Molecular Genetics and Genomics (2019) 294:57–68
https://doi.org/10.1007/s00438-018-1484-8

ORIGINAL ARTICLE

Development of sequence-based markers for seed protein content in pigeonpea

Jimmy Obala[1,2] · Rachit K. Saxena[1] · Vikas K. Singh[1] · C. V. Sameer Kumar[1] · K. B. Saxena[1] · Pangirayi Tongoona[2] · Julia Sibiya[2] · Rajeev K. Varshney[1]

Received: 1 February 2018 / Accepted: 22 August 2018 / Published online: 1 September 2018

Abstract

Pigeonpea is an important source of dietary protein to over a billion people globally, but genetic enhancement of seed protein content (SPC) in the crop has received limited attention for a long time. Use of genomics-assisted breeding would facilitate accelerating genetic gain for SPC. However, neither genetic markers nor genes associated with this important trait have been identified in this crop. Therefore, the present study exploited whole genome re-sequencing (WGRS) data of four pigeonpea genotypes (~ 12X coverage) to identify sequence-based markers and associated candidate genes for SPC. By combining a common variant filtering strategy on available WGRS data with knowledge of gene functions in relation to SPC, 108 sequence variants from 57 genes were identified. These genes were assigned to 19 GO molecular function categories with 56% belonging to only two categories. Furthermore, Sanger sequencing confirmed presence of 75.4% of the variants in 37 genes. Out of 30 sequence variants converted into CAPS/dCAPS markers, 17 showed high level of polymorphism between low and high SPC genotypes. Assay of 16 of the polymorphic CAPS/dCAPS markers on an F_2 population of the cross ICP 5529 (high SPC) × ICP 11605 (low SPC), resulted in four of the CAPS/dCAPS markers significantly ($P < 0.05$) co-segregated with SPC. In summary, four markers derived from mutations in four genes will be useful for enhancing/regulating SPC in pigeonpea crop improvement programs.

Keywords Seed protein content · *Cajanus cajan* · Whole-genome resequencing · Next generation sequencing · Sequence variants · Common variant analysis

Introduction

Pigeonpea (*Cajanus cajan* L.) is one of the important legume crops in sub-tropical and tropical regions of the world. It is an often cross pollinated species with 11 pairs of chromosomes ($2n = 2x = 22$) and a genome size of 833.07 Mbp (Varshney et al. 2012). It is the only cultivated food legume of the tribe Phaseoleae, sub-tribe Cajaninae, family Fabaceae (*Leguminosae*) and sub-family Papilionoideae (Greilhuber and Obermayer 1998). Global area under pigeonpea cultivation continues to increase annually (Akibode and Maredia 2011) standing at 5.6 million ha in the year 2016 with a production of ~ 4.0 million tons (FAO 2016). Pigeonpea has diverse uses including food, feed, fodder, building material and fuel wood, in addition to its contribution to biological nitrogen fixation (Rao et al. 2010). It is also a cash crop that supports the livelihoods of millions of resources-poor farmers in Asia and Africa (Mula and Saxena 2010). As a source of food it provides dietary protein to more than a billion people globally (Krishnan et al. 2017).

Considering the importance of total seed protein content (SPC) in global food and nutritional security, there is a need to produce more protein per unit area to meet the present and future dietary protein demands (Saxena and Sawargaonkar 2015). However, breeding objectives in pigeonpea have, for

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-018-1484-8) contains supplementary material, which is available to authorized users.

✉ Rajeev K. Varshney
r.k.varshney@cgiar.org

[1] International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502324, India

[2] University of KwaZulu-Natal, African Center for Crop Improvement, Scottsville, Pietermaritzburg 3209, South Africa

a long time, almost entirely focused on increasing yield and crop adaptability (Odeny 2007; Mligo and Craufurd 2005; Upadhyaya et al. 2007a). Very little or no attention has been given to the nutritional quality of the pigeonpea seed in terms of genetic enhancement, yet it has been reported that adequate genetic variability for SPC exists within the cultivated genepool that can be harnessed for trait improvement (Upadhyaya et al. 2007b). Availability of genomic resources in pigeonpea such as a reference genome (Varshney et al. 2012) and whole genome re-sequencing (WGRS) data (Kumar et al. 2016; Varshney et al. 2017) provides an opportunity to improve productivity and quality traits in the crop through modern/molecular breeding approaches such as genomics-assisted breeding (GAB) for accelerated genetic gains. However, the first step in GAB is the identification of molecular markers or candidate genes associated with the trait(s) of interest (Feng et al. 2014), which in turn provides the breeder with a critical tool to modify those traits (Janninks 2001).

The recent developments in next generation sequencing (NGS) technologies provide rapid and cost-effective methods to identify sequence variants and candidate genes underlying qualitative and quantitative traits (Silva et al. 2012; Xu et al. 2014). In the presence of a reference genome sequence, WGRS data of one or a few individuals can be used to identify variants associated with phenotype of interest as demonstrated in human (Rios et al. 2010; Roach et al. 2010; Sobreira et al. 2010) and in crop plants such as rice (Lim et al. 2014; Silva et al. 2012), maize (Xu et al. 2014) and recently in pigeonpea (Varshney et al. 2017). The sequence variant can then be used as markers in breeding programs aimed at improving the trait(s) of interest (Cabezas et al. 2015). To identify sequence variants and associated candidate genes using WGRS data, common variant (CV) and clustering analyses have been proposed and used (Silva et al. 2012; Xu et al. 2014). However, Silva et al. (2012) did not find any significant performance difference between the two analysis methods while Xu et al. (2014) found the CV analysis to be more efficient than the clustering approach. A major assumption in filtering variants from NGS data for the purpose of selection of candidate gene-based sequence variants is that the causative variant likely leads to change on the protein level, so changes such as nonsense, missense, splicing and frameshift variants are prioritized (Coonrod et al. 2013). Further prioritization of sequence variants may be based on information on gene function in relation to the phenotype (Gilissen et al. 2012).

NGS technologies as used for generating WGRS data of the parental lines produces short reads, which may result in to misalignments to the reference genome (Church et al. 2011). Thus, validation of sequence variants identified from NGS-based approaches must be done to determine the analytical sensitivity and specificity by comparing NGS test results to those obtained from an independently validated method such as Sanger sequencing. Sanger sequencing is less prone to sequencing errors than NGS (Machado et al. 2011) and has preferentially been used to validate the presence of sequence variants such as single nucleotide polymorphisms (SNPs) by sequencing the fragments containing the candidate variants. The final testing of the role of candidate gene mutations can be carried out by conventional co-segregation analysis in structured population such as F_2, or by sequence variant-phenotype associations in germplasm collections or natural populations, or in functional experiments (Pflieger et al. 2001; Grattapaglia 2008; Sobreira et al. 2010; Gilissen et al. 2012).

In view of above, the present study has been undertaken to identify sequence variants and candidate genes for SPC in pigeonpea by: (1) identifying sequence variants from WGRS data that play role in seed storage protein accumulation, (2) identifying corresponding candidate genes with sequence variants, (3) validating presence of the sequence variants in candidate genes through Sanger sequencing, and (4) determining the association of the sequence variants/candidate genes with SPC in segregating mapping population.

Materials and methods

Plant material

Five pigeonpea genotypes (HPL 24, ICP 5529, ICP 11605, ICPL 87119 and UQ 50) from cultivated (*C. cajan*) pool and one genotype (ICPW 90) from wild relative species (*C. scarabaoiedes*) were investigated (Table 1). The WGRS data of HPL 24, ICP 5529, ICP 11605 and draft genome of ICPL 87119 were used for the identification of putative candidate sequence variants and genes. HPL 24 and ICP 11605 were used to validate presence of sequence variants through Sanger sequencing. UQ 50 and ICPW 90 were included as independent genetic background for checking amplification of the primers. They also facilitated comparison of read alignments across multiple individuals, which have the potential to filter out sequence variants that are an artifact of inaccurate read alignments (Bansal et al. 2010). To assess the co-segregation of the identified sequence variants with SPC, two parental lines (ICP 5529 and ICP 11605) with contrasting SPC values, and their segregating F_2 population were used.

Seed protein content phenotyping

Five pigeonpea genotypes as well as one wild relative and 188 F_2 progenies of the cross between ICP 5529 × ICP 11605 grown under field conditions were used in the present study. Pigeonpea genotypes and wild relative genotype were sown

Table 1 Pigeonpea lines and segregating population used for the identification and validation of candidate genes for seed protein content

Pedigree	Description
HPL 24	Breeding line with high SPC. WGRS data available Kumar et al. (2016)
ICPW 90	*C. scarabaeoides* (a wild relative of *C. cajanus*). Presumably previously used to develop high SPC breeding lines
UQ 50	Breeding line with moderate SPC. WGRS data available Kumar et al. (2016)
ICP 5529	Landrace with high SPC. WGRS data available Kumar et al. (2016)
ICP 11605	Germplasm line with low SPC. WGRS data available Kumar et al. (2016)
ICPL 87119	Germplasm line with low SPC. Reference genome available Varshney et al. (2012)
ICP 5529 × ICP 11605	F_2 mapping population segregating for SPC

SPC seed protein content, *WGRS* whole genome resequencing

in single rows each while the F_2s were in 19 rows. Each row was 4 m long with row to row and plant to plant spacing of 75 cm and 30 cm, respectively. To avoid insect pollinators, the materials were grown under nylon nets. Agronomic practices included application of 100 kg/ha of diammonium phosphate as basal fertilizer without any top dressing, 2 and 4 L/ha of pendimethalin and paraquat dichloride pre-emergence herbicides, respectively, provision of two irrigations, one each at planting and pod filling stages, and two weedings one each at early vegetative and podding stages. Pod borers (*Maruca vitrata* Fab. and *Helicoverpa armigera* Hub.) were controlled by spraying with acephate and spinosad insecticides at rates of 1.0 kg/ha and 0.2 L/ha, respectively at 15-day intervals from flowering to podding stages. At maturity individual pods from individual plants were carefully hand-harvested leaving out plants at the beginning and at the end of each row and those at the field borders to avoid border effects. Sun drying was done for 1 week before threshing and another 1 week after threshing to ensure uniform reduction in seed moisture content.

Ten grams of mature dry clean seeds of three plants each per genotype and 188 F_2 plants were analyzed at the Charles Renard Analytical Laboratory at ICRISAT, India. Before grinding, seeds were oven-dried at 60 °C for 48 h. The dried seed samples were ground into powder in a mill with Teflon chambers. The ground samples were again kept in an oven at 60 °C overnight. Samples and appropriate blanks were digested simultaneously in duplicate (i.e. two independent analyses) using tri-acid digestion procedure as described in Upadhyaya et al. (2016). Aliquots were obtained from the digests and used to estimate the total nitrogen (N) using a San++Automated Wet Chemistry Analyzer (Skalar, Breda, The Netherlands). Seed protein content of a sample was estimated by multiplying its N (%) content by factor 6.25.

Sequence variant detection

Existing WGRS data of HPL 24, ICP 5529 and ICP 11605 (Kumar et al. 2016), were cleaned and trimmed to remove poor quality bases using Sickle (Joshi and Fass 2011). The cleaned data were aligned onto version 1.0 of the pigeonpea reference genome (Varshney et al. 2012) using Bowtie 2 version 2.0 (Langmead and Salzberg 2012). Unique hits were retained for further analyses in the Binary Alignment/Map (BAM) (Li et al. 2009) files. The BAM files were processed using the IndelRealigner component of the genome analysis toolkit (GATK) version 4.0 suite (DePristo et al. 2011) and sequence variants were detected using the UnifiedGenotyper of GATK version 4.0 (DePristo et al. 2011). A position in a genotype was reported as a sequence variant if the Phred quality score for the base was ≥ 30 and if the number of sequence reads aligned in each of the lines against the reference genome was ≥ 5. Only one sequence variant was retained and reported if two or more sequence variants were present in a 5-bp window. The sequence variants obtained in last step were then subjected to the common variant analysis (CV) (Silva et al. 2012) to identify candidate variants and genes.

Common variant (CV) analysis

The CV analysis was performed as follows: sequence variations within high and within low SPC genotypes were compared. Sequence variants for which the allelic calls in HPL 24 was the same as in ICP 5529 but contrasting with that in ICP 11605 and ICPL 87119 (in which the calls in ICP 11605 was the same as that in ICPL 87119) were retained for further analysis. The sequence variants were subjected to their effects using snpEff program (Cingolani et al. 2012). Annotation of the genes containing sequence variants was carried out using BLASTX against SWISS-PROT and TREMBL databases. Corresponding gene ontologies were extracted using UniprotKB database (The UniProt Consortium 2008). Where UniprotKB database returns an uncharacterized protein, the *C. cajan* gene ID was submitted to LegumeIP v2.0 (Li et al. 2012) to search for gene/protein function category within the integrated legume database. Potential causal variants that result in non-synonymous changes in the coding DNA sequence (CDS) regions were identified by filtering out intergenic, intronic and synonymous variants. Heterozygous calls were also removed from the list of sequence variants. A final selection of the candidates was based on information on gene function in relation to SPC. This was achieved by using the protein name associated with *C. cajan* gene together with either 'seed storage protein' or 'seed protein content' or 'grain protein content' as search terms to obtain original publication containing gene information. A gene was considered as a candidate only if there was experimental evidence from the publication that it plays a role in storage protein metabolism or reported as falling within confidence intervals of a QTL for seed or grain protein content.

Sanger sequencing

Genomic DNA (gDNA) was isolated from young trifoliate leaves using CTAB method (Mace et al. 2003) and then column purified using NucleoSpin Plant II kit (Macherey–Nagel, Düren, Germany) following the manufacturer's instructions. Sequences of approximately 350 bps flanking either side of the identified sequence variant sites were extracted using the pigeonpea reference genome. Polymerase chain reaction (PCR) primers of length 21–24 bp and T_m of 56–59.5 °C were designed from each 601 bp sequence using BatchPrimer3 v1.0 primer design software tool (You et al. 2008).

PCR was performed for each of the selected variants in a total volume of 30 µL containing 21.9 µL of ddH$_2$O, 10× Taq polymerase buffer, 2.0 µL of 2 mM dNTPs, 10 pmol/µL of each of the forward and reverse primers, 0.06 µL of Taq polymerase and 2.0 µL of 20 ng/µL gDNA. A touchdown PCR (Korbie and Mattick 2008) was used as follows: initial denaturation at 95 °C for 5 min followed by (1) 5 cycles consisting of (1) 94 °C for 15 s, (2) 62 °C for 20 s and (3) 72 °C for 30 s, (2) 35 cycles consisting of (1) 94 °C for 15 s, (2) 54 °C for 30 s and (3) 72 °C for 30 s and a final extension of 72 °C for 20 min. PCR products were run in 3.5% Nusieve agarose gel. Gels were stained with ethidium bromide (0.5 µg/mL) and visualized under UV light in a transilluminator (Bio-Rad, California, USA).

Only PCR products showing single bands across the four genotypes were further processed for Sanger sequencing. PCR cleanup reactions were then performed by mixing 20 µL of PCR products with 1.1 µL of ExoSAP-IT (USB Corporation, Cleveland, Ohio) and incubating the mixture for 45 min at 37 °C followed by 15 min at 80 °C. Ten µL of each of the cleaned PCR products was vacuum dried and end-sequenced using forward and reverse primers at Macrogen Korea (https://dna.macrogen.com/eng/). The two sequences generated by the forward and reverse primers from each genotype were combined into genotype-specific contigs. The genotype-specific contigs from all the four genotypes were compared with the reference sequence of Asha (ICPL 87119) at the originally targeted sequence variant position using DNA Baser (DNA Baser Sequence Assembler v4.23, Heracle BioSoft, http://www.DnaBaser.com).

CAPS and dCAPS primer design, PCR amplification and restriction digestion

Cleaved amplified polymorphic sequence (CAPS) and derived-CAPS (dCAPS) primers were designed by submitting 22–24 bp sequences flanking the sequence variant position for both 'wild-type' and 'mutant-type' alleles (Lee 2012) using online software dCAPS Finder 2.0 (Neff et al. 2002). Because the dCAPS Finder software generates only either a forward or reverse primer sequence in the case of dCAPS, the complementary strand of any chosen dCAPS primer was designed by submitting the 601-bp long reference fragment containing the appropriate sequence variant allele (either wild type or mutant type) to Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) with the default settings (Lestari and Koh 2013). PCR amplification and gel visualisation for the CAPS and dCAPS markers were performed as described above under "Sanger sequencing" section. Restriction digestion was performed in 30 µL reaction volume containing 17 µL of ddH$_2$O, 1.0 µL restriction enzyme (RE), 2.0 µL RE buffer and 10 µL PCR product. The digestion mixture was incubated at 37–50 °C for 2–3 h and held at 0–80 °C for 20 min depending on RE and the manufacturer's instructions.

Integration of CAPS/dCAPS markers in to genetic map and single marker analysis

The CAPS/dCAPS genotyping data generated from 188 F$_2$ plants derived from cross ICP 5529 × ICP 11605 were combined with a GBS-derived single nucleotide polymorphism (SNP) data already available on the same population (ICP 5529 × ICP 11605) (Saxena et al. 2017). To assess co-segregation of the CAPS/dCAPS markers with SPC, single marker regression analysis (SMA) was carried out in Excel 2013 (Microsoft) using the F$_2$ CAPS/dCAPS marker genotypes as independent variables and the F$_2$ phenotypes as dependent variables.

Results

Sequence variations

Sequencing data on genotypes obtained from Kumar et al. (2016) were used for alignment with the draft genome and sequence variant detection. All the detected sequence variants were subjected to CV analysis as mentioned in "Materials and methods" section. As a result, a total of 32,964 sequence variants in 1,417 genes were found between the high (HPL 24, ICP 5529) and low (ICP 11605, ICPL 87119) SPC groups (Table 2; ESM 1). Intergenic regions had the highest proportion of sequence variants (83.4%) followed by sequence variants present in intronic (12.4%) and exonic (3.8%) regions. Within the exonic regions, there were 485 synonymous SNPs (sSNPs), 718 non-synonymous (nsSNPs), 26 stop-gains and one each of stop-loss and start-loss mutations. Other sequence variant types identified in the exons included splice-sites (0.03%), indels (0.03%) and frameshifts (0.07%). Non-synonymous SNPs were more abundant with an average nsSNPs to sSNPs ratio of 1.48, which is close to 1.46 estimated previously (Kumar et al. 2016). The number of genes per chromosome/pseudomolecule ranged from 14 in CcLG05 to 125 in CcLG02 with the unanchored scaffolds containing the highest number (674) of the genes (Table 2). To identify potential causal sequence variants that induce protein coding alterations, the present study focused on non-synonymous sequence variants. The nonsynonymous variants included nsSNPs, stop-gains, splice sites, frame-shifts and indel-mutations in the coding regions.

Candidate genes for seed protein content

A total of 108 nonsynonymous sequence variants in 57 genes were identified in relation to SPC metabolism (ESM 1; ESM 2; ESM 3). The sequence variants present in the 57 pigeonpea genes were spread over all pseudomolecules (CcLGs) with an exception of CcLG05 and several scaffolds (ESM 2). The distribution of selected sequence variants and corresponding genes across chromosomes was not uniform. For example, a maximum of 25 sequence variants in nine genes were found on CcLG01 whereas 1, 3, 5, 17 and 19 sequence variants and 1, 3, 4, 9 and 4 genes were detected on CcLG09, CcLG11, CcLG03, CcLG02 and CcLG07, respectively (ESM 2). A considerable number of sequence variants and genes (14 and 9, respectively) were present in nine unanchored scaffolds. The 57 identified candidate genes could be placed in 19 functional categories based on GO molecular function (Fig. 1). The functional groups which were highly represented in terms of selected genes include aspartic-type endopeptidase (protease), ATP binding/ATPase, DNA binding, iron ion binding, metal iron binding and chitinase activity with 17, 15, four, three, three and two genes, respectively (Fig. 1). The remaining functional categories contained one gene each (Fig. 1).

Validation of candidate sequence variants

Primer pairs were designed to amplify 108 sequence variant-containing fragments from 57 genes. A total of 86 sequence variant-containing gene fragments could be amplified and further processed for Sanger sequencing. Sixty-nine sequence fragments from 42 genes were successfully Sanger-sequenced (no missing genotype data) across the validation panel of two genotypes, namely ICP11605 (with low SPC) and HPL24 (with high SPC) (ESM 4). The ICP 11605 allele would be expected to match with the reference assembly allele of Asha (ICPL 87119) since ICPL 87119 is a low SPC genotype itself while the HPL 24 allele should match to the alternative allele. Accordingly, not all PCR-generated sequence variant-specific alleles for the test genotypes were consistent with those from the WGRS data and the reference genome sequence (ESM 4). By comparing ICP 11605 (low SPC) and ICP 5529 (high SPC) alleles with the reference genome and the WGRS-derived alternative alleles, respectively, the presence of a total of 52 (75.4%) SNPs was confirmed out of the 69 successfully Sanger-sequenced fragments. However, a SNP locus at position 17,486,133 bp on CcLG01 had a different alternative sequence variant allele, i.e. A–C instead of A–T (ESM 4).

Conversion of sequence variants to CAPS/dCAPS assays

A combined set of 61 sequence variants including 52 variants confirmed through Sanger sequencing and nine variants which had poor quality of Sanger sequencing data were converted in to CAPS/dCAPS assays (ESM 5). As a result, 59 sequence variants could be converted in to CAPS/dCAPS assays, and no suitable restriction sites could be found in sequence fragments containing two remaining sequence variants. Of the 59 CAPS/dCAPS only 28 were successfully amplified and digested on six pigeonpea genotypes (HPL 24, ICP 11605, ICPL 87119, ICP 5529, ICP 8863 and ICP 14209) and remaining 31 were either not amplified or failed to digest (ESM 6). Pair-wise analysis of CAPS/dCAPS genotyping data on six pigeonpea genotypes provided the highest number of polymorphic markers between the high/low parental pairs such as HPL 24/ICP 11605 with 17 markers, HPL 24/ICPL87119 (16 markers) and ICP 5529/1CP 11605 (16 markers) (ESM 6). The lowest number of polymorphic markers was between high/high such as in HPL 24/ICP 5529 (01), moderate/moderate, e.g. in ICP 8863/ICP 14209 (03) and low/low, e.g. in ICP 11605/ICPL 87119 (03) (ESM 6). Of the CAPS/dCAPS assays derived from nine sequence variant-containing fragments with poor/no Sanger sequencing reads, two markers (spc002 and spc107) showed polymorphism in six of the parental pairs involving low/high SPC (Fig. 2; ESM 6).

Markers associated with SPC

Sixteen polymorphic CAPS/dCAPS markers in parental pair ICP 5529 and ICP 11605 (ESM 6) were combined with GBS-derived SNPs data in the population to construct an F_2

Table 2 Summary of type and number of detected variants and genes and their distribution in different genomic regions of pigeonpea

Chr/CcLG	Total SNPs	Exonic region					Splice sites	Intronic	Indels	Frame shifts	Intergenic	Het	No genes
		sSNPs	nsSNPs	Stop-gain	Stop-loss	Start-loss							
CcLG01	1721	35	46	1	0	0	2	166	0	1	1470	342	74
CcLG02	2430	42	51	1	0	0	0	420	1	2	1913	692	125
CcLG03	1425	18	30	0	0	0	0	196	2	3	1176	405	99
CcLG04	925	15	22	1	0	0	0	141	0	1	744	168	65
CcLG05	171	3	6	0	0	0	0	34	1	0	127	75	14
CcLG06	726	17	18	1	0	0	0	108	1	1	580	341	66
CcLG07	1105	15	25	0	0	0	1	147	0	3	914	306	62
CcLG08	1436	16	31	0	0	0	0	202	0	0	1187	245	61
CcLG09	514	6	14	0	1	0	0	79	0	0	414	178	24
CcLG10	1016	11	23	0	0	0	0	106	0	0	876	526	48
CcLG11	2564	40	57	1	0	0	0	251	0	0	2215	692	105
Scaffolds[a]	18,931	268	395	21	0	1	7	2244	5	11	15,364	6091	674
Total	32,964	485	718	26	1	1	10	4086	10	22	26,979	10061	1417
Distribution (%)	100	1.47	2.18	0.08	0.00	0.00	0.03	12.40	0.03	0.07	81.84	30.52	

Chr chromosome, *nsSNPs* nonsynonymous SNPs, *sSNPs* synonymous SNPs, *Het* all sequence variants coded in the high protein group as Het, K, M, S, R, Y and W

[a] Includes variants without variant effects thus resulting in unbalanced totals for variants from scaffolds

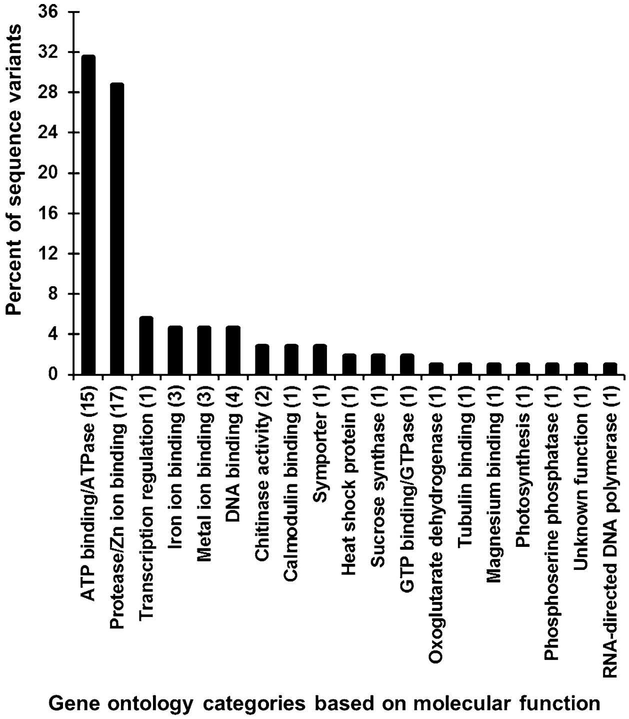

Fig. 1 Grouping of common variant (CV)-selected candidate genes based on GO molecular function. Number in parenthesis on horizontal axis represents the number of genes in the category. (Color figure online)

genetic map (Saxena et al. 2017; Fig. 3). Eleven of the 16 markers could be mapped on to the genetic map with two markers each on CcLG01 and CcLG04, four (CcLG02), one (CcLG07) and two (CcLG08) (Fig. 3). Two of the markers, spc002 and spc107, derived from mutations in the NADH-GOGAT (*C.cajan_04622*) and a copper transporter gene (*C.cajan_05609*) on CcLG02 were found 2.7 and 7.6 cM distances away, respectively, from a QTL explaining 9.0% of the phenotypic variation in SPC. Another marker (spc100) derived from a mutation in a BLISTER gene (*C. cajan_06086*) on the same CcLG02 was 7.8 cM away from a major QTL explaining 11.5% of the phenotypic variation for SPC (Fig. 3) (Obala 2017). Further, single marker analysis (SMA) using regression of F_2 genotype (ESM 7) and phenotype (ESM 8) found four of the 11 assayed CAPS/dCAPs to have significant association with SPC (Table 3). Three of the markers were on CcLG02 and included spc003 derived from a mutation in the NADH-GOGAT gene, spc107 derived from a mutation in a copper transporter gene and spc100 derived from a BLISTER gene. The fourth marker, spc017 was derived from a receptor-like protein kinase gene on CcLG08.

Discussion

The observed quantitative phenotypic variation of seed protein content (SPC) among pigeonpea genotypes (Upadhyaya et al. 2007b; Obala 2017) reflect the complex nature of the trait consistent with observations in other crop plants such as soybean (Hwang et al. 2014; Zhang et al. 2015), pea (Burstin et al. 2007; Krajewski et al. 2012), wheat (Blanco et al. 2012), maize (Guo et al. 2013; Yang et al. 2014) and rice (Mahender et al. 2016). For a quantitative character like SPC, the conventional QTL mapping to identify marker trait associations is laborious, time-consuming and costly (Singh et al. 2016a). Such a scenario is worsened by the low level of polymorphism in pigeonpea, which makes identification of polymorphic markers a daunting task (Saxena et al. 2010). Modern NGS-based genomics approaches that

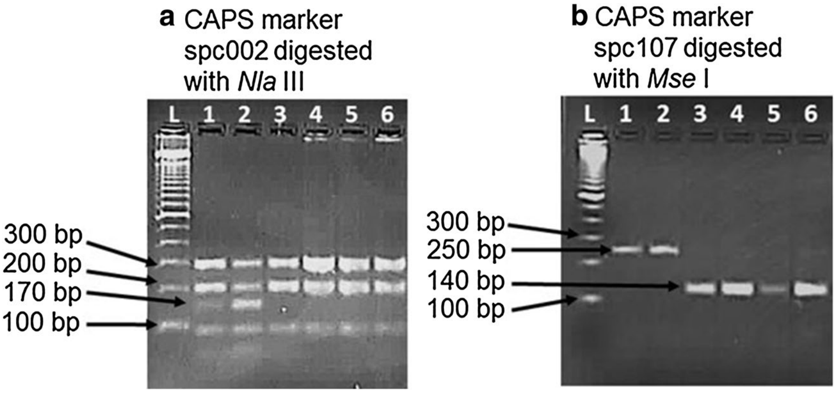
a CAPS marker spc002 digested with *Nla* III
L 1 2 3 4 5 6
300 bp
200 bp
170 bp
100 bp
b CAPS marker spc107 digested with *Mse* I
L 1 2 3 4 5 6
300 bp
250 bp
140 bp
100 bp

Fig. 2 Two CAPS markers indicating the presence of sequence variants in two PCR amplified fragments for which the presence of the variants were not previously confirmed by Sanger sequencing due poor results. **a** CAPS marker spc002 derived from a sequence variant in glutamate synthase gene, **b** CAPS marker spc107 derived from a sequence in a copper transporter gene. The two markers distinguished between high (lanes 1 and 2) and low to moderate (lanes 3–6) seed protein content genotypes, L: 100 bp DNA, 1: HPL 24, 2: ICP 5529, 3: ICP 11605, 4: ICPL 87119, 5: ICP 8863, 6: ICP 14209. (Color figure online)

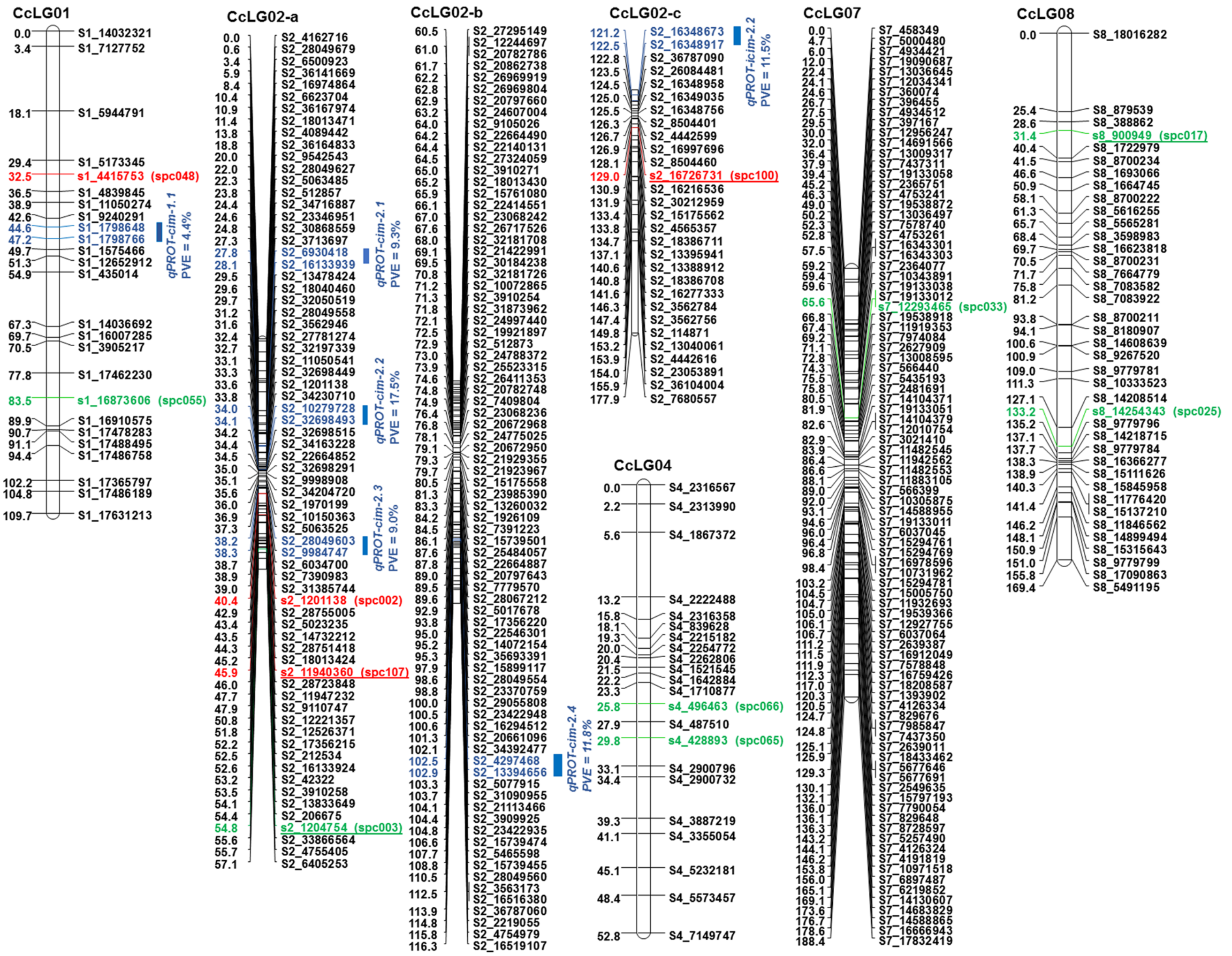

Fig. 3 Five linkage groups from a genetic map of an F_2 mapping population ICP 5529 × ICP 11605 (Saxena et al. 2017) indicating positions of CAPS/dCAPS markers and QTLs associated with seed protein content (Obala 2017). Bars indicate position of QTL. Coloured markers represent sequence variants along with name of CAPS/dCAPS marker in parenthesis, red markers are those located < 10 cM from a QTL, green are markers located > 10 cM from a QTL. Underlined markers showed significant associations with SPC in an F_2 population of the cross ICP 5529 × ICP 11605. (Color figure online)

Table 3 Genic-CAPS/dCAPS markers with significant association with SPC in an F_2 mapping population of the cross ICP 5529 × ICP 11605

Chr.	Gene ID	Marker (type)	Enzyme	R^2 (%)	F-prob	Gene name
CcLG02	*C.cajan_04622*	spc003 (CAPS)	NIa*III*	3.5	0.011	NADH-GOGAT
CcLG02	*C.cajan_05609*	spc107 (CAPS)	Mse*I*	3.7	0.008	Copper transporter
CcLG08	*C.cajan_15445*	spc017 (dCAPS)	Pme*l*	2.2	0.043	Protein kinase
CcLG02	*C.cajan_06086*	spc100 (CAPS)	NIa*III*	2.8	0.023	BLISTER

involve re-sequencing genomes of genotypes contrasting in trait phenotype together with detection of nonsynonymous sequence variants have been found to be efficient for rapid identification of potential candidate genes controlling complex traits in pigeonpea and other crops (Silva et al. 2012; Xu et al. 2014; Singh et al. 2016a, b, 2017). The results obtained from our previous NGS-based trait mapping studies (Singh et al. 2016b, 2017) have encouraged us to use similar approach for identification of candidate genes/markers associated with SPC in pigeonpea.

In the present study, the re-sequencing data obtained from Kumar et al. (2016) was subjected to a common variant filtering strategy (Silva et al. 2012) to detect sequence variants that potentially play a role in SPC variation in pigeonpea. Our initial prioritisation of the CV-detected candidate sequence variants/genes was on the basis of predicted impact of the variants on protein function. This led us to select nonsynonymous, stop, frame-shift, splice-site and indel mutations. A final selection of the candidates was based on information on gene function in relation to the SPC. Eventually, 108 sequence variants in 57 genes were selected and considered for further analysis. The 57 genes belong to 19 GO-molecular function categories. A number of the genes or their homologues have been implicated in the control of SPC variation in other plant species. Such genes include those of sucrose synthase (Zeng et al. 2016) on CcLG01 at position 4,415,753 bp, glutamate synthase (NADH-GOGAT) (Schoenbeck et al. 2000; Nigro et al. 2013) on CcLG02 at position 1,204,754 bp, basic 7S globulin on CcLG02 at position 8,895,098 bp (Yamada et al. 2014), 2-oxoglutarate dehydrogenase (Araújo et al. 2013) on CcLG02 at position 36,162,648 bp, ABC transporter (Upadhyaya et al. 2016) on CcLG03 at position 20,453,445 and 20,477,859 bp, and asparagine synthetase (Lam et al. 2003; Pandurangan et al. 2012) at position 14,801 bp on scaffold 132,767.

Several of the putative candidate genes detected in the present study, although with no known proof that they increase or decrease SPC accumulation, have been reported to play a role in storage protein biosynthesis through various metabolic pathways. For example, genes of the proteolytic pathway such as the aspartic-type endopeptidase (proteases) (EC 3.4.23.-) and RNA-directed DNA polymerase (Reverse transcriptase; EC 2.7.7.49) have been reported to play a role in proteolysis and processing of seed storage proteins (Pereira et al. 2008). Similarly, a number of transcription factors such as Heat shock proteins, e.g. Hsp 40 (Bolon et al. 2010; Ohta et al. 2013), Protein ETHYLENE INSENSITIVE-3 (EIN3) (Cohen et al. 2014), GTP-binding subunit (Lestari et al. 2013), WRKY transcription factor and Myb related proteins have been implicated as broad-range regulators of gene expression (Rahaie et al. 2013). As a considerable number of genes identified from the pigeonpea WGRS data had been previously reported in literature to play roles in SPC in several crops underscores the probable role of these genes in conditioning SPC in pigeonpea. It also indicates a correct selection and grouping of the genotypes used for the detection of the candidate variants and genes in the present study.

To ensure certainty in the existence of the variants detected in the genes, a validation through Sanger sequencing was done. Up to 75.4% of tested sequence variants were found to be correct between one low (ICP 11605) and one high (ICP 5529) SPC genotypes. Both ICP 11605 and ICP 5529 were originally used for sequence variant prediction from the WGRS data (see "Materials and methods"). In comparing results of the present study with that of earlier similar studies, the sequence variant prediction rate from the Illumina WGRS data as verified by Sanger sequencing is close to 79–97% in soybean (Hyten et al. 2010a; Deschamps et al. 2010), but lower than 86% in common bean (Hyten et al. 2010b), > 80% in Tausch's goatgrass (You et al. 2011) and 96.4% in rice (Deschamps et al. 2010). It is, however, higher than the 35.3% in chickpea (Azam et al. 2012). Factors that may contribute to the low sequence variant prediction accuracy in the present study include draft genome assembly and errors associated with sequence alignment, genotype and variant calling (Olson et al. 2015) and use of small datasets (Azam et al. 2012). In addition, the read depths of 9.68–14.03 of the WGRS datasets (Kumar et al. 2016) used for the identification of putative sequence variants may be considered to be relatively low and may also have contributed to the realized sequence variant prediction accuracy. Nonetheless, with an accuracy of 75.4%, 81 out of 108 final selected sequence variants can be expected to be valid and may be useful in genetic studies and breeding applications aimed at improving SPC in pigeonpea. To test this hypothesis, and further verify the presence of the sequence variants, a set of 59 sequence variants were converted into CAPS/dCAPS assays and tested for polymorphism on six (two low, two high and two moderate SPC) genotypes. The highest number of polymorphic markers observed in the high vs low than in the high vs moderate or high vs high SPC genotypes provided confirmation of the potential usefulness of the genic-derived CAPS/dCAPS markers.

With an objective to test for co-segregation of the markers with SPC, 16 polymorphic CAPS/dCAPS markers between parents ICP 5529 and ICP 11605 were assayed on an F_2 mapping population of the two parents. Through SMA, four markers: spc003, spc100, spc107 and spc017 derived from mutations in four genes (NADH-GOGAT, BLISTER, copper transporter and receptor-like protein kinase, respectively), showed significant association with SPC. Of the four genes, a higher expression of NADH-GOGAT in two durum wheat has been associated with higher grain protein content (Nigro et al. 2013). While a BLISTER gene is reported to localize within a major SPC QTL on chromosome 20 of soybean (Lestari et al. 2013). The receptor-like protein kinases have been shown to be differentially expressed between low and high SPC near isogenic lines of soybean (Bolon et al. 2010). However, in the case of the copper transporter gene, no report exists that indicates its functional or positional relationship to SPC in any plant, and may, therefore, be considered novel.

In this study, several sequence variants showing protein changes in genes with possible roles in SPC accumulation were identified in pigeonpea by exploiting WGRS data generated through NGS. A high proportion of the sequence variants were confirmed using Sanger sequencing. Conversion of a subset of the sequence variants into gel-based CAPS/dCAPS markers provided further validation of the variants, and co-segregation analysis of CAPS/dCAPS markers with SPC confirmed potential use of the sequence variants as markers in GAB aimed at improving SPC in pigeonpea. The sequence variants could be added to marker panels for genomic prediction or genotyping arrays for routine use in breeding programs. An example of such strategy is the approach used by Cabezas et al. (2015) in which they effectively employed candidate gene-based sequence variant studies as a means to pre-select relevant markers and aid genomic selection in maritime pine breeding programs. Nonetheless, the actual function of the changed proteins resulting from the DNA sequence variants still needs verification. Gene knockouts using genome editing technologies (Gaj et al. 2016), as well as gene expression analysis (Lovén et al. 2012), could verify the involvement of the sequence variants and associated genes in seed storage protein metabolic pathways in pigeonpea. In addition, it is possible that other types of causative sequence variants have been overlooked in the panel of selected genes as a result of the strategies used to prioritize the candidate variants. Thus, future studies should also include causative variants in the non-coding regions of the targeted putative SPC candidate genes, which are not included in the exon, or in other genes that are not in the panel of putative SPC candidate genes that could have been overlooked by our approach. The potential for other genetic mechanisms, such as copy number variation, large indels, or structural genomic variants to contribute to the underlying mutations should also be investigated.

Acknowledgements Authors are thankful to the United States Agency for International Development (USAID) and ICRISAT for funding various projects related to pigeonpea genomics at ICRISAT. We also thank Mr. Aamir Khan (ICRISAT) for his help in processing the WGRS data and Ms. Anu Chitikineni (ICRISAT) for coordinating Sanger sequencing and primer ordering. This work has been undertaken as part of the CGIAR Research Program on Grain Legumes and Dryland Cereals. ICRISAT is a member of CGIAR Consortium.

Compliance with ethical standards

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of interest Jimmy Obala declares that he has no conflict of interest. Rachit K. Saxena declares that he has no conflict of interest. Vikas K. Singh declares that he has no conflict of interest, C.V. Sameer Kumar declares that he has no conflict of interest. K.B. Saxena declares that he has no conflict of interest, Pangirayi Tongoona declares that he has no conflict of interest. Julia Sibiya declares that she has no conflict of interest. Rajeev K. Varshney declares that he has no conflict of interest.

Data availability statement All data generated or analysed during this study are included in this published article and its supplementary information files.

References

Akibode CS, Maredia M (2011) Global and regional trends in production, trade and consumption of food legume crops. Report submitted to the Standing Panel on Impact Assessment (SPIA) of the CGIAR Science Council. FAO, Rome

Araújo WL, Trofimova L, Mkrtchyan G, Steinhauser D, Krall L, Graf A, Fernie AR, Bunik VI (2013) On the role of the mitochondrial 2-oxoglutarate dehydrogenase complex in amino acid metabolism. Amino Acids 44:683–700

Azam S, Thakur V, Ruperao P, Shah T, Balaji J, Amindala B, Farmer AD, Studholme DJ, May GD, Edwards D, Jones JD, Varshney RK (2012) Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (*Cicer arietinum*; Fabaceae), a crop species without a reference genome. Am J Bot 99:186–192

Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ, Frazer KA (2010) Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res 20:537–545

Blanco A, Mangini G, Giancaspro A, Giove S, Colasuonno P, Simeone R, Signorile A, De Vita P, Mastrangelo L, Cattivelli AM, Gadaleta A (2012) Relationships between grain protein content and grain yield components through quantitative trait locus analyses in a recombinant inbred line population derived from two elite durum wheat cultivars. Mol Breed 30:79–92

Bolon YT, Joseph B, Cannon SB, Graham MA, Diers BW, Farmer AD, May GD, Muehlbauer GJ, Specht JE, Tu ZJ, Weeks N, Xu WW, Shoemaker RC, Vance PC (2010) Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biol 10:41

Burstin J, Marget P, Huart M, Moessner A, Mangin B, Duchene C, Desprez B, Munier-Jolain N, Duc G (2007) Developmental genes have pleiotropic effects on plant morphology and source capacity, eventually impacting on seed protein content and productivity in pea. Plant Physiol 144:768–781

Cabezas JA, González-Martínez SC, Collada C, Guevara MA, Boury C, de María N, Eveno E, Aranda I, Garnier-Géré PH, Brach J, Alía R, Plomion C, Cervera MT (2015) Nucleotide polymorphisms in a pine ortholog of the Arabidopsis degrading enzyme cellulase KORRIGAN are associated with early growth performance in *Pinus pinaster*. Tree Physiol 35:1000–1006

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GRS, Albracht D, Kremitzki M, Rock S, Kotkiewicz H, Kremitzki C, Wollam A, Trani L, Fulton L, Fulton R, Matthews L, Whitehead S, Chow W, Torrance J, Dunn M, Harden G, Threadgold G, Wood J, Collins J, Heath P, Griffiths G, Pelan S, Grafham D, Eichler EE, Weinstock G, Mardis ER, Wilson RK, Howe K, Flicek P, Hubbard T (2011) Modernizing reference genome assemblies. PLoS Biol 9:e1001091

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Ruden DM, Lu X (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w118. Fly 6:1–13

Cohen H, Israeli H, Matityahu I, Amir R (2014) Seed-specific expression of a feedback-insensitive form of cystathionine-g-synthase in Arabidopsis stimulates metabolic and transcriptomic responses associated with desiccation stress. Plant Physiol 166:1575–1592

Coonrod EM, Durtschi JD, Margraf RL, Voelkerding KV (2013) Developing genome and exome sequencing for candidate gene identification in inherited disorders: an integrated technical and bioinformatics approach. Arch Pathol Lab Med 137:415–433

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis KS, Gabriel B, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498

Deschamps S, la Rota M, Ratashak JP, Biddle P, Thureen D, Farmer A, Luck S, Beatty M, Nagasawa N, Michael L, Llaca V, Sakai H, May G, Lightner J, Campbell MA (2010) Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina Genome Analyzer. Plant Genome 3:53–68

FAO (2016) FAOSTAT. Food and Agriculture Organisation of the United Nations. http://faostat3.fao.org. Acccessed 28 Aug 2018

Feng X, Yu X, Tong J (2014) Novel single nucleotide polymorphisms of the insulin-like growth factor-1 gene and their associations with growth traits in common carp (*Cyprinus carpio* L.). Int J Mol Sci 15:22471–22482

Gaj T, Sirk SJ, Shui S-L, Liu J (2016) Genome-editing technologies; principles and applications. Cold Spring Harb Perspect Biol 8:a023754

Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20:490–497

Grattapaglia D (2008) Genomics of Eucalyptus, a global tree for energy, paper and wood. In: Moore P, Ming R (eds) Genomics of tropical crop plants. Springer, New York, pp 257–295

Greilhuber J, Obermayer R (1998) Genome size variation in *Cajanus cajan* (Fabaceae): a reconsideration. Plant Syst Evol 212:135–141

Guo Y, Yang X, Chander S, Yan J, Zhang J, Song T, Li J (2013) Identification of unconditional and conditional QTL for oil, protein and starch content in maize. Crop J 1:34–42

Hwang E-Y, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genom 15:1

Hyten DL, Cannon SB, Song Q, Weeks N (2010a) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genom 11:38

Hyten DL, Song Q, Fickus EW, Quigley CV (2010b) High throughput SNP discovery and assay development in common bean. BMC Genom 11:475

Janninks J (2001) Using interconnected populations to find quantitative trait loci. http://www.reeis.usda.gov/web/crisprojectpages/0189427-using-interconnected-populations-to-find-quantitative-trait-loci.html. Accessed 24 Feb 2016

Joshi NA, Fass JN (2011) Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. https://github.com/najoshi/sickle. Accessed 20 Jan 2016

Korbie DJ, Mattick JS (2008) Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat Protoc 3:1452–1456

Krajewski P, Bocianowski J, Gawłowska M, Kaczmarek Z, Pniewski T, Święcicki W, Wolko B (2012) QTL for yield components and protein content: a multi-environment study of two pea (*Pisum sativum* L.) populations. Euphytica 183:323–336

Krishnan HB, Natarajan SS, Oehrle NW, Garrett WM, Darwish O (2017) Proteomic analysis of pigeonpea (*Cajanus cajan*) seeds reveals the accumulation of numerous stress-related proteins. J Agric Food Chem 65:4572–4581

Kumar V, Khan AW, Saxena RK, Garg V, Varshney RK (2016) First-generation HapMap in *Cajanus* spp. reveals untapped variations in parental lines of mapping populations. Plant Biotech J 14:1673–1681

Lam H-M, Wong P, Chan H-K, Yam K-M, Chen L, Chow C-M, Coruzzi G-M (2003) Overexpression of the ASN1 gene enhances nitrogen status in seeds of Arabidopsis. Plant Physiol 132:926–935

Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Lee S (2012) Converting SNPs to CAPS and dCAPS marker using dCAPS Finder. http://articles.extension.org/pages/32594/converting-snps-to-caps-and-dcaps-marker-using-dcaps-finder. Accessed 17 July 2016

Lestari P, Koh HJ (2013) Development of new CAPS/dCAPS and SNAP markers for rice eating quality. HAYATI J Biosci 20:15–23

Lestari P, Van K, Lee J, Kang YJ, Lee S-H (2013) Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. Front Plant Sci 4:176

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li J, Dai X, Liu T, Zhao PX (2012) LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes. Nucleic Acids Res 40:D1221–D1229

Lim J-H, Yang H-J, Jung K-H, Yoo S-C, Paek N-C (2014) Quantitative trait locus mapping and candidate gene analysis for plant architecture traits using whole genome re-sequencing in rice. Mol Cells 37:149–160

Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA (2012) Revisiting global gene expression analysis. Cell 151:476–482

Mace ES, Buhariwalla HK, Crouch JH (2003) A high throughput DNA extraction protocol for molecular breeding programs. Plant Mol Biol Rep 21:459–459

Machado M, Magalhães WC, Sene A, Araújo B, Faria-Campos AC, Chanock SJ, Scott L, Oliveira G, Tarazona-Santos E, Rodrigues MR (2011) Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies. Investig Genet 2:3

Mahender A, Anandan A, Pradhan SK, Pandit E (2016) Rice grain nutritional traits and their enhancement using relevant genes and QTLs through advanced approaches. SpringerPlus 5:2086

Mligo JK, Craufurd PQ (2005) Adaptation and yield of pigeonpea in different environments in Tanzania. Field Crop Res 94:43–53

Mula MG, Saxena KB (2010) Lifting the level of awareness on pigeonpea—a global perspective. International Crops Research Institute for the Semi-Arid Tropics, Patancheru

Neff MM, Turk E, Kalishman M (2002) Web-based primer design for single nucleotide polymorphism analysis. Trend Genet 18:613–661

Nigro D, Gu YQ, Huo N, Marcotuli I, Blanco A, Gadaleta A, Anderson OD (2013) Structural analysis of the wheat genes encoding NADH-dependent glutamine-2-oxoglutarate amidotransferases and correlation with grain protein content. PLoS One 8:e73751

Obala J (2017) Study of inheritance and identification of molecular markers for seed protein content in pigeonpea (*Cajanus cajan*). PhD Thesis, University of KwaZulu-Natal, South Africa

Odeny DA (2007) The potential of pigeonpea (*Cajanus cajan* (L.) Millsp.) in Africa. Nat Resour Forum 31:297–305

Ohta M, Wakasa Y, Takahashi H, Hayashi S, Kudo K, Takaiwa F (2013) Analysis of rice ER-resident J-proteins reveals diversity and functional differentiation of the ER-resident Hsp70 system in plants. J Exp Bot 64:5429–5441

Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet 6:235

Pandurangan S, Pajak A, Molnar SJ, Cober ER, Dhaubhadel S, Hernández-Sebastiá C, Kaiser WM, Nelson RL, Huber SC, Marsolais F (2012) Relationship between asparagine metabolism and protein concentration in soybean seed. J Exp Bot 63:3173–3184

Pereira CS, da Costa DS, Pereira S, Nogueira FD, Albuquerque PM, Teixeira J, Faro C, Pissarra J (2008) Cardosins in postembryonic development of cardoon: towards an elucidation of the biological function of plant aspartic proteinases. Protoplasma 232:203–213

Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review. Mol Breed 7:275–291

Rahaie M, Xue GP, Schenk PM (2013) The role of transcription factors in wheat under different abiotic stress. In: Vahdati K, Leslie C (eds) Abiotic stress-plant responses and applications in agriculture. InTechOpen, London. https://doi.org/10.5772/54795

Rao PP, Birthal PS, Bhagavatula S, Bantilan MCS (2010) Chickpea and pigeonpea economies in Asia: facts, trends and outlook. International Crops Research Institute for the Semi-Arid Tropics, Patancheru

Rios J, Stein E, Shendure J, Hobbs HH, Cohen JC (2010) Identification by whole genome resequencing of gene defect responsible for severe hypercholesterolemia. Hum Mol Genet 19:4313–4318

Roach JC, Glusman G, Smit AFA, Hu VCD, Hubley R, Shannon RT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636–639

Saxena KB, Sawargaonkar SL (2015) Genetic enhancement of seed proteins in pigeonpea—methodologies, accomplishments and opportunities. Int J Sci Res 4:254–258

Saxena RK, Prathima C, Saxena KB, Hoisington DA, Singh NK, Varshney RK (2010) Novel SSR markers for polymorphism detection in pigeonpea (*Cajanus* spp.). Plant Breed 129:142–148

Saxena RK, Obala J, Sinjushin A, Sameer-Kumar CV, Saxena KB, Varshney RK (2017) Characterization and mapping of Dt1 locus which co-segregates with CcTFL1 for growth habit in pigeonpea. Theor Appl Genet 130:1773–1784

Schoenbeck MA, Temple SJ, Trepp GB, Blumenthal JM, Samac DA, Gantt SJ, Hernandez G, Vance CP (2000) Decreased NADH-glutamate synthase activity in nodules and flowers of alfalfa (*Medicago sativa* L.) transformed with an antisense glutamate synthase transgene. J Exp Bot 51:29–39

Silva J, Scheffler B, Sanabria Y, de Guzman C, Galam D, Farmer A, Woodward J, May G, Oard J (2012) Identification of candidate genes in rice for resistance to sheath blight disease by whole genome sequencing. Theor Appl Genet 124:63–74

Singh VK, Khan AW, Saxena RK, Kumar V, Kale SM, Sinha P, Chitikineni A, Pazhamala LT, Garg V, Sharma M, Sameer-Kumar CV, Parupalli S, Vechalapu S, Patil S, Muniswamy S, Ghanta A, Yamini KN, Dharmaraj PS, Varshney RK (2016a) Next-generation sequencing for identification of candidate genes for Fusarium wilt and sterility mosaic disease in pigeonpea (*Cajanus cajan*). Plant Biotechnol J 14:1183–1194

Singh VK, Khan AW, Jaganathan D, Thudi M, Roorkiwal M, Takagi H, Garg V, Kumar V, Chitikineni A, Gaur PM, Sutton T, Terauchi R, Varshney RK (2016b) QTL-seq for rapid identification of candidate genes for 100-seed weight and root/total plant dry weight ratio under rainfed conditions in chickpea. Plant Biotechnol J 14:2110–2119

Singh VK, Khan AW, Saxena RK, Sinha P, Kale SM, Parupalli S, Kumar V, Chitikineni A, Vechalapu S, Sameer-Kumar CV, Sharma M, Ghanta A, Yamini KN, Muniswamy S, Varshney RK (2017) Indel-seq: a fast-forward genetics approach for identification of trait-associated putative candidate genomic regions and its application in pigeonpea (*Cajanus cajan*). Plant Biotechnol J 15:906–914

Sobreira NLM, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, Ge D, Shianna KV, Smith JP, Maia JM, Gumbs CE, Pevsner J, Thomas G, Valle D, Hoover-Fong JE, Goldstein DB (2010) Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet 6:e1000991

The UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:D190–D195

Upadhyaya HD, Reddy KN, Sastry DVSSR, Gowda CLL (2007a) Identification of photoperiod insensitive sources in the world collection of pigeonpea at ICRISAT. J SAT Agric Res 3:46–49

Upadhyaya HD, Reddy KN, Gowda CLL, Silim SN (2007b) Patterns of diversity in pigeonpea (*Cajanus cajan* (L.) Millsp.) germplasm collected from different elevations in Kenya. Genet Resour Crop Evol 54:1787–1795

Upadhyaya HD, Bajaj D, Narnoliya L, Das S, Kumar V, Gowda CLL, Sharma S, Tyagi AK, Parida SK (2016) Genome-wide scans for delineation of candidate genes regulating seed protein content in chickpea. Front Plant Sci 7:302

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, Whaley AM, Farmer AD, Sheridan J, Iwata A, Tuteja R, Penmetsa RV, Wu W, Upadhyaya HD, Yang S, Shah T, Saxena KB, Michael T, McCombie WR, Yang B, Zhang G, Yang H, Wang J, Spillane C, Cook DR, May GD, Xu X, Jackson SA (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

Varshney RK, Saxena RK, Upadhyaya HD, Khan AW, Yu Y, Kim C, Rathore A, Kim D, Kim J, An S, Kumar V, Anuradha G, Yamini KN, Zhang W, Muniswamy S, Kim J, Penmetsa RV, von Wettberg E, Datta SK (2017) Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. Nat Genet 49:1082–1088

Xu J, Yuan Y, Xu Y, Zhang G, Guo X, Wu F, Wang Q, Rong T, Pan G, Cao M, Tang Q, Gao S, Liu Y, Wang J, Lan H, Lu Y (2014) Identification of candidate genes for drought tolerance by whole-genome re-sequencing in maize. BMC Plant Biol 14:83

Yamada T, Mori Y, Yasue K, Maruyama N, Kitamura K, Abe J (2014) Knockdown of the 7S globulin subunits shifts distribution of nitrogen sources to the residual protein fraction in transgenic soybean seeds. Plant Cell Rep 33:1963–1976

Yang GH, Dong YB, Li YL, Wang QI, Shi QL, Zhou Q (2014) QTL verification of grain protein content and its correlation with oil content by using connected RIL populations of high-oil maize. Genet Mol Res 13:881–894

You FM, Huo N, Gu YQ, Luo M-C, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinform 9:253

You FM, Huo N, Deal KR, Gu YQ, Luo M-C, McGuire PE, Dvorak J, Anderson OD (2011) Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. BMC Genom 12:59

Zeng Y-D, Sun J-L, Bu S-H, Deng K-S, Tao T, Zhang Y-M, Zhang T-Z, Du X-M, Zhou B-L (2016) EcoTILLING revealed SNPs in GhSus genes that are associated with fiber- and seed-related traits in upland cotton. Sci Rep 6:29250

Zhang YH, Liu MF, He JB, Wang YF, Xing GN, Li Y, Yang SP, Zhao TJ, Gai JH (2015) Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr.]. Theor Appl Genet 128:1061–1072