Journal
Association of Applied Biologists
Society for Experimental Biology

Plant Biotechnology Journal (2016) **14**, pp. 1183–1194 doi: 10.1111/pbi.12470

Next-generation sequencing for identification of candidate genes for *Fusarium* wilt and sterility mosaic disease in pigeonpea (*Cajanus cajan*)

Vikas K. Singh[1], Aamir W. Khan[1], Rachit K. Saxena[1], Vinay Kumar[1], Sandip M. Kale[1], Pallavi Sinha[1], Annapurna Chitikineni[1], Lekha T. Pazhamala[1], Vanika Garg[1], Mamta Sharma[1], Chanda Venkata Sameer Kumar[1], Swathi Parupalli[1], Suryanarayana Vechalapu[1], Suyash Patil[1], Sonnappa Muniswamy[2], Anuradha Ghanta[3], Kalinati Narasimhan Yamini[3], Pallavi Subbanna Dharmaraj[2,†] and Rajeev K. Varshney[1,4,*]

[1]*International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India*
[2]*Agricultural Research Station (ARS)-Gulbarga, University of Agricultural Sciences (UAS), Raichur, Karnataka, India*
[3]*Institute of Biotechnology, Professor Jayshankar Telangana State Agricultural University (PJTSAU), Telangana, India*
[4]*School of Plant Biology and Institute of Agriculture, The University of Western Australia, Crawley, WA, Australia*

Received 26 December 2014;
revised 29 July 2015;
accepted 6 August 2015.
**Correspondence* (Tel 91 40 30713305;
fax 91 40 30713074;
email r.k.varshney@cgiar.org)
†Deceased.

Keywords: *Fusarium* wilt, sterility mosaic disease, bulked segregant analysis, whole-genome re-sequencing, SNP index, nonsynonymous SNPs.

Summary

To map resistance genes for *Fusarium* wilt (FW) and sterility mosaic disease (SMD) in pigeonpea, sequencing-based bulked segregant analysis (Seq-BSA) was used. Resistant (R) and susceptible (S) bulks from the extreme recombinant inbred lines of ICPL 20096 × ICPL 332 were sequenced. Subsequently, SNP index was calculated between R- and S-bulks with the help of draft genome sequence and reference-guided assembly of ICPL 20096 (resistant parent). Seq-BSA has provided seven candidate SNPs for FW and SMD resistance in pigeonpea. In parallel, four additional genotypes were re-sequenced and their combined analysis with R- and S-bulks has provided a total of 8362 nonsynonymous (ns) SNPs. Of 8362 nsSNPs, 60 were found within the 2-Mb flanking regions of seven candidate SNPs identified through Seq-BSA. Haplotype analysis narrowed down to eight nsSNPs in seven genes. These eight nsSNPs were further validated by re-sequencing 11 genotypes that are resistant and susceptible to FW and SMD. This analysis revealed association of four candidate nsSNPs in four genes with FW resistance and four candidate nsSNPs in three genes with SMD resistance. Further, *In silico* protein analysis and expression profiling identified two most promising candidate genes namely *C.cajan_01839* for SMD resistance and *C.cajan_03203* for FW resistance. Identified candidate genomic regions/SNPs will be useful for genomics-assisted breeding in pigeonpea.

Introduction

Pigeonpea (*Cajanus cajan* L. Millsp.) is an economically important grain legume crop in the developing countries of the tropical and subtropical regions of the world (Varshney *et al.*, 2012a). The crop productivity of pigeonpea is severely affected by biotic stresses such as *Fusarium* wilt (FW) and sterility mosaic disease (SMD). *Fusarium* wilt is caused by *Fusarium udum,* and SMD is caused by a pigeonpea sterility mosaic virus (PSMV) and transmitted by the eriophyid mite, *Aceria cajani.* These two biotic stresses of pigeonpea could result in complete yield loss (Reddy *et al.*, 2012). The annual losses due to FW and SMD have been reported to be US $ 113 million (Saxena *et al.*, 2010a).

To develop FW- and SMD-resistant pigeonpea lines through molecular breeding, identification of genomic regions/QTLs or candidate genes responsible for resistance to diseases is an important step. Once a marker (or candidate gene) associated with resistance is identified and validated; marker-assisted selection (MAS) can be used for introgression of resistance in susceptible genotypes (Varshney *et al.*, 2012b). However, traditional QTL mapping approach that involves identification of parental polymorphisms and genotyping the entire population with polymorphic markers is time-consuming and labour intensive (Abe *et al.*, 2012). To map simply inherited traits like disease resistance, bulked segregant analysis (BSA) approach was proposed by Michelmore *et al.* (1991). Bulked segregant analysis approach involves screening of the extreme bulks along with the parents with a large number of markers, and subsequently, polymorphic markers showing the similar pattern in the bulks with respect to their corresponding parental genotypes are used to screen the entire population. This approach has been extensively used for trait mapping in a number of crops (see Semagn *et al.*, 2010).

Advent of the next-generation sequencing (NGS) technologies and due to declining cost in per sample sequencing has drastically accelerated the pace with which candidate genes/genomic regions were identified (Schneeberger *et al.*, 2009). Consequently, many recent approaches were proposed using BSA combined with whole-genome re-sequencing (WGRS) for rapid identification of candidate genes of interest termed as 'fast forward genetics' (Mokry *et al.*, 2011). Next-generation sequencing-based BSA approaches were successfully applied to model crop, *Arabidopsis* (~135 Mb of genome size) for identification of candidate genes for growth habit and colour of leaves (Schneeberger *et al.*, 2009; approach SHOREmap), cell wall composition (Austin *et al.*, 2011; approach next-generation mapping), suppressor mutant (Uchida *et al.*, 2011; approach

SNPing; Hartwig *et al.*, 2012; approach isogenic mapping-by-sequencing) and gametophyte lethal mutation (Lindner *et al.*, 2012; approach SNP-ratio mapping). Such methods have also been successfully applied in crop plants like rice (~389 Mbp of genome size) for identification of candidate genes for pale green leaves, semi dwarfism (Abe *et al.*, 2012; approach MutMap), blast resistance (Takagi *et al.*, 2013a; approach MutMap-Gap) and lethal phenotype associated with plant development (Fekih *et al.*, 2013; approach MutMap+).

Recently, QTL-seq method was proposed in rice, which is a powerful approach for handling quantitative traits (Takagi *et al.*, 2013b). This approach was successfully utilized for mapping genomic region for blast resistance and seedling vigour in rice (Takagi *et al.*, 2013b), flowering associated QTL in cucumber (Lu *et al.*, 2014) and seed size and root trait ratio in chickpea (unpublished). Similarly, WGRS-based BSA approach based on G′ statistics was utilized for identification of candidate genomic region for cold tolerance in rice (Yang *et al.*, 2013). In addition, WGRS of contrasting parents has been used for identification of nonsynonymous SNPs (nsSNPs) to map the candidate genes for sheath blight resistance in rice and drought tolerance in maize (Silva *et al.*, 2012; Xu *et al.*, 2014).

Fusarium wilt resistance in pigeonpea has been found to be controlled by different gene actions in various genetic backgrounds, ranging from single to multiple genes with complementary to duplicate gene actions (Saxena, 2008). On the other hand, SMD resistance has been found to be governed by a single gene or two recessive genes (Gnanesh *et al.*, 2011). Therefore, for identification of SNPs for multiple genes, associated with FW and SMD resistance in pigeonpea, we have analysed extreme bulks of resistant (R-bulk) and susceptible (S-bulk) RILs using Seq-BSA approach. Furthermore, nonsynonymous SNPs (nsSNPs) approach has been used to complement Seq-BSA approach. Re-sequencing of additional genotypes, *In silico* protein analysis and transcription profiling have validated and shortlisted candidate genomic regions/SNPs conferring resistance to FW and SMD in pigeonpea.

Results

This study uses a combined approach of Seq-BSA and WGRS-based nsSNPs for identification of candidate resistance genes for FW and SMD. The detailed approach has been illustrated in Figure 1.

FW and SMD screening of genotypes

Six genotypes namely ICPL 20096, ICPL 332, ICPL 20097, ICP 8863, ICPB 2049 and ICPL 99050 along with one mapping population (ICPL 20096 × ICPL 332) comprising of 188 lines were screened for resistance to FW and SMD in this study. Our screening results showed resistance and/or susceptibility of these genotypes to either or both FW and SMD in multilocation phenotyping. Three genotypes (ICPL 20096, ICPL 20097 and ICPL 99050) were found to be resistant to both FW and SMD, and their disease reaction ranged from 0% to 4.4% for FW and 0.0% for SMD. However, one genotype namely ICPL 332 was found highly susceptible for both FW and SMD with disease reaction of 100% for each disease. The genotypes, ICP 8863 and ICPB 2049, were found to be resistant only to FW (0.6%) and SMD (0%), respectively. The detailed observations are presented in Table 1. Similarly, screening of the mapping population showed disease incidence score from 0% to 100% for FW and 0% to 100% for SMD.

Sequencing-based BSA (Seq-BSA) approach

Construction and sequencing of R- and S-bulks

Based on the phenotyping data on 188 RILs, 16 resistant and 16 susceptible RILs to FW as well as SMD were selected for the constitution of R- and S-bulks (Figure 2). The phenotypic disease score of 16 RILs of the R-bulk ranged from 0% to 5.0% (for FW) and 0% to 2.6% (for SMD). Similarly, the phenotypic disease score of selected RILs of the S-bulk varied from 80.0% to 100% (for FW) and 72.62% to 100% (for SMD). Each DNA bulk (R- and S-bulk) along with the resistant parent (ICPL 20096) was subjected for whole-genome re-sequencing (WGRS) using Illumina (MiSeq) sequencing platform. As a result, a total of 8.99 Gb (14.85 X coverage) sequence data for R-bulk, 8.43 Gb (13.91 X coverage) for S-bulk and 9.27 Gb data (15.30 X coverage) for ICPL 20096 were generated (Table 1).

Genomewide SNP index analysis

A total of 37.53 million filtered reads of the resistant parent (ICPL 20096) were used for developing reference-guided assembly. This resulted in genome coverage of 94.46%. The cleaned sequence data for the R- (36.38 million reads) and S-bulks (33.91 million reads) were then used for mapping onto the developed reference-guided assembly of ICPL 20096. Subsequently, SNP index was calculated from R- and S-bulks using QTL-seq pipeline (http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap). Based on the stringent selection criteria of read depth ≥7 in both bulks and SNP index ≥0.3 in either of the bulks, 35 877 SNPs were identified on all 11 linkage groups (Figures S1 and S2). However, only 4139 (11.54%) SNPs were found to have homozygous allele calls in both bulks (Table S1). Finally, seven candidate SNPs with delta SNP index = −1 were selected that were present on five different chromosomes (CcLG02, CcLG07, CcLG08, CcLG10 and CcLG11) (Table 2 and Figure 3). To identify the causative SNPs for FW and SMD resistance, chi-square test was conducted for the seven SNPs along with neighbouring SNPs (one on each flanking side) in sequence data of both R- and S-bulks. The analysis revealed that the probability of these SNPs to

Figure 1 Schematic representation of the combined approach of sequencing-based bulked segregant analysis (Seq-BSA) and nsSNPs substitution for identification of candidate genes for *Fusarium* wilt (FW) and sterility mosaic disease (SMD) resistance in pigeonpea. (a) Two contrasting parents, ICPL 20096 (R) and ICPL 332 (S) were crossed to develop F_7 RILs segregating for FW and SMD resistance through single-seed descent method. (b, c) Phenotypic score of RILs for FW and SMD resistance resulted in the selection of highly resistant and highly susceptible RILs to form the resistant (R) and susceptible (S) bulks. (d) These two bulks along with resistant parent were subjected to whole-genome re-sequencing (WGRS) for identification of SNPs and SNP index through QTL-seq pipeline. (e) Candidate genome regions were identified based on the SNP index (0 and 1). (f and g) WGRS was performed on the four contrasting parents (ICPL 20097, ICP 8863, ICPL 99050 and ICPB 2049) to identify nsSNPs. (h) WGRS data of the contrasting parents and the bulks defined the nsSNPs associated to the genomic regions to FW and SMD resistance. (i–l) Based on the WGRS data and nsSNPs analysis, candidate genes were subsequently selected for functional validation.

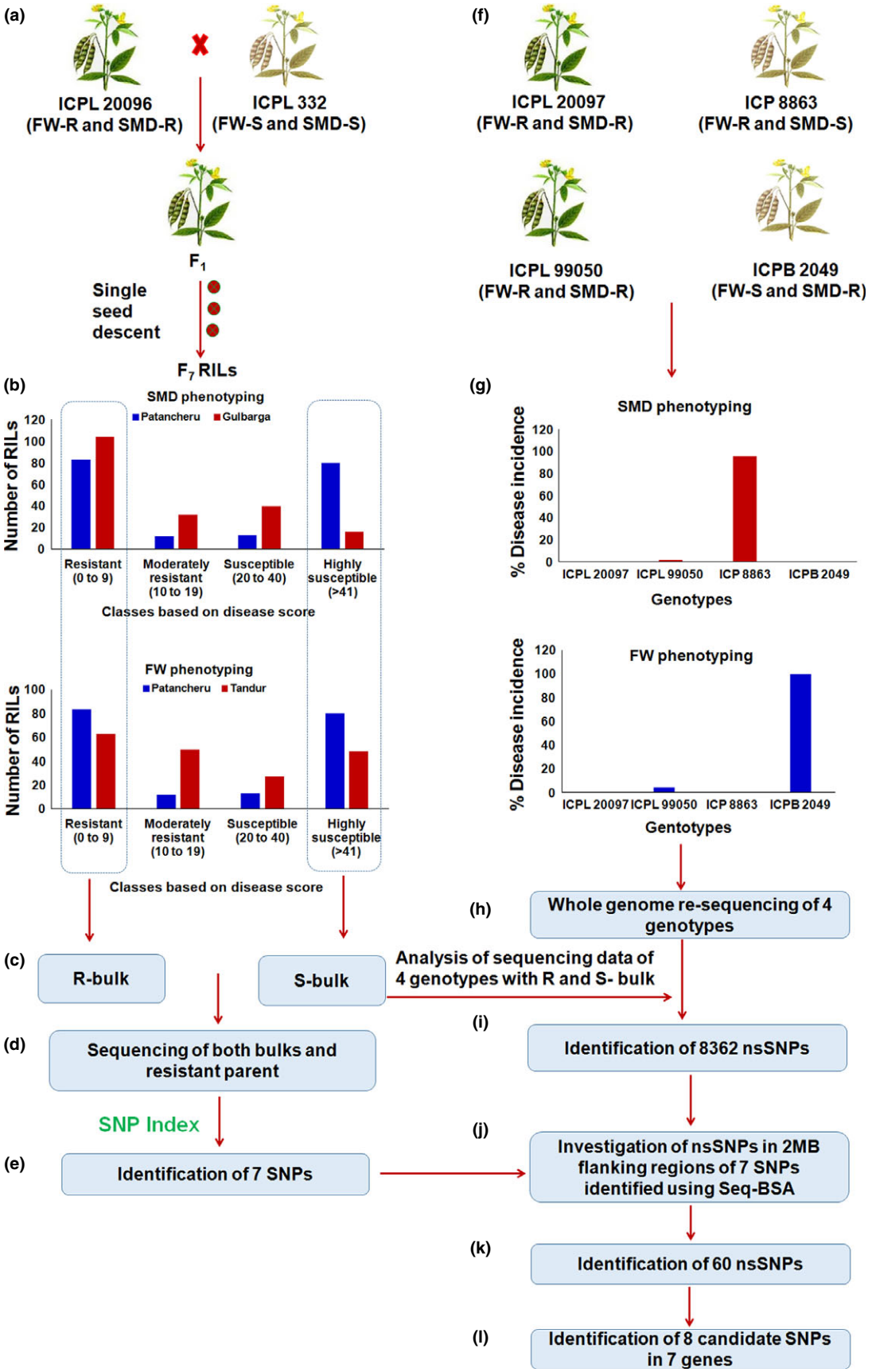

Table 1 List of pigeonpea genotypes with phenotypic score and summary of Illumina sequencing of parental lines and bulks

Sample	Pedigree	Phenotyping for FW		Phenotyping for SMD		Illumina Sequencing			
		Patancheru (average of 2012–2013 and 2013–2014)	Gulbarga (average of 2012–2013 and 2013–2014)	Patancheru (average of 2012–2013 and 2013–2014)	Tandur (average of 2012–2013 and 2013–2014)	Data generated (Gb)	% Alignment	% coverage	Mean depth
Genotypes									
ICPL 20096*,†,§,¶	ICPL 87119 × ICP 12746	4.20 (R)	–	0.0 (R)	–	9.27	90.61	89.21	13.4
ICP 8863‡,¶	Selection from landrace	0.6 (R)	–	95.4 (HS)	–	6.72	95.41	88.96	9.97
ICPB 2049‡,¶	CMS 2039 × ICP 6697	99.9 (HS)	–	0.0 (R)	–	9.98	92.94	89.10	14.88
ICPL 20097‡,¶	ICPL 87119 × ICP 12746	0.0 (R)	–	0.0 (R)	–	7.81	91.84	88.66	11.51
ICPL 99050‡,¶	C11 × Banda Palera	4.4 (R)	–	0.0 (R)	–	10.36	91.64	89.48	15.12
ICPL 332†,§,¶	Selection from ICP 1903	100.0 (HS)	–	100.0 (HS)	–	9.60	93.02	89.58	14.25
ICPL 87119¶,**	C11 × ICP1-6-W3-W	0.0 (R)	–	0.0 (R)	–	–	–	–	–
HPL 24¶	Baigani × *A. scarabaeoides*	0.0 (R)	–	–	–	7.39	95.16	89.19	11.32
ICPL 85063¶	Selection from ICPL 1903	100 (HS)	–	0.0 (R)	–	9.94	93.31	89.31	14.87
ICPL 87¶	T 21 × ICP 6393	47 (HS)	–	–	–	6.01	94.96	88.55	9.20
ICPL 88039¶	Selection from ICPL 161	94.0 (HS)	–	–	–	6.35	93.28	86.77	9.59
Mapping population									
188 RILs	ICPL 20096 × ICPL 332	0%–100%	0%–100%	0%–100%	0%–100%	–	–	–	–
R-bulk†,‡,¶,††	ICPL 20096 × ICPL 332	0.0%–5.0%		0.0%–2.6%	–	8.99	99.77	88.33	12.30
S-bulk†,‡,¶,‡‡	ICPL 20096 × ICPL 332	80.0%–100%		72.62%–100%	–	8.43	80.56	89.15	14.76

*ICPL 20096 short reads were aligned to the publicly available pigeonpea genome of Asha (ICPL 87119) (http://www.icrisat.org/gt-bt/iipg/genomedata.zip).

†Parents and bulks used for Seq-BSA analysis.

‡Parents and bulks used for nonsynonymous SNP (nsSNPs) analysis.

§Parents used for expression profiling studies.

¶Parents and bulks used for identified nsSNPs validation.

**Reference genome assembly of publicly available pigeonpea genotype ICPL 87119 genome (Asha) was used in this study (Varshney *et al.*, 2012a).

††R-bulk was constituted using 16 resistant RILs identified at both locations (Patancheru and Gulbarga).

‡‡S-bulk was constituted using 16 susceptible RILs identified at both locations (Patancheru and Gulbarga).

Note: genotypes were categorized in different groups based on per cent disease incidence (PDI): resistant (0%–9.99% of PDI), moderately resistance (10%–19.99% of PDI), susceptible (20%–40% of PDI) and highly susceptible (>40% of PDI).

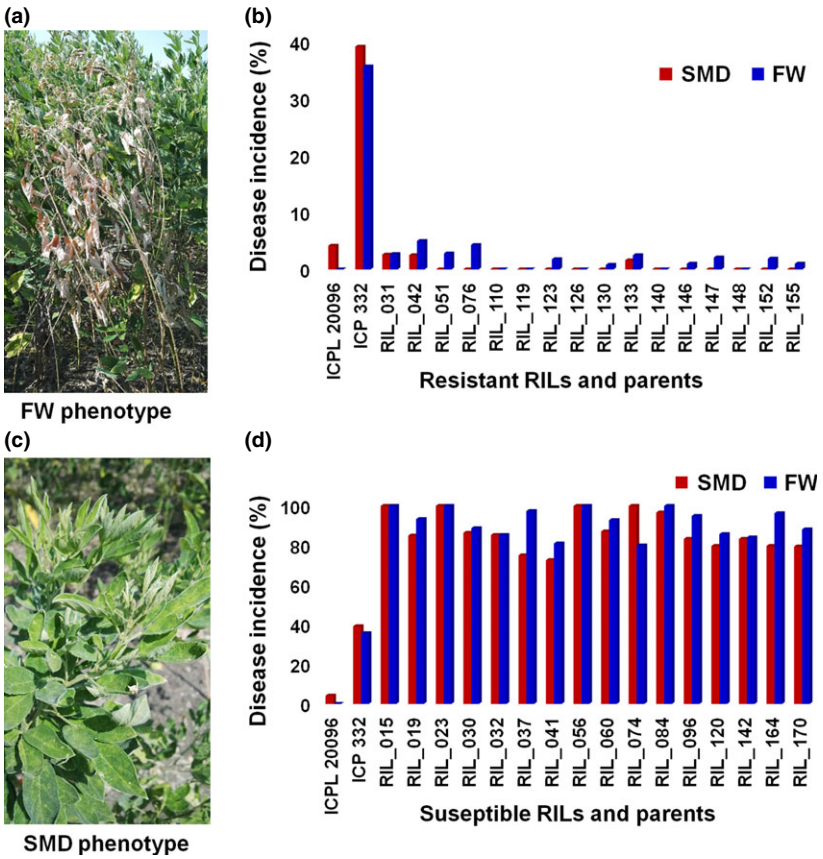

Figure 2 Phenotyping of RILs and parents for *Fusarium* wilt (FW) and sterility mosaic disease (SMD) resistance. (a) FW is seed and soil borne fungal disease caused by *Fusarium udum*. *Fusarium* wilt causes the complete death of plant and reduction in yield due to loss of leaf turgidity, interveinal clearing and mild chlorosis to the bright yellow colour of leaves. (b) SMD is a viral disease caused by *Pigeonpea sterility mosaic virus* (PSMV). This disease can be identified as patches of bushy, pale green plants. Due to excess vegetative growth, without growing into reproductive phase, this condition is known as *green plague* of pigeonpea. (c) Based on the phenotyping of RILs for FW and SMD resistance at two different locations, a total of 16 resistant RILs were used to develop resistant bulk (R-bulk). (d) A total of 16 susceptible RILs with percent disease score comparable to susceptible parent were used to produce susceptible bulk (S-bulk).

follow 1 : 1 binomial distribution with a *P*-value <0.01 did not fit the expected pattern of co-segregation, indicating their causativeness (Table S2).

Nonsynonymous SNPs (nsSNPs) substitution approach

WGRS of additional set of resistant and susceptible genotypes

Four other genotypes viz. ICPL 20097 (resistant to FW and SMD), ICP 8863 (resistant to FW and susceptible to SMD), ICPB 2049 (susceptible to FW and resistant to SMD) and ICPL 99050 (resistant to FW and SMD) were re-sequenced at >10 X coverage (Table 1). In brief, 6.72–10.36 Gb data with 11.09–17.09 X coverage were generated. Alignment of cleaned data from these genotypes and R- and S-bulks from ICPL 20096 × ICPL 332 population mentioned in Seq-BSA to the draft genome sequence indicated a higher level mapping in the range of 91.64% to 95.41%. In terms of depth coverage, alignment of sequence reads for these samples with respect to the draft genome was found in the range of 9.97 to 15.12 mean depth.

Table 2 Identification of SNPs between resistant and susceptible bulks using Seq-BSA approach

Linkage group	Position	Resistant parent base	R-bulk base	Read depth of R-bulk (X coverage)	Phred quality score of R-bulk	SNP index of R-bulk	S-bulk base	Read depth of S-bulk (X coverage)	Phred quality score of S-bulk	SNP index of S-bulk	Δ SNP index
CcLG02	26 551 810	T	T	7	48	0	A	7	48	1	−1
CcLG07	16 064 896	G	G	8	51	0	C	8	51	1	−1
CcLG07	18 411 642	G	G	11	60	0	A	9	54	1	−1
CcLG08	354 473	G	G	10	57	0	C	7	48	1	−1
CcLG10	7 815 091	G	G	9	54	0	A	7	48	1	−1
CcLG11	19 958 148	A	A	9	54	0	C	10	57	1	−1
CcLG11	34 310 320	C	C	7	48	0	A	7	48	1	−1

SNP index = 0 means bulked DNA representing resistant parent genome.

SNP index = 1 means bulked DNA representing susceptible parent.

Δ SNP index = −1 bulked DNA representing resistant parent genome.

Phred quality score ≥40: Probability of incorrect base call 1 in 10 000 (99.99%); Phred quality score ≥50: Probability of incorrect base call 1 in 100 000 (99.999%); Phred quality score ≥60: Probability of incorrect base call 1 in 1 000 000 (99.9999%).

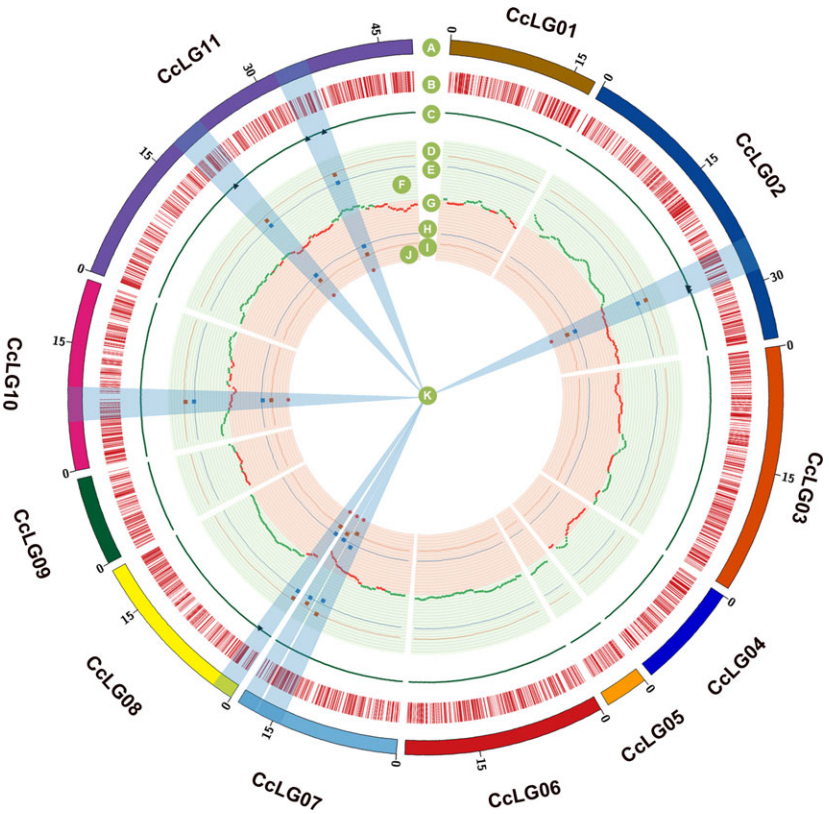

Figure 3 Global distribution of Δ SNP index and nonsynonymous SNPs. (a) Psuedomolecules of reference genome Asha adopted from Varshney *et al.* (2012a). (b) Genomewide distribution of nonsynonymous SNPs (nsSNPs) identified between resistant and susceptible genotypes and bulks. (c) Positions of identified candidate genes with nsSNPs in the vicinity of identified genomic regions through delta SNP index. (d) Upper probability values at 99% confidence ($P < 0.01$). (e) Upper probability values at 95% confidence ($P < 0.05$). (f) Region in green colour representing SNP index ranging from 0 to 1. (g) Genomewide delta SNP index, including those genomic regions with 0 and 1 SNP index, that is same in resistant parent (RP) and resistant bulk (R-bulk) but entirely different in susceptible bulk (S-bulk). These particular positions are marked with red dots along with their upper and lower confidence interval values at 99% and 95% probability values. (h) Lower probability values at 99% confidence ($P < 0.01$). (i) Lower probability values at 95% confidence ($P < 0.05$). (j) Region in red colour representing SNP index ranging from 0 to −1. (k) 2-Mb selected genomic regions flanked both sides to each identified genomic positions with 0 to 1 SNP index for identification of candidate nsSNPs in the target regions.

Genomewide nsSNPs

Detailed analysis of WGRS data sets for above four samples and two bulks identified 226 393 SNPs, which ranged from 3915 (CcLG05) to 43 367 (CcLG11) with an average 20 581 SNPs per linkage group (Table S3). The majority of these SNPs (42.18%) were found in the intergenic region while only ~3% SNPs were present in the exonic region (Table S4). Of 226 393 SNPs, only 8362 SNPs were found to be nsSNPs.

Combined approach of Seq-BSA and nsSNP analysis

To obtain converging evidences from the Seq-BSA and nsSNP analysis for causative SNPs, 2-Mb regions of the seven candidate SNPs identified through Seq-BSA were investigated for the presence of nsSNPs identified in the second approach. A total of 60 nsSNPs were detected in seven genomic regions identified using Seq-BSA (Figure 3). Of these 60 nsSNPs, 16 SNPs were found on each CcLG02 and CcLG11 while seven, eight and 13 SNPs were present on CcLG08, CcLG10 and CcLG07, respectively. Subsequently, haplotype analysis was carried out for 60 nsSNPs in all resistant (ICPL 20097 and ICP 99050) and susceptible (ICP 8863 and ICP 2049B) genotypes along with the resistant and susceptible bulks. Of 60 nsSNPs, eight nsSNPs (four on CcLG02, one on CcLG08 and three on CcLG11) showed specific haplotype in all resistant genotypes and R-bulk, while the other (alternate) allele in all susceptible genotypes and S-bulk. These candidate nsSNPs (haplotypes) were found in seven candidate genes present on three linkage groups (CcLG02, CcLG08 and CcLG11).

Association of nsSNPs/genes with resistance to FW/SMD

With an objective to identify association of nsSNPs/genes with resistance to a particular disease, eight nsSNPs were compared in two different combinations of genotypes one each for FW and SMD. For FW, the re-sequencing data of three resistant genotypes (ICPL 99050, ICPL 20097 and ICP 8863) and one susceptible genotype (ICPB 2049) were compared. Similarly, in the case of SMD, the re-sequencing data of three resistant genotypes (ICPL 99050, ICPL 20097 and ICPB 2049) and one susceptible genotype (ICP 8863) were compared (Table 3). Based on WGRS data sets, the SNP identified in candidate gene *C.cajan_07067* (at position 27 324 239 and 27 324 261 bp) had T (R-bulk) to G (S-bulk) substitution at both SNP position and based upon the allele calls in other genotypes this gene was found to be associated with SMD resistance. Similarly, after comparative analyses in other genotypes T (R-bulk) to G (S-bulk) substitution for the candidate gene *C.cajan_07078* (PHD finger protein) and G (R-bulk) to A (S-bulk) substitution for the candidate gene *C.cajan_07124* (rRNA-processing protein) were found specific for FW resistance.

Comparative analysis of candidate genes *C.cajan_15535* (copia-like retrotransposable) and *C.cajan_01839* (serine–threonine protein phosphatase) showed C (R-bulk) to G (S-bulk) and A (R-bulk) to C (S-bulk) substitution, respectively, and based on the allele calls in other genotypes, these two genes were found specific for SMD resistance. Analysis of the remaining candidate genes *C.cajan_02962* (NADH dehydrogenase) and *C.cajan_03203* (retrovirus-like polyprotein) had T (R-bulk) to C (S-bulk) and C (R-bulk) to A (S-bulk) substitutions, respectively, and based on the allele calls in resistant (ICPL 20096, ICPL 99050, ICPL 20097 and ICP 8863) and susceptible (ICPB 2049) genotypes, both of the genes were found to be specific to FW resistance.

Validation of identified SNPs

In addition to using the re-sequencing data from above-mentioned genotypes, a total of 39.29 Gb data with 9.20–14.88 X coverage

Table 3 Association of nsSNPs to the candidate genes responsive to FW and SMD diseases

			Seq-BSA approach			nsSNPs substitution approach							
						FW				SMD			
Linkage group	Genes	nsSNPs position (bp)	ICPL 20096 (R* to FW & SMD)	R-bulk† (R* to FW & SMD)	S-bulk‡ (S* to FW & SMD)	ICPL 99050 (R*)	ICPL 20097 (R*)	ICP 8863 (R*)	ICPB 2049 (HS§)	ICPL 99050 (R*)	ICPL 20097 (R*)	ICPB 2049 (R*)	ICP 8863 (HS§)
FW associated nsSNPs													
CcLG02	*C.cajan_07078*	27 426 866	T	T	G	T	T	T	G	T	T	G	T
CcLG02	*C.cajan_07124*	27 861 114	G	G	A	G	G	G	A	G	G	A	G
CcLG11	*C.cajan_02962*	32 606 065	T	T	C	T	T	T	C	T	T	C	T
CcLG11	*C.cajan_03203*	35 228 097	C	C	A	C	C	C	A	C	C	A	C
SMD associated nsSNPs													
CcLG02	*C.cajan_07067*	27 324 239	T	T	G	T	T	G	T	T	T	T	G
CcLG02	*C.cajan_07067*	27 324 261	T	T	G	T	T	G	T	T	T	T	G
CcLG08	*C.cajan_15535*	2 014 125	C	C	G	C	C	G	C	C	C	C	G
CcLG11	*C.cajan_01839*	19 958 148	A	A	C	A	A	C	A	A	A	A	C

*R: resistant genotype.

†R-bulk: resistant bulk for FW and SMD.

‡S-bulk: susceptible bulk for FW and SMD.

§HS: highly susceptible genotype.

Note: genotypes were categorized in different groups based on per cent disease incidence (PDI): resistant (0%–9.99% of PDI), moderately resistance (10%–19.99% of PDI), susceptible (20%–40% of PDI) and highly susceptible (>40% of PDI).

were generated for six additional genotypes (Table 1). Re-sequencing data from above-mentioned six genotypes and raw sequence reads from draft genome assembly (ICPL 87119) were used to test the association of nsSNPs with the targeted traits. For the validation of FW associated SNPs, two resistant (ICPL 87119, HPL 24) and four susceptible (ICPL 85063, ICPL 332, ICPL 87, ICPL 88039) genotypes were added. In the case of SMD, re-sequencing data from two resistant (ICPL 87119, ICPL 85063) and one susceptible (ICPL 332) genotypes were added for the validation of SMD-associated SNPs. In brief, sequence data of 11 genotypes (six resistant: ICPL 20096, ICPL 99050, ICPL 20097, ICPL 8863, ICPL 87119 and HPL 24; five susceptible: ICPL 85063, ICPL 332, ICPB 2049, ICPL 87 and ICPL 88039) were used to test and validate the association of four nsSNPs for FW resistance. T to G substitution for the gene *C.cajan_07078* and G to A substitution for the gene *C.cajan_07124* were observed in all the resistant and susceptible genotypes, respectively (Table S5). Similarly, T (resistant genotypes) to C (susceptible genotypes) and C (resistant genotypes) to A (susceptible genotypes) substitution were observed for the other two candidate genes, *C.cajan_02962* and *C.cajan_03203,* respectively (Table S5). This analysis unequivocally provided four SNPs in four different candidate genes associated with FW resistance.

Likewise, association of four SNPs targeting three candidate genes (*C.cajan_07067, C.cajan_15535* and *C.cajan_01839*) with SMD was also validated using re-sequencing data of eight genotypes (six resistant: ICPL 20096, ICPL 99050, ICPL 20097, ICPB 2049, ICPL 87119 and ICPL 85063; two susceptible: ICPL 8863 and ICPL 332). SNPs at positions 27 324 239 bp and 27 324 261 bp on CcLG02 targeting the same candidate gene, *C.cajan_07067,* showed T to G substitution for both the SNPs in all the resistant and susceptible genotypes (Table S6). Analysis of SNP for the genes *C.cajan_15535* and *C.cajan_01839* revealed C to G and A to C substitution, respectively, for resistant and susceptible genotypes. These results clearly suggested association of four SNPs in the three candidate genes with SMD resistance.

Functional annotation of candidate genes

The candidate gene *C.cajan_07067* on CcLG02 has two nonsynonymous substitutions at 27 324 239 bp and 27 324 261 bp positions. The nsSNP at 27 324 239 bp had T (in resistant genotypes) to G (in susceptible genotypes) substitution, which leads to change in amino acid from isoleucine (ATT) to methionine (ATG). Another nsSNP at 27 324 261 bp position had a similar T (in resistant genotypes) to G (in susceptible genotypes) substitution but codes for different amino acid serine (TCC) to alanine (GCC). Functional annotation of this candidate gene reveals its role in serine–threonine protein kinase. Similarly, another candidate gene *C.cajan_07078* on CcLG02 had T (in resistant genotypes) to G substitution (in susceptible genotypes), which leads to change the codon from ATG (methionine) to CTG (leucine). The functional annotation of this candidate gene showed similarity to PHD finger protein. *C.cajan_07124* is a third candidate gene on CcLG02 and had G (in resistant genotype) to A (in susceptible genotype) substitution. This substitution leads to change in amino acid from glycine (GGC) to serine (AGC). This gene showed functional similarity to rRNA-processing protein. The candidate gene *C.cajan_15535* on CcLG08 has C (in resistant genotypes) to G (in susceptible genotypes) substitution, which leads to a silent change from CAA (glutamine) to GAA (glutamine). This gene is functionally characterized as copia-like retrotransposable.

The chromosome CcLG11 had three SNP positions 19 958 148 bp, 32 606 065 bp and 35 228 097 bp in three candidate genes *C.cajan_01839, C.cajan_02962* and *C.cajan_03203, respectively*. The candidate gene *C.cajan_01839* had A (in resistant genotypes) to C (in susceptible genotypes) substitution resulting effect on exon 1 and leads to change into

the codon from TAC (tyrosine) to TCC (serine). Functional characterization of *C.cajan_01839* reveals their role in serine–threonine protein phosphatase an important candidate gene for defence mechanism. At SNP position 32 606 065 bp candidate gene, *C.cajan_02962* had T (in resistant genotypes) to C (in susceptible genotypes) substitution, which leads to change in the amino acid from proline (CCA) to glutamine (CAA) and characterized as *NADH* dehydrogenase. In the same way, another candidate gene *C.cajan_03203* had C (in resistant genotypes) to A (in susceptible genotypes) substitution. The change in nucleotide substitution led to conversion in amino acid from leucine (TTG) to phenylalanine (TTT) and characterized as retrovirus-like polyprotein. Detailed functional annotation of these seven candidate genes is presented in Table S7.

In silico structural analysis of the candidate genes

To understand the structural variation in the candidate genes, nonsynonymous SNPs substitution effects were calculated. For each nsSNPs, the mutation effect was calculated based on proven score value (cut-off = −2.50). As a result, of seven nsSNPs, five mutations did not show any deleterious effect in the protein structure analysis. Therefore, the five genes causing no changes in protein structure were not selected for further study. The remaining two mutations targeting candidate genes, namely *C.cajan_01839* (serine–threonine protein phosphatase) (Figure S3) and *C.cajan_03203* (retrovirus-like polyprotein) (Figure S4), showed deleterious effect on the protein structure with proven score value of −8.56 and −3.69, respectively (Table S7). Proteins 3D structure of the two genes was modelled with >90% accuracy to analyse conformational changes in the translated proteins. In the case of *C.cajan_01839*, secondary protein structure revealed 1% variation in β-sheet model of the resistant and susceptible genotypes. For another gene *C.cajan_03203,* a variation of 32% for resistant and 31% for susceptible genotypes was observed in the α-helix secondary protein structure.

Expression profiling of candidate genes

Above-mentioned approaches provide converging evidences about causativeness of the *C.cajan_03203* gene with FW and the *C.cajan_01839* gene with SMD. For confirming this at functional level, quantitative real-time PCR (qRT-PCR) for these genes was undertaken on root (for FW) and leaf (for SMD) tissues of the resistant (ICPL 20096) and susceptible (ICPL 332) genotypes for both FW and SMD. The *C.cajan_03203* gene showed 3.45-fold down-regulation in root tissues in the FW-susceptible genotype as compared to 0.50-fold down-regulation in the FW-resistant genotype. Similarly, the *C.cajan_01839* gene showed 2.32-fold up-regulation in leaf tissues in the SMD-susceptible genotype as compared to 0.83-fold up-regulation in the SMD-resistant genotype (Figure S5).

Discussion

Conventional methods of trait mapping required genotyping of all the individuals of the developed mapping population. This process, however, is laborious, time-consuming and costly. Additionally, due to low level of polymorphism in some crop species like, pigeonpea and groundnut, identification of polymorphic markers is another challenging task (Pandey *et al.*, 2011; Saxena *et al.*, 2010b). To overcome these issues, in the recent past, NGS-based technologies have been successfully utilized in trait mapping (see Varshney *et al.*, 2014). Recently sequenced pigeonpea genome opened the new avenues to enable the NGS-based breeding for rapid trait mapping similar to other crops (Varshney *et al.*, 2012a). Next-generation sequencing technologies can generate large number of short reads in less time, and with the help of powerful bioinformatics tools, it is possible to assemble the reads for variant (SNPs) calling between and among genotypes. This approach can be useful for identification of higher number of genomewide SNPs, which can be used in number of applications such as in trait mapping, MAS, etc. (Varshney *et al.*, 2009). In the near future, NGS-derived WGS-based approach is expected to be the approach for trait mapping. In this study, we have applied two WGRS-based methods, that is Seq-BSA and nonsynonymous SNPs substitution to find out the candidate genes for two most dreaded diseases namely FW and SMD for enabling genomics-assisted breeding.

To find out the SNPs associated with the trait of interest, we used the concept of traditional BSA approach as proposed by Michelmore *et al.* (1991). This method has been used to map large number of simply inherited traits (Semagn *et al.*, 2010). Due to continuous advances and reduction in cost of sequencing technologies, we have used BSA approach by sequencing the extreme bulks instead of screening with molecular markers. We constituted and sequenced two pools of 16 plants each for resistance and susceptible to these two diseases and referred the approach as Seq-BSA. Accordingly, the probability (as calculated based on Michelmore *et al.*, 1991; $2\ (1 - [1/4]^{16})(1/4)^{16}$) of an unlinked locus being polymorphic between bulks of 16 such individuals would be equivalent to 4.65×10^{-10}. Consequently, on the basis of SNP index analysis, we have identified the SNPs that were monomorphic for resistant parent and resistant bulk (SNP index = 0) but entirely different in susceptible bulk (SNP index = 1) with read depth of ≥7. This approach directly reduced the number of SNPs from 4139 to seven (0.0016%) (with SNP index 1 and 0). Out of seven SNPs identified, only one SNP was present in the genic region (*C.cajan_01839*) while the other SNPs were present in the nongenic region. Our findings are similar to the previous studies in which NGS-based pooled analysis does not allow the identification of direct candidate genes but provides information on putative candidate genes in form of associated SNPs (Hartwig *et al.*, 2012).

As the parental genotypes are segregating for two traits, it was difficult to associate the identified SNPs with one or both (FW, SMD) diseases. Therefore, nsSNP substitution-based approach was utilized. This approach was successfully utilized for mapping drought tolerance in maize (Xu *et al.*, 2014) and sheath blight resistance in rice (Silva *et al.*, 2012). This approach identified 8362 nsSNPs between the R- and S-bulks which are too far to be associated with the target trait. In the present study, we combined both Seq-BSA and nsSNP approach that narrowed down putative eight SNPs in seven candidate genes on three different chromosomes (CcLG02, CcLG08 and CcLG11) of pigeonpea. However, we did not identify any nsSNP within flanking regions of the SNPs identified through Seq-BSA on CcLG07 and CcLG10. This may be attributed to the low sequencing depth in these regions. These speculations, however, need to be confirmed by re-sequencing the particular regions at higher depth.

For classification of identified genes in terms of their association with resistance to FW and SMD, detailed haplotype analysis in diverse set of lines, *In silico* structural (protein) analysis and transcript profiling finally provided causativeness of *C.cajan_03203* gene to FW and *C.cajan_01839* gene to SMD. For instance, the single nucleotide mutation in these genes was

predicted with deleterious effects for translation at the protein level that was also reflected in the form of conformational change in the secondary structure of these genes. qRT-PCR-based expression studies also showed a similar expression in the susceptible genotype (ICPL 332). The gene *C.cajan_03203* codes for a retrovirus-like polyprotein, which is known to be involved in plant defence against pathogens (Grandbastien, 2014). This gene showed >2-fold up-regulation in the susceptible genotype carrying the mutation as compared to the resistant genotype. Retrovirus-like polyprotein are the mobile genetic elements which can replicate and transpose from one position in the genome to the other position either by RNA intermediate (Class-I) driven reverse transcription or by direct transposition (Class-II) (Grandbastien, 2014). Many of the plant retrotransposons reported to date are transcriptionally activated by various abiotic and biotic factors (Grandbastien, 1998). Fungal pathogens like *Trichoderma viride* and *Cladosporium fulvum* or inoculation with various viral and bacterial pathogens have been shown to activate the retrotransposons (Mhiri *et al.*, 1997; Pouteau *et al.*, 1994).

While the gene *C.cajan_01839*, showing causativeness with SMD, has been annotated as serine–threonine protein phosphatase, which is known to be involved in the regulation of specific signal transduction cascades. The mutation in this gene also predicted to have a deleterious effect on the susceptible genotype. This SNP may also be responsible for showing significant down-regulation of the gene in the susceptible genotype and could possibly lead to a defect in the response to SMD. Serine–threonine protein phosphatase is the known principal classes of plant defence genes (Farkas *et al.*, 2007). This gene has been shown to play an important role in defence mechanism of Tobacco mosaic virus resistance (Dunigan and Madlender, 1995).

It is evident from this study that NGS-based approaches do increase not only the precision and power but also saves time in the identification of candidate genes for targeted traits as compared to the conventional mapping methods. We anticipate the accelerated use of NGS-based mapping strategies in all those species, where draft genome sequence has become available. The present study identifies one gene (*C.cajan_03203*) associated with FW and one gene (*C.cajan_01839*) with SMD in pigeonpea. The diagnostic SNPs in these genes can be assayed in some cost-effective marker platform such as CAPS (cleaved amplified polymorphic sequences) or KASP Assay for use in the MAS. Molecular breeding using such genes will help the development of superior lines with enhanced resistance to FW and SMD that will eventually enhance crop productivity of pigeonpea.

Experimental procedures

Plant materials

A total of 11 pigeonpea genotypes were selected based on their FW and SMD responses identified in our previous experiments (Saxena *et al.*, 2010a) (Table 1). Among the selected genotypes, ICPL 20096, ICPL 99050, ICPL 20097, ICPL 8863, ICPL 87119 and HPL 24 were resistant to FW, and ICPL 20096, ICPL 99050, ICPL 20097, ICPB 2049, ICPL 87119, ICPL 85063 were resistant to SMD. Furthermore, ICPL 85063, ICPL 332, ICPB 2049 ICPL 87 and ICPL 88039 were susceptible to FW, and ICP 8863 and ICPL 332 were susceptible to SMD. Two genotypes ICPL 20096 (FW and SMD resistant) and ICPL 332 (FW and SMD susceptible) with contrasting phenotypes were crossed, and the confirmed F_1s were selfed through single-seed descent method. Finally, a total of 188 F_7 recombinant inbred lines (RILs) were produced.

Phenotyping for FW and SMD resistance

The 11 contrasting genotypes for FW and SMD resistance used in the present study were phenotyped in sick plot nursery at Patancheru (Telangana State, India) during crop season 2012–2013 and 2013–2014. Similarly, the RILs (ICPL 20096× ICPL 332) along with contrasting parents were sown in the sick plot nursery Patancheru and Gulbarga (Karnataka State, India) for FW and Patancheru and Tandur (Telangana State, India) for SMD during crop season 2012–2013 and 2013–2014 in three replications using randomized complete block design (RCBD). The experimental plots were four metres long with row to row spacing of 75 cm and a 20 spacing cm between plants. For satisfactory evaluation, the selection of FW-sick plots was on the basis of disease incidence seen every year and for SMD 'Leaf Stapling Technique' (Nene and Reddy, 1976) at two leaf stages was followed. The observations for FW and SMD incidence were recorded at 30 and 90 days after sowing (DAS), and susceptible and resistant RILs were identified using the scale described by Singh *et al.* (2003). Based on the per cent disease incidence (PDI) score, the RILs and their parental lines were classified into four categories (i) resistant (0%–9.99% of PDI), (ii) moderately resistance (10%–19.99% of PDI), (iii) susceptible (20%–40% of PDI), and (iv) highly susceptible (>40% of PDI).

Construction of sequencing libraries and Illumina sequencing

From 188 RILs, a total of 32 RILs (16 susceptible and 16 resistant) and 10 genotypes (ICPL 20096, ICP 8863, ICPB 2049, ICPL 20097, ICPL 99050, ICPL 332, HPL 24, ICP 85063, ICPL 87 and ICPL 88039) were selected for sequencing. The sequencing data of one genotype namely, ICPL 87119, were used from the already published genome (Varshney *et al.*, 2012a). Genomic DNA was isolated from two to three young leaves from selected RILs and 10 genotypes using NucleoSpin Plant II kit (Macherey-Nagel, Düren, Germany). Two DNA pools, one resistant bulk (R-bulk) and one susceptible bulk (S-bulk), were prepared by mixing equimolar concentration of DNA samples from resistant and susceptible RILs.

The Illumina libraries for these two bulks and 10 genotypes were prepared using TruSeq DNA sample Prep kit LT (set A) FC-121-2001. Two microgram DNA from each sample was sheared using Bioruptor® NGS (Diogenode, Liege, Belgium), end repaired and adapter ligated. Size selection of libraries was performed using 2% agarose gel to obtain a target insert size of 500–600 bp and purified for further analysis. Further, the libraries were enriched using adaptor compatible PCR primers. The size distribution of amplified DNA libraries was checked on an Agilent Technologies 2100 bioanalyzer using a High Sensitivity chip (Agilent Technologies, Palo Alto, CA, USA). The DNA libraries were sequenced on Illumina MiSeq platform using MiSeq Reagent Kit v2 (500 cycles) (Illumina Inc., San Diego, CA, USA) to generate 250 base paired-end reads.

Alignment of short reads of bulks for Seq-BSA

QTL-seq pipeline (http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap, developed by Iwate Biotechnology Research Center, Japan) was used for calculating SNP indices. Briefly, the cleaned reads of resistant parent (ICPL 20096) were first aligned to the reference genome (Asha) using inbuilt BWA aligner (Li *et al.*, 2009). Coval was used for postprocessing and filtering of the alignment files (Kosugi *et al.*, 2013). The variants called for the resistant parent were then used to develop reference-guided

assembly of the resistant parent (ICPL 20096; resistant parent) by substituting the bases with confidence variants calls in the genome. The reads from R- and S-bulks were then aligned, and variants were called for both the bulks against the developed assembly.

SNP index was calculated at each SNP position for both the bulks as suggested by Abe *et al.* (2012) using the formula:

$$\text{SNP index (at a position)} = \text{Count of alternate base} / \text{Count of reads aligned}$$

The SNPs with read depth <7 in both the bulks and SNP index <0.3 in either of the bulks were filtered out and SNPs with homozygous alleles in both the bulks were used for ΔSNP index calculation using formula:

$$\Delta\text{SNP index} = \text{SNP index in R-bulk} - \text{SNP index in S-bulk}$$

Only, SNP positions with ΔSNP index = −1 (i.e. the allele called in R-bulk was same as that of resistant parent while contrastingly different in S-bulk) were considered as the causal SNPs responsible for the trait of interest. Additionally, chi-square analysis was performed for both the bulks based on the read depth to test the level of significance of associated SNPs at $P < 0.01$.

Identification of nonsynonymous SNPs

The WGRS data of four parents, namely ICPL 20097 (R-FW and R-SMD) and ICP 8863 (R-FW and S-SMD), ICPB 2049 (S-FW and R-SMD) and ICPL 99050 (R-FW and R-SMD), segregating for FW and SMD along with R- and S-bulks were used for nsSNPs identification. The data were aligned against the reference genome of Asha (Varshney *et al.*, 2012a) using Bowtie 2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). The BAM files thus obtained were used for SNP identification using Samtools 1.0 (Li *et al.*, 2009). The SNPs obtained were annotated using SnpEff tool (http://snpeff.sourceforge.net/), and nsSNPs were subsequently identified using the stringent criteria as described in Silva *et al.* (2012). Further, nsSNPs present in 2-Mb region flanking the SNP positions identified from Seq-BSA study were selected, and the functions of associated genes were predicted by searching the respective protein sequences against nonredundant (nr) database using BLASTP program implemented in Blast2GO software (Conesa *et al.*, 2005). The results from both the approaches were combined to identify disease-specific nsSNPs.

qRT-PCR for expression profiling

One resistant (ICPL 20096) and one susceptible (ICPL 332) genotypes for both FW and SMD, respectively, were used to validate the functionality of two putative candidate genes. Primers were designed for each candidate genes using Primer 3 software for qRT-PCR experiment using standard criteria (Rosen and Skaletsky, 2000). The list of primers used in the qRT-PCR analysis is provided in Table S8.

FW and SMD stresses were imposed on 10 days old seedlings of ICPL 20096 and ICPL 332 grown in two sets for each stress. Root dip inoculation (FW) and leaf staple techniques (SMD) were followed for stress imposition under glasshouse conditions, and tissues were harvested after seven days of stress. Total RNA was isolated from roots (FW) and leaves (SMD) using XcelGen Plant RNA Mini Kit (Xcelris Genomics, Gujarat, India), while cDNA was synthesized using SuperScript® III First-Strand Synthesis SuperMix (Invitrogen, Life Technologies, Thermo Fisher Scientific Corporation, Waltham, MA, USA) for qRT-PCR. The qRT-PCR reactions were performed using SYBR green master mix in 96-well plates with two technical replicates and three biological replicates. The qRT-PCR reaction was performed as mentioned previously (Mir *et al.*, 2014). The housekeeping gene *Actin* was used as an endogenous control to normalize the variations in the cDNA samples. The data were compiled from the mean C_t values of all the biological replicates after normalizing with the C_t values of the endogenous control. The relative transcriptional level in terms of fold change was calculated using the $2^{-\Delta\Delta Ct}$ method (Livaka and Schmittgen, 2001). Tukey's post hoc multiple comparison test using SPSS (version 16.0; SPSS Inc., Chicago, IL, USA) was used for data analysis at $P < 0.05$ to present significant values statistically. The different and similar letters were considered as statistically nonsignificant.

Protein structure analysis

Phyre 2 (Protein Homology/analogY Recognition Engine v2.0 server (www.sbg.bio.ic.ac.uk/phyre2) was used for 3D modelling structures of proteins. Phyre 2 uses the alignment of hidden Markov models via HH search to significantly improve the accuracy of alignment to known 3D structure models. It also incorporates an *ab initio* folding simulation called Poing to model regions of proteins with no detectable homology (Kelley and Sternberg, 2009). Best models were selected based on superfamilies, confidence key, coverage and amino acid identities.

Acknowledgements

The work was supported by the United States Agency for International Development (USAID) – India Mission and Department of Agriculture and Cooperation, Ministry of Agriculture, Government of India. This work has been undertaken as part of the CGIAR Research Program on Grain Legumes. ICRISAT is a member of CGIAR Consortium.

Conflict of interest

The author(s) declare that they have no competing interests.

References

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., Matsumura, H., Yoshida, K., Mitsuoka, C. and Tamiru, M. (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* **30**, 174–178.

Austin, R.S., Vidaurre, D., Stamatiou, G., Breit, R., Provart, N.J., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang, P.W., McCourt, P. and Guttman, D.S. (2011) Next-generation mapping of Arabidopsis genes. *Plant J.* **67**, 715–725.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization, and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Dunigan, D.D. and Madlender, J.C. (1995) Serine/threonine protein phosphatase is required for tobacco mosaic virus mediated cell death. *Virology*, **207**, 460–466.

Farkas, I., Dombradi, V., Miskei, M., Szabados, L. and Koncz, C. (2007) Arabidopsis PPP family of serine/threonine phosphatases. *Trends Plant Sci.* **12**, 169–176.

Fekih, R., Takagi, H., Tamiru, M., Abe, A., Natsume, S., Yaegashi, H., Sharma, S., Sharma, S., Kanzaki, H., Matsumura, H., Saitoh, H., Mitsuoka, C., Utsushi, H., Uemura, A., Kanzaki, E., Kosugi, S., Yoshida, K., Cano, L., Kamoun, S.

and Terauchi, R. (2013) MutMap+: genetic mapping and mutant identification without crossing in rice. *PLoS ONE*, **8**, e68529.

Gnanesh, B.N., Ganapathy, K.N., Ajay, B.C. and Byregowda, M. (2011) Inheritance of sterility mosaic disease resistance to Bangalore and Patancheru isolates in pigeonpea (*Cajanus cajan* (L.) Millsp.). *Electron J. Plant Breed.* **2**, 218–223.

Grandbastien, M.A. (1998) Activation of plant retrotransposons under stress conditions. *Trends Plant Sci.* **3**, 1360–1385.

Grandbastien, M.A. (2014) LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim. Biophys. Acta*, **1894**, 403–416.

Hartwig, B., James, G.V., Konrad, K., Schneeberger, K. and Turck, F. (2012) Fast isogenic mapping-by-sequencing of ethyl methane sulfonate-induced mutant bulks. *Plant Physiol.* **160**, 591–600.

Kelley, L.A. and Sternberg, M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–371.

Kosugi, S., Natsume, S., Yoshida, K., MacLean, D., Cano, L., Kamoun, S. and Terauchi, R. (2013) Coval: improving alignment quality and variant calling accuracy for next-generation sequencing data. *PLoS ONE*, **8**, e75402.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) 1000 genome project data processing subgroup the sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lindner, H., Raissig, M.T., Sailer, C., Shimosato-Asano, H., Bruggmann, R. and Grossniklaus, U. (2012) SNP-ratio mapping (SRM): Identifying lethal alleles and mutations in complex genetic backgrounds by next-generation sequencing. *Genetics*, **191**, 1381–1386.

Livaka, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C (T)) method. *Methods*, **25**, 402–408.

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., Sun, J., Zhang, Z., Weng, Y. and Huang, S. (2014) QTL-seq identifies an early flowering QTL located near *Flowering Locus T* in cucumber. *Theor. Appl. Genet.* **127**, 1491–1499.

Mhiri, C., Morel, J.B., Vernhettes, S., Casacuberta, J.M., Lucas, H. and Grandbastien, M.A. (1997) The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. *Plant Mol. Biol.* **33**, 257–266.

Michelmore, R.W., Paran, I. and 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. *Proc. Natl Acad. Sci. USA*, **88**, 9828–9832.

Mir, R.R., Kudapa, H., Srikanth, S., Saxena, R.K., Sharma, A., Azam, S., Saxena, K., Varma Penmetsa, R. and Varshney, R.K. (2014) Candidate gene analysis for determinacy in pigeonpea (*Cajanus spp.*). *Theor. Appl. Genet.* **127**, 2663–2678.

Mokry, M., Nijman, I., Van Dijken, A., Benjamins, R., Heidstra, R., Scheres, B. and Cuppen, E. (2011) Identification of factors required for meristem function in Arabidopsis using a novel next generation sequencing fast forward genetics approach. *BMC Genom.* **12**, 256.

Nene, Y.L. and Reddy, M.V. (1976) A new technique to screen pigeonpea for resistance to sterility mosaic. *Trop. Grain Legume Bull.* **5**, 23.

Pandey, M., Gautami, B., Jayakumar, T., Sriswathi, M., Upahyaya, H.D., Gowda, M.V.C., Radhakrishnan, T., Bertioli, D.J., Knapp, S.J., Cook, D.R. and Varshney, R.K. (2011) Highly informative genic and genomic SSR markers to facilitate molecular breeding in cultivated groundnut (*Arachis hypogaea*). *Plant Breed.* **131**, 139–147.

Pouteau, S., Grandbastien, M.A. and Boccara, M. (1994) Microbial elicitors of plant defence response activate transcription of a retrotransposon. *Plant J.* **5**, 532–545.

Reddy, M.V., Raju, T.N., Sharma, S.B., Nene, Y.L., McDonald, D., Pande, S. and Sharma, M. (2012) *Handbook of Pigeonpea Diseases (Revised)*. Information Bulletin No. 42. ICRISAT, Patancheru, pp 64.

Rosen, S. and Skaletsky, H.J. (2000) Primer 3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386.

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

Saxena, R.K., Saxena, K.B., Kumar, R.V., Hoisington, D.A. and Varshney, R.K. (2010a) Simple sequence repeat-based diversity in elite pigeonpea genotypes for developing mapping populations to map resistance to *Fusarium* wilt and sterility mosaic disease. *Plant Breed.* **129**, 135–241.

Saxena, R.K., Prathima, C., Saxena, K.B., Hoisington, D.A., Singh, N.K. and Varshney, R.K. (2010b) Novel SSR markers for polymorphism detection in pigeonpea (*Cajanus* spp.). *Plant Breed.* **129**, 142–148.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.E., Weigel, D. and Andersen, S.U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods*, **6**, 550–551.

Semagn, K., Bjornstad, A. and Xu, Y. (2010) The genetic dissection of quantitative traits in crops. *Electronic J. Biotechnol.* **13**, 1–45.

Silva, J., Scheffler, B., Sanabria, Y., De Guzman, C., Galam, D., Farmer, A., Woodward, J., May, G. and 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, I.P., Vishwadhar, and Dua, R.P. (2003) Inheritance of resistance to sterility mosaic in pigeonpea (*Cajanus cajan*). *Indian J. Agric. Sci.* **73**, 414–417.

Takagi, H., Uemura, A., Yaegashi, H., Tamiru, M., Abe, A., Mitsuoka, C., Utsushi, H., Natsume, S., Kanzaki, H., Matsumura, H., Saitoh, H., Yoshida, K., Cano, L.M., Kamoun, S. and Terauchi, R. (2013a) MutMap-Gap: whole-genome resequencing of mutant F2 progeny bulk combined with *de novo* assembly of gap regions identifies the rice blast resistance gene *Pii*. *New Phytol.* **200**, 276–283.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M. and Takuno, S. (2013b) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.* **74**, 174–183.

Uchida, N., Sakamoto, T., Kurata, T. and Tasaka, M. (2011) Identification of EMS-induced causal mutations in a non-reference *Arabidopsis thaliana* accession by whole genome sequencing. *Plant Cell Physiol.* **52**, 716–722.

Varshney, R.K., Nayak, S.N., May, G.D. and Jackson, S.A. (2009) Next generation sequencing technologies and their application for crop genetics and breeding. *Trends Biotechnol.* **27**, 522–530.

Varshney, R.K., Chen, W., Li, Y., Bharti, A.K., Saxena, R.K., Schlueter, J.A., Donoghue, M.T., Azam, S., Fan, G., Whaley, A.M., Farmer, A.D., Sheridan, J., Iwata, A., Tuteja, R., Penmetsa, R.V., Wu, W., Upadhyaya, H.D., Yang, S.P., Shah, T., Saxena, K.B., Michael, T., McCombie, W.R., Yang, B., Zhang, G., Yang, H., Wang, J., Spillane, C., Cook, D.R., May, G.D., Xu, X. and Jackson, S.A. (2012a) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89.

Varshney, R.K., Ribaut, J.M., Buckler, E.S., Tuberosa, R., Rafalski, J.A. and Langridge, P. (2012b) Can genomics boost productivity of orphan crops? *Nat. Biotechnol.* **30**, 1172–1176.

Varshney, R.K., Terauchi, R. and McCouch, S.R. (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biol.* **12**, e1001883.

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. and Lu, Y. (2014) Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC Plant Biol.* **14**, 83.

Yang, Z., Huang, D., Tang, W., Zheng, Y., Liang, K., Cutler, A.J. and Wu, W. (2013) Mapping of quantitative trait loci underlying cold tolerance in rice seedlings via high-throughput sequencing of pooled extremes. *PLoS ONE*, **8**, e68433.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 Read depth of R- and S- bulks calculated across the 11 chromosome of pigeonpea.

Figure S2 SNP counts per window across the 11 chromosome of pigeonpea.

Figure S3 Protein sequence and nsSNP effect analysis for *C.cajan_01839* gene.

Figure S4 Protein sequence and nsSNP effect analysis for *C.cajan_03203* gene.

Figure S5 Expression profiling of two genes.

Table S1 Chromosome wise SNPs distribution between resistant (R) and susceptible (S) bulks.
Table S2 Chi-square analysis for selected SNPs.
Table S3 Genome wide identification of nsSNPs between resistant and susceptible genotypes along with their respective bulks.
Table S4 Distribution of nsSNPs among resistant and susceptible genotypes and both the bulks.
Table S5 Validation of nsSNPs to the candidate genes responsive to fusarium wilt resistance.
Table S6 Validation of nsSNPs to the candidate genes responsive to sterility mosaic disease resistance.
Table S7 Non-synonymous SNP substitution effects and expression fold difference for selected associated genes.
Table S8 List of primer pairs used for qRT-PCR analysis of target genes.

