W., Desai, A., Jiang, H., Zhuang, W., Guo, B., Liao, B., & Varshney, R. (2020). *Arachis hypogea* gene expression atlas (AhGEA) for fastigiata subspecies of cultivated groundnut to accelerate functional and translational genomics applications. *Plant Biotechnology Journal*, *18*, 2187–2200. https://doi.org/10.1111/pbi.13374

Smaczniak, C., Immink, R. G. H., Angenent, G. C., & Kaufmann, K. (2012). Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development*, *139*, 3081–3098. https://doi.org/10.1242/dev.074674

Sun, Q., Qu, J., Yu, Y., Yang, Z., Wei, S., Wu, Y., Yang, J., & Peng, Z. (2019). *TaEPFL1*, an EPIDERMAL PATTERNING FACTOR-LIKE (EPFL) secreted peptide gene, is required for stamen development in wheat. *Genetica*, *147*, 121–130. https://doi.org/10.1007/s10709-019-00061-7

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L. M., Kamoun, S., & Terauchi, R. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal*, *74*, 174–183. https://doi.org/10.1111/tpj.12105

Taylor, I., Wang, Y., Seitz, K., Baer, J., Bennewitz, S., Mooney, B. P., & Walker, J. C. (2016). Analysis of phosphorylation of the receptor-like protein kinase HAESA during Arabidopsis floral abscission. *PLoS One*, *11*, e0147203. https://doi.org/10.1371/journal.pone.0147203

Uhlken, C., Horvath, B., Stadler, R., Sauer, N., & Weingartner, M. (2014). MAIN-LIKE1 is a crucial factor for correct cell division and differentiation in *Arabidopsis thaliana*. *The Plant Journal*, *78*, 107–120. https://doi.org/10.1111/tpj.12455

van den Burg, H. A., Kini, R. K., Schuurink, R. C., & Takken, F. L. (2010). Arabidopsis small ubiquitin-like modifier paralogs have distinct functions in development and defense. *The Plant Cell*, *22*, 1998–2016. https://doi.org/10.1105/tpc.109.070961

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., & Sorrells, M. E. (2021). Designing future crops: genomics-assisted breeding comes of age. *Trends in Plant Sciences*, *26*, 631–649. https://doi.org/10.1016/j.tplants.2021.03.010

Varshney, R. K., Mohan, S. M., Gaur, P. M., Gangarao, N. V. P. R., Pandey, M. K., Bohra, A., Sawargaonkar, S. L., Chitikineni, A., Kimurto, P. K., Janila, P., & Saxena, K. B. (2013). Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnology Advances*, *31*, 1120–1134. https://doi.org/10.1016/j.biotechadv.2013.01.001

Varshney, R. K., Pandey, M. K., Bohra, A., Singh, V. K., Thudi, M., & Saxena, R. K. (2019). Toward sequence-based breeding in legumes in the postgenome sequencing era. *Theoretical and Applied Genetics*, *132*, 797–816. https://doi.org/10.1007/s00122-018-3252-x

Varshney, R. K., Pandey, M. K., Janila, P., Nigam, S. N., Sudini, H., Gowda, M. V. C., Sriswathi, M., Radhakrishnan, T., Manohar, S. S., & Patne, N. (2014). Marker-assisted introgression of a QTL region to improve rust resistance in three elite and popular varieties of peanut (*Arachis hypogaea* L.). *Theoretical and Applied Genetics*, *127*, 1771–1781. https://doi.org/10.1007/s00122-014-2338-3

Venuprasad, R., Aruna, R., & Nigam, S. N. (2011). Inheritance of traits associated with seed size in groundnut (*Arachis hypogaea* L.). *Euphytica*, *181*, 169–177. https://doi.org/10.1007/s10681-011-0390-5

Wang, Z., Huai, D., Zhang, Z., Cheng, K., Kang, Y., Wan, L., Yan, L., Jiang, H., Lei, Y., & Liao, B. (2018). Development of a high-density genetic map based on specific length amplified fragment sequencing and its application in quantitative trait loci analysis for yield-related traits in cultivated peanut. *Frontiers in Plant Science*, *9*, 827. https://doi.org/10.3389/fpls.2018.00827

Wils, C. R., & Kaufmann, K. (2017). Gene-regulatory networks controlling inflorescence and flower development in *Arabidopsis thaliana*. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1860*, 95–105. https://doi.org/10.1016/j.bbagrm.2016.07.014

Xiao, W., Brown, R. C., Lemmon, B. E., Harada, J. J., Goldberg, R. B., & Fischer, R. L. (2006). Regulation of seed size by hypomethylation of maternal and paternal genomes. *Plant Physiology*, *142*, 1160–1168. https://doi.org/10.1104/pp.106.088849

Zhang, S., Hu, X., Miao, H., Chu, Y., Cui, F., Yang, W., Wang, C., Shen, Y., Xu, T., Zhao, L., Zhang, J., & Chen, J. (2019). QTL identification for seed weight and size based on a high-density SLAF-seq genetic map in peanut (*Arachis hypogaea* L.). *BMC Plant Biology*, *19*, 537. https://doi.org/10.1186/s12870-019-2164-5

Zhao, Y., Ma, J., Li, M., Deng, L., Li, G., Xia, H., Zhao, S., Hou, L., Li, P., Ma, C., & Yuan, M. (2020). Whole-genome resequencing-based QTL-seq identified *AhTc1* gene encoding a R2R3-MYB transcription factor controlling peanut purple testa colour. *Plant Biotechnology Journal*, *18*, 96–105. https://doi.org/10.1111/pbi.13175

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., Chang, W. C., Zhang, L., Zhang, X., Tang, R., Garg, V., Wang, X., Tang, H., Chow, C. N., Wang, J., Deng, Y., Wang, D., Khan, A. W., Yang, Q., & Varshney, R. K. (2019). The *Arachis hypogaea* genome elucidates legume karyotypes, polyploid evolution and crop domestication. *Nature Genetics*, *51*, 865–876. https://doi.org/10.1038/s41588-019-0402-2

SUPPORTING INFORMATION

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