RESEARCH ARTICLE

Population structure and association mapping studies for important agronomic traits in soybean

BHUPENDER KUMAR[1], AKSHAY TALUKDAR[2]*, INDU BALA[3], KHUSHBU VERMA[2], SANJAY KUMAR LAL[2], RAMESH LAL SAPRA[2], B. NAMITA[2], SUBHASH CHANDER[2] and RESHU TIWARI[2]

[1]*Cummings's Laboratory, Directorate of Maize Research, Pusa Campus, New Delhi 110 012, India*
[2]*Indian Agricultural Research Institute, New Delhi 110 012, India*
[3]*Molecular Cytogenetics and Tissue Culture Laboratory, Department of Crop Improvement, CSK Himachal Pradesh Agricultural University, Palampur 176 062, India*

Abstract

The present study was carried out with a set of 96 diverse soybean genotypes with the objectives of analysing the population structure and to identify molecular markers associated with important agronomic traits. Large phenotypic variability was observed for the agronomic traits under study indicating suitability of the genotypes for association studies. The maximum values for plant height, pods per plant, seeds per pod, 100-seed weight and seed yield per plant were approximately two and half to three times more than the minimum values for the genotypes. Seed yield per plant was found to be significantly correlated with pods per plant ($r = 0.77$), 100-seed weight ($r = 0.35$) and days to maturity ($r = 0.23$). The population structure studies depicted the presence of seven subpopulations which nearly corresponded with the source of geographical origin of the genotypes. Linkage disequilibrium (LD) between the linked markers decreased with the increased distance, and a substantial drop in LD decay values was observed between 30 and 35 cM. Genomewide marker-traits association analysis carried out using general linear (GLM) and mixed linear models (MLM) identified six genomic regions (two of them were common in both) on chromosomes 6, 7, 8, 13, 15 and 17, which were found to be significantly associated with various important traits *viz*., plant height, pods per plant, 100-seed weight, plant growth habit, average number of seeds per pod, days to 50% flowering and days to maturity. The phenotypic variation explained by these loci ranged from 6.09 to 13.18% and 4.25 to 9.01% in the GLM and MLM studies, respectively. In conclusion, association mapping (AM) in soybean could be a viable alternative to conventional QTL mapping approach.

[Kumar B., Talukdar A., Bala I., Verma K., Lal S. K., Sapra R. L., Namita B., Chander S. and Tiwari R. 2014 Population structure and association mapping studies for important agronomic traits in soybean. *J. Genet.* **93**, 775–784]

Introduction

Soybean (*Glycine max* (L.) Merr.) is the most important oilseed crop in India. At present, it occupies an area of about 10 million hectares with an average production of 12.28 metric tonnes. To meet its increasing demand in the local and international markets, elevating production through genetic improvement of the cultivars is gaining top priority countrywide. Traditional breeding approaches have contributed significantly towards the improvement of desirable traits in soybean, such as seed protein content (Cober and Voldeng 1999; Chung *et al.* 2003), grain yield (Elmore *et al.* 2001) and seed filling period (Smith and Nelson 1986). However, as most of the agronomically important traits are quantitatively inherited, hence genes responsible for variation of these traits are hard to detect (Neumann *et al.* 2011). The development of new cultivars with desirable component traits could be facilitated by marker-assisted selection (MAS), where the estimation of the positions and effects of quantitative traits loci (QTL) are of prime importance (Stich *et al.* 2008).

In plant genetics, conventional linkage mapping of genes normally involves segregating populations derived from parents with contrasting phenotypes and/or genotypes. This approach has been used in various plant species over the last two decades for tagging and cloning of gene(s) (Price 2006; Holland 2007). However it has relatively low resolution (Mather *et al.* 2004) and is time-consuming as it requires generation of large size mapping population and their evaluation in multiple environments to obtain robust phenotypic data (Jun *et al.* 2008). It rarely offer opportunities to detect

*For correspondence. E-mail: akshay.talukdar1@gmail.com.

Keywords. phenotypic variation; agronomic traits; population structure; linkage disequilibrium; general linear model; mixed linear model.

multiple alleles in a locus. Further, cost of propagating and evaluating the lines is also not less (Doerge 2002). In contrast, the association mapping (AM) approach offers opportunities to exploit genetic variation in natural populations with high-resolution mapping of complex traits (Zhu *et al.* 2008). It relies on the linkage disequilibrium (LD) which is maintained over generations between loci which are genetically linked to one another (Neumann *et al.* 2011). If LD exists between a marker and locus associated with a trait, then specific marker alleles of haplotypes can be associated with phenotypes at a high level of statistical significance (Cardon and Bell 2001). This approach however, is not free from shortcomings. Population structure and genetic relatedness among the individuals may lead to spurious associations. It is, therefore, important to study the structure of the population with respect to individual's membership to the population and their genetic relatedness among pairs of all the individuals used in the study (Pritchard *et al.* 2000). Population structure have been studied in various crop species, namely, maize (Remington *et al.* 2001), rice (Garris *et al.* 2003; Zhang *et al.* 2011), wheat (Kruger *et al.* 2004), barley (Mather *et al.* 2004) and soybean (Jun *et al.* 2008; Wang *et al.* 2008).

In panels with highly divergent individuals and assumed random mating, only polymorphisms with extremely tight linkage of a locus with desirable phenotypic effects are likely to be significantly associated with a given trait (Remington *et al.* 2001). In the last decade, AM has been used for mapping various qualitative and quantitative traits in plants, namely, flowering time in maize (Thornsberry *et al.* 2001), disease resistance in rice (Garris *et al.* 2003; Agrama and Eizenga 2008; Zhao *et al.* 2011), iron deficiency chlorosis and protein content in soybean (Charlson *et al.* 2003, 2005; Wang *et al.* 2008), maysin and chlorogenic acid content in corn (Szalma *et al.* 2005; Sun *et al.* 2010), yield traits in barley (Kraakman *et al.* 2004) and earliness in wheat (Gouis *et al.* 2012). It has been identified as a tool to resolve complex trait variations down to the sequence level (Nordborg and Tavare 2002), and to identify novel mutations causing specific phenotypes (Palaisa *et al.* 2004). Association analysis in 219 accessions of soybean detected 41 single nucleotide polymorphism (SNP)-trait associations, with 20 of these SNPs being associated with seed-size traits and seven with seed-shape traits (Hu *et al.* 2013). Further, AM studies identified five simple sequence repeat (SSR) markers namely, Satt231, Satt411, Satt489, Satt276 and Satt434 significantly associated with photoperiod insensitivity in soybean explaining 15.8, 18.7, 10.1, 9.98 and 11.9% of the trait variations, respectively (Singh *et al.* 2008). AM of salt tolerance and alkaline index in 257 cultivars of soybean with 135 SSR markers identified 83 QTL-by-environment (QE) interactions for salt tolerance index and 86 QE interactions for alkaline index (Zhang *et al.* 2014). However, molecular mapping of agronomic traits viz., plant height, plant growth habit, days to 50% flowering and days to maturity has largely been neglected in the Indian soybean germplasm. Therefore, in this study, AM approach was used to identify molecular markers linked with traits of agronomic importance in Indian and some exotic germplasm collections of soybean.

Materials and methods

Plant materials

A set of 96 soybean genotypes was selected from 500 germplasm lines maintained at Genetics Division, Indian Agricultural Research Institute (IARI), New Delhi, India (table 1). This set of diverse germplasm was used as the association mapping panel (AMP), which comprised of 54 genotypes from India, 30 from Taiwan, 11 from USA and one from New Guinea. Indian genotypes included one indigenous collection (IC) and 53 high yielding varieties of soybean released for cultivation in different states of India.

Field experiments and phenotypic evaluation

Phenotypic evaluation of the AMP was carried out during kharif season of 2008 and 2009 following augmented block design at the experimental field of IARI, New Delhi, India (altitude 228.1 m; 28.38°N; 77.12°E). The four check varieties namely, DS9712, DS9814, SL525 and JS335 were randomized in each block to estimate the block effect. Each genotype was sown in two rows of 4-m length with 45 cm × 10 cm spacing, and recommended package and practices were followed. Data were recorded on five plants, selected randomly from middle portion of the rows. The traits evaluated in this study included plant height (cm), number of pods per plant, number of seeds per pod, 100-seed weight (g), seed yield per plant (g), days to 50% flowering, days to maturity and plant growth habit (determinate/indeterminate). Descriptive statistics and analysis of variance for all the traits were carried out using SPSS software ver. 17.

Genomewide screening

Genomewide screening of the AMP was performed with 121 SSRs markers spanning across the whole soybean genome (on average six markers/chromosome) (table 2). Sequence of the primers, and their map position was similar to Cregan *et al.* (1999), and was downloaded from http://www.soybase.org. Genomic DNA from a pooled sample of five plants was isolated using CTAB method (Saghai-Maroof *et al.* 1984). Purity and quantity of isolated DNA sample was estimated using spectrophotometer. For PCR amplification, the programme was set for initial denaturation at 94°C for 2 min followed by 39 cycles consisting of denaturation, primer annealing and extension at 94, 45–55 and 72°C, respectively for 1 min each. The PCR master-mix of 20 μL contained 50 ng of template DNA, 1 μM each of forward and reverse primers, 1.5 mM dNTPs, 1 U *Taq* polymerase and 1× buffer with $MgCl_2$. It was run in standard thermocycler, Gene Amp® PCR system 9700 (Applied Biosystem, Foster City,

Table 1. List of soybean genotypes used as AMP with their source and subpopulation in which they are grouped during population structure study.

	Genotype	Source	Subpopulation		Genotype	Source	Subpopulation
1	SKAF635	Mandsaur	A	49	VLS 57	Almora	D
2	SKAF2202	Mandsaur	A	50	SL 444	Ludhiana	D
3	SKAF106	Mandsaur	A	51	L 291	Taiwan	D
4	SKAF750-1	Mandsaur	A	52	DS 9712	Delhi	D
5	SKA2008	Mandsaur	A	53	Himso 1598	Palampur	D
6	SKAF415	Mandsaur	A	54	EC 456554	USA	D
7	JS335	Jabalpur	A	55	PKV 25	Akola	D
8	KG83-1A	Kasbe Digraj	B	56	PK 1169	Pantnagar	E
9	IC244409	India	B	57	NRC 1180	Indore	E
10	DS2009	Delhi	B	58	EC 439619	Taiwan	E
11	DS2011	Taiwan	B	59	PK 1223	Pantnagar	E
12	DS9816	Delhi	B	60	PK 1135	Pantnagar	E
13	EC457254	USA	B	61	EC 44303	Taiwan	E
14	EC472211	Taiwan	B	62	EC 439618	Taiwan	E
15	DS2006	Delhi	B	63	PK 1241	Pantnagar	E
16	EC471427	Taiwan	B	64	PK 1225	Pantnagar	E
17	UPSL534	Pantnagar	B	65	DS 9819	Delhi	F
18	EC458354	USA	B	66	DS 9817	Delhi	F
19	MAUS162	Parbhani	B	67	SL 528	Ludhiana	F
20	PK7427-B	Pantnagar	B	68	PS 1392	Pantnagar	F
21	EC456580	USA	B	69	PK 292	Pantnagar	F
22	EC472171	Taiwan	B	70	PK 1041	Pantnagar	F
23	JS(SH) 93-01	Sehore	B	71	SL 633	Ludhiana	F
24	EC456535	USA	B	72	EC 9467	USA	F
25	EC472217	Taiwan	B	73	PS 1394	Pantnagar	F
26	EC456626	USA	B	74	DS 9821	Delhi	F
27	JS96	Jabalpur	B	75	DS 9720	Delhi	F
28	PK1080	Pantnagar	B	76	SL 710	Ludhiana	F
29	EC472239	Taiwan	B	77	EC 471784	Taiwan	F
30	EC472229	Taiwan	B	78	DS 9801	Delhi	F
31	EC472184	Taiwan	B	79	DS 9814	Delhi	F
32	G2144	Taiwan	B	80	DS 9820	Delhi	F
33	EC471276	Taiwan	C	81	EC 456574	USA	F
34	EC456597	USA	C	82	EC 456549	USA	F
35	EC458356	USA	C	83	SL 637	Ludhiana	F
36	EC472220	Taiwan	C	84	EC 472127	Taiwan	G
37	UPSM534	Pantnagar	C	85	EC 472145	Taiwan	G
38	EC472228	Taiwan	C	86	EC 472095	Taiwan	G
39	EC471809	Taiwan	C	87	EC 472101	Taiwan	G
40	EC472218	Taiwan	C	88	EC 472103	Taiwan	G
41	SL525	Ludhiana	D	89	G 2130	Taiwan	G
42	PS1042	Pantnagar	D	90	G 2132	Taiwan	G
43	SL427	Ludhiana	D	91	EC 439617	Taiwan	G
44	SL432	Ludhiana	D	92	EC 472126	Taiwan	G
45	PK1347	Pantnagar	D	93	EC 472118	Taiwan	G
46	MAUS164	Parbhani	D	94	EC 389392	Taiwan	G
47	SL459	Ludhiana	D	95	EC 113397	N. Guinea	G
48	SL46	Ludhiana	D	96	M 135	India	G

Pantnagar, Ludhiana, Delhi, Indore, Akola, Parbhani, Almora, Palampur, Kasbe Digraj, Mandsaur, Sehore and Jabalpur are the soybean research centres located in India. A, B, C, etc. represents different subpopulations identified in this study (subpopulations have shown in different colour in figure 2).

USA). The amplified products were analysed through electrophoresis in metaphore-agarose gel.The alleles were scored as 1 for present and 0 for absent. The size of each allele was also recorded in base pairs (bp) as per the requirement of various analytical packages used. The polymorphism information content (PIC) was determined as described by Senior and Henn (1993) which is given as $PIC = 1 - \sum P_{ij}^2$, where P_{ij} is the frequency of jth allele at ith locus summed across all alleles in the locus. Alleles with frequency of less than 0.20 were considered as rare alleles and such allele representing a particular genotype was considered as a unique allele for that genotype.

Study of population structure

For analysing the population structure and kinship, a subset of 66 randomly inherited polymorphic SSR markers were

Table 2. List of SSR markers used for whole genomewide screening, population structure and kinship analysis.

Chr. no.	LG	SSR marker
1	D1a	**Satt408*** (106.69), Satt407* (99.59), Satt370(60.99), **Satt532*** (49.07), **Satt342*** (48.14), Satt531* (40.87)
2	D1b	Satt_289(131.92), **Satt459*** (118.62), Satt600* (75.41), **Satt290*** (73.35), **Satt266*** (59.61), Satt558* (43.91), Satt698* (38.04)
3	N	**Satt022*** (102.06), **Satt255*** (76.49), **Satt549*** (70.60), **Satt530*** (32.85)
4	C1	**Satt180*** (127.77), Satt524* (120.12), Satt361(75.52), **Satt646*** (70.52), Satt396* (24.11), Sct_186(9.02), **Satt565***(0.00)
5	A1	**Satt211*** (95.96), Satt619*(69.21), **Satt648***(59.18), Satt717*(51.95), **Satt593*** (25.56)
6	C2	**Satt202*** (126.24), **Satt460*** (117.77), Satt643(94.65), Satt170* (70.56), Sat_336(51.84), Satt432* (38.05), **Aw734043*** (4.22)
7	M	Satt336(133.83), **Satt308*** (130.76), Satt618* (111.06), **Satt463*** (50.10), Satt435* (38.94), **Sat_389*** (0.00)
8	A2	Satt378(165.73), **Satt429*** (162.3), **Satt538*** (159.03), **Satt228*** (154.11), **Satt409*** (145.57), Satt377(116.64), Satt707(116.62), Satt119* (92.43), **Sat_199*** (84.09), Satt187* (54.92), Satt177* (36.77)
9	K	**Satt196*** (104.79), **Satt260*** (80.12), Satt499(71.01), Satt240* (52.88), Satt337* (47.38), **Sat_087*** (4.85)
10	O	**Satt581*** (106.63), Satt592(100.38), **Satt331*** (93.37), **Sat_282*** (63.81), Satt347* (42.29), Satt653* (38.09)
11	B1	**Satt453*** (123.96), Satt665* (96.36), **Satt197*** (46.39), **Satt251*** (36.48), **Sat_156*** (35.00), Satt411(30.87)
12	H	**Satt434*** (105.74), Satt142(86.49), **Satt302*** (81.04), **Satt635*** (4.88), Satt666 (0.59),
13	F	Satt522* (119.19), **Satt554*** (111.89), **Satt335*** (77.70), **Satt114*** (63.69)
14	B2	**Satt687*** (113.61), **Satt560*** (97.92), **Satt534*** (87.59), **Satt272*** (71.68), Satt601(67.73), Satt168* (55.20), Satt_287* (31.88), **Satt467*** (17.77)
15	E	**Satt230*** (71.31), **Satt685*** (56.70), **Satt720*** (20.80)
16	J	Sat_**224*** (75.13), Satt620(53.71), Sat_366(52.84), **Satt183*** (42.51), **Satt529*** (41.90), Satt674* (15.95), Satt405* (12.41)
17	D2	**Satt386*** (125.00), **Satt186*** (105.45), GMHSP179* (99.04), Satt301* (93.71), **Satt543*** (88.02), Satt311 (84.62)
18	G	Sct_**187*** (107.11), Satt612* (80.38), **Satt288*** (76.77), **Satt566*** (49.91)
19	L	**Satt373*** (107.24), Satt664(92.66), Satt006* (92.00), Sat_286* (87.42), **Satt481*** (54.57), Satt418(30.93), Sat_405(29.62), Sat_408(1.31)
20	I	Sat_**324*** (84.48), **Satt671*** (72.09), **Satt239*** (36.94), Satt367* (27.98), Satt451 (20.34)

A total of 121 markers were used for whole genome screening of which 97 were polymorphic * used for LD study. A subset of 66 polymorphic markers (in bold) were used for population structure (Q) and kinship (K) analysis. Figures within parentheses are the locations (cM) in the soybean genome map. LG linkage group. Map positions correspond to the composite map of soybean by Cregan *et al*. (1999).

selected from the set of 121 markers (table 2). Average interval of the selected markers was 32 cM. The Q matrix was constructed following the model-based approach as described by Pritchard *et al*. (2000) and Falush *et al.* (2003) which is implemented in the software Structure (Pritchard *et al.* 2000; http://pritch.bsd.uchicago.edu). For analysis of the data, set parameters of the population admixture model and correlated frequency of alleles were considered. The hypothetical subpopulations were considered as K= 2–15 and, the package was run with 3 independent runs for each K. Length of burn-in period and number of iterations were set at 1, 50,000. The genetic distances among the K structure clusters were computed applying the neighbour-joining algorithm to the matrix of allele-frequency divergence among clusters in PHYLIP (phylogeny inference package) ver. 3.6 (Felsenstein 2005). The kinship matrix was generated for MLM study using Tassel3 package.

Analysis of LD and marker-traits associations

For analysing LD between SSR loci, values for r^2 (Hill and Robertson 1968) and D′ (Farnir *et al.* 2000) were calculated using the software package Tassel3 (http://www.maizegenetics.net) with permutations test of 10,000. The pairs of loci were considered to be in significant LD at $P < 0.01$. Associations between markers and traits variation were established through Tassle3 software using GLM with Q (individuals' membership in the population) and MLM with Q+K matrix generated during population studies (Bradbury *et al.* 2007).

Table 3. Summary of the descriptive statistics for the seven phenotypic traits examined during 2008 and 2009.

Trait	Year	Minimum	Maximum	Mean	SD
Plant height (cm)	2008	30.00	85.90	54.78	13.68
	2009	25.20	93.20	52.83	13.95
Pods per plant (no.)	2008	30.45	100.20	55.11	14.89
	2009	23.40	112.80	53.62	17.24
Seed per pod (no.)	2008	2.00	3.00	2.36	0.26
	2009	1.74	3.00	2.30	0.25
100-seed weight (g)	2008	3.40	13.29	8.33	2.00
	2009	4.67	15.23	8.21	1.72
Seed yield per plant (g)	2008	4.85	24.56	10.45	3.67
	2009	4.20	21.63	9.96	3.55
50% flowering (no. of days)	2008	32.00	61.00	43.90	5.28
	2009	34.00	58.00	42.63	4.52
Days to maturity (no. of days)	2008	88.00	124.00	103.56	7.90
	2009	86.00	126.00	100.67	8.042

Results

Phenotypic variation

The descriptive statistics namely, mean, range (minimum and maximum) and standard deviation for the phenotypic traits observed during 2008 and 2009 are given in table 3. Marked variations were recorded among the genotypes for all the traits in both the years of study. Example, days to 50% flowering ranged from 32 (MAUS164) to 62 (EC472229) days during 2008 and 34 (EC471276) to 63 (EC472229) days during 2009. The value of days to maturity ranged from 89 (DS9817) to 124 days (EC472218), and 86 (SL 46) to 126 days (EC472228) during 2008 and 2009, respectively. High range of phenotypic variations observed in the traits can be attributed to the diversity of the genotypes included in the study. Pearson's correlation studies indicated highly significant ($P \leq 0.01$) correlation of pods per plant ($r = 0.77$), 100-seed weight ($r = 0.35$) and days to maturity ($r = 0.23$) with seed yield per plant (table 4). Further, highly significant ($P \leq 0.01$) correlation ($r = 0.32$) was observed between days to maturity and plant growth habit. The analysis of variance (ANOVA) revealed significant genotypic and year effect for most of the traits under study (table 5).

Analysis of population structure

To obtain the appropriate value for K, the pointer used was such that the values of ln Pr $(X|K)$ would stop varying significantly much from its succeeding as compared to its preceding (Pritchard *et al.* 2000; http://pritch.bsd.uchicago.edu). Further, grouping of the genotypes into subpopulations based on their geographical origin was also taken into account while deciding the optimum subpopulation. Considering all these pointers, K = 7 was found to be optimum indicating that the experimental population was composed of seven subpopulations (figure 1). These substructures partially coincided with the geographical origin of the genotypes. The phylogenetic relationship of the substructures has been depicted in figure 2. Both the figures clearly depict that each subgroup is diverse from others. Group 'G', which appears to be the most diverse consisted of the genotypes exclusively of Taiwan origin. The germplasm introduced from Taiwan have been used in breeding programmes of India since long back. Therefore, a few Indian soybean genotypes showed its closeness to Taiwan soybean and was grouped with it. Similarly, involvement of Taiwan soybean in other groups also indicated its gene introgression into soybean genotypes of other places. The group 'A' contained the popular Indian variety

Table 4. Correlation studies between yield related traits examined during 2008 and 2009.

	PdPLT	SPPd	TSWT	SYLDP	50% FLW	DAM	PGH
PHT	0.460**	0.327**	−0.183 ∗ ∗	0.11*	0.256**	0.298**	0.485**
PdPLT		0.1	−0.127	0.770**	0.143*	0.032	0.336**
SPPd			−0.090	0.12*	0.333**	0.348**	0.503**
TSWT				0.348**	−0.126	−0.141*	−0.332**
SYLDP					0.135*	0.23**	0.213*
50% FLW						0.56**	0.485**
DAM							0.32**

PHT, plant height; PdPLT, pods per plant; SPPd, seeds per pod; TSWT, test weight (100-seed weight); SYLDP, seed yield per plant; 50% FLW, 50% flowering; DAM, days to maturity; PGH, plant growth habit (determinate/indeterminate).
** Significant at $P \leq 0.01$; *significant at $P \leq 0.05$.

Table 5. Analysis of variance for yield and related traits based on two year data recorded during 2008 and 2009.

Source	df	PHT	PdPLT	SPPd	TSWT	SYLDP	50% FLW	DAM
G	95	322.862**	480.000**	0.097**	5.663**	27.547**	44.456**	113.258**
B	3	61.136*	55.213	0.04	2.007	11.509*	4.333	2.781
Y	1	65.24*	86.759	0.141*	1.048	18.307*	11.23	73.23*

G, genotype; B, block; Y, year; df, degree of freedom; PHT, plant height; PdPLT, pods per plant; SPPd, seeds per pod; TSWT, test weight (100-seed weight); SYLDP, seed yield per plant; 50% FLW, 50% flowering; DAM, days to maturity.
** Significant at $P \leq 0.01$; *significant at $P \leq 0.05$.

JS335 and its descendents, and appeared to be the least diverse group. Genotypes from USA and Pantnagar clustered together. It indicated their genetic relatedness owing to introduction of soybean from USA and their involvement in breeding at Pantnagar.

Genomewide LD and association study

During whole genome screening of the genotypes, 97 out of 121 SSR markers appeared to be polymorphic (table 2). A total of 286 alleles ranging from two to six with an average of 2.36 alleles per locus were identified in the population. Out of the total alleles, 90 were rare alleles (having frequency < 0.20) (Kumar *et al.* 2014), however were treated as general alleles in this study. The PIC values of the polymorphic markers ranged from 0.06 to 0.75 with an average of 0.32. The LD plot for r^2 and D' values between the markers indicated existence of significant LD between the linked as well as unlinked markers. The values of r^2 between pairs of markers ranged from 0 to 0.35, while D' values ranged from 0 to as high as 1.00. Values of LD between the linked markers decreased with the increase in distance between them. A substantial drop in LD was observed between 30 and 35 cM, suggesting that resolution can be achieved as low as 35 cM level.

The number of significant marker-trait associations was more in GLM than MLM. The GLM based marker-traits association analysis identified five loci, AW734043, Satt463, Satt538, Satt114 and Satt301 (some markers have association to more than one traits) on chromosomes 6, 7, 8, 13 and 17, respectively, which were found to be significantly associated with different traits such as plant height, pods per plant, 100-seed weight, plant growth habit, days to 50% flowering and days to maturity (table 6; figure 3). However, the association studies carried out using MLM identified three loci Satt463, Satt685 and Satt301 (two were common in GLM) on chromosomes 7, 15 and 17, respectively, to be significantly associated with important traits namely, days to maturity, plant height and average number of seeds per pod (table 6). The phenotypic variations demonstrated by these loci ranged from 6.09 to 13.18% and 4.25 to 9.01% for GLM and MLM studies, respectively. One marker Satt685 was found to be significantly associated ($P \leq 0.05$) with average number of seeds per pod in MLM only. The same marker was found to be associated in GLM at $P < 0.2$. However, Satt301 and Satt463 were found to be associated with days to maturity, plant height and plant growth habit in both GLM and MLM (table 6; figure 3).

Discussion

Association mapping is a powerful tool to establish the marker-trait associations. Its applicability as well as resolution power has already been established in a number of crops including soybean. This approach also provides estimates of phenotypic variations demonstrated by each locus. In this study, it detected large variations for various phenotypic traits. Usually, the Indian soybean has a narrow genetic base and is least variable. However, due to inclusion of diverse genotypes including indigenous and exotic collections of germplasm and improved varieties of soybean, wider variability was observed in this study. Govindarao (2010) also reported detection of wider variations in morphological traits through variations in composition of the population under study. Owing to the variability present, the population

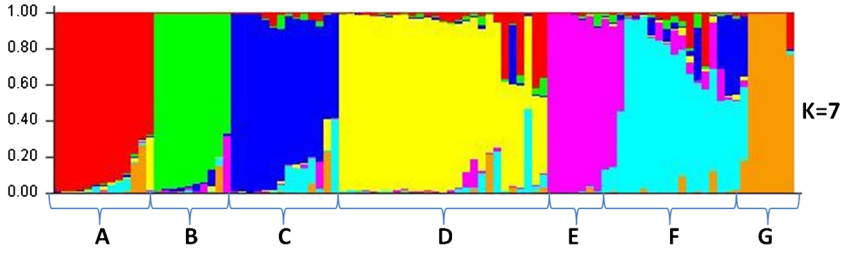

Figure 1. Graphical depiction of the seven subpopulations detected in the population structure study. Partial substructure has been observed in the population corresponding to geographical origin of the genotypes.

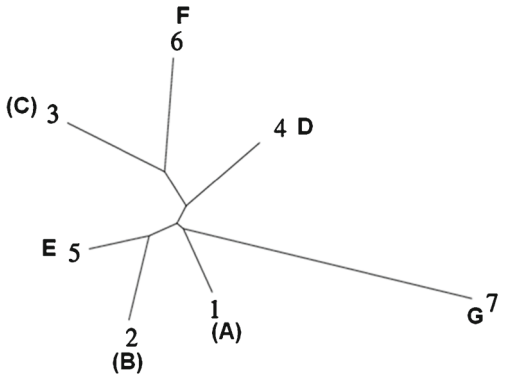

Figure 2. A phylogenetic tree representing the K structure clusters computed by neighbour-joining algorithm in PHYLIP ver. 3.6 (Felsenstein 2005). Genotypes involved in each subpopulation have been mentioned in table 1.

appeared to be a panel suitable for association mapping studies, provided the population structure is studied. Character association studies indicated possible association of a number of traits with yield. Direct selection of such traits e.g., pods per plant, 100-seed weight, etc., would contribute towards genetic enhancement of yield.

Population stratification contributes towards false positive results (Cardon and Bell 2001; Yu *et al.* 2006). The population structure with respect to geographical origin contribute towards pseudoassociations (Maskri *et al.* 2012). Therefore, adequate methods need to be implemented to control the effects of population structure in order to avoid high rate of type I error (Agrama *et al.* 2007). In this study, model-based analysis of population structure was conducted, which indicated the presence of seven subpopulations in the genotypes studied. These subpopulations largely corresponded to the major geographic regions of their origin or collection. The genotypes with prefix 'SKAF' clustered with 'JS335', the most popular soybean variety of India. The two were collected from the same geographical locality i.e., Madhya Pradesh, India. Like JS335, the 'SKAF'-series genotypes were also highly susceptible to yellow mosaic virus (YMV) (Kumar *et al.* 2014), indicating their genetic closeness to 'JS335'. Similarly, Indian genotypes clustered with genotypes from Taiwan and USA indicating that they are genetically close. In fact, Indian soybean gene pool comprises primarily of the germplasm introduced from USA and China (Taiwan), and hence such results are obvious. The overlapping of structures by a number of genotypes indicated that such accessions showed shared ancestry. These accessions probably had a complex breeding history involving intercrossing and introgression between germplasms from diverse backgrounds, followed by strong selection pressure for desirable traits (Mather *et al.* 2004). Wang *et al.* (2008) reported the existence of five subpopulations in the genotypes used for mapping iron deficiency chlorosis (IDC) in soybean. Knowledge of population structure and ancestral background would facilitate selection of parental lines in soybean breeding programmes.

LD which is explained by nonrandom association between polymorphisms at different loci is the basis of association mapping studies (Zhu *et al.* 2008). In this study, LD between linked markers decreased with increased distances, and a substantial drop was noticed between 30 and 35 cM. It, thus, indicated that high mapping resolution is possible to achieve as low as the 35 cM level. Long range LD observed in this study demonstrated the potential for genomewide association mapping with fewer markers in soybean; however such steps are certain to compromise the resolution. Strong LD was found to sustain up to a genetic distance of only ~10 cM. However, in rice strong LD did not decay up to 20–30 cM (Agrama *et al.* 2007). Genomewide LD decay was found to occur within 200–1,500 bp in maize (Tenaillon *et al.* 2001), 3 cM in sugar beet (Kraft *et al.* 2000), 50 cM in sorghum (Hamblin *et al.* 2004), 10–50 cM in barley (Malysheva-Otto *et al.* 2006) and 10–20 cM in durum wheat (Maccaferri *et al.* 2005). Thus, the highly selfpollinated crops including soybean, rice, wheat, etc., maintain LD to a longer genetic distance as compared to cross-pollinated crops.

Identifying genetic variants that underlie complex traits is an important and challenging goal in plant genetics. The association mapping approach presents opportunities

Table 6. Genomewide associations between SSR markers and important agronomic traits studied using GLM and MLM.

Trait	Marker name	Marker position (cM)	Chr. no.	% phenotypic variation (r^2)	
				GLM (Q)	MLM (Q+K)
Days to maturity	Satt114	63.69	13	11**	NS
	Satt463	50.10	7	8.75**	6.20**
	AW734043	4.22	6	8.66**	NS
Days to 50% flowering	AW734043	4.22	6	9.06**	NS
Pods per plant	Satt463	50.10	7	8.32*	NS
Plant height	Satt301	93.71	17	10.85**	6.23**
	Satt463	50.10	7	7.31*	4.25*
Plant growth habit	Satt301	93.71	17	13.18**	8.43**
100-seed weight	Satt538	159.63	8	6.09*	NS
Average no. of seed per pod	Satt685	56.70	15	NS	9.01**

** Significant at $P \leq 0.01$, *significant at $P \leq 0.05$.

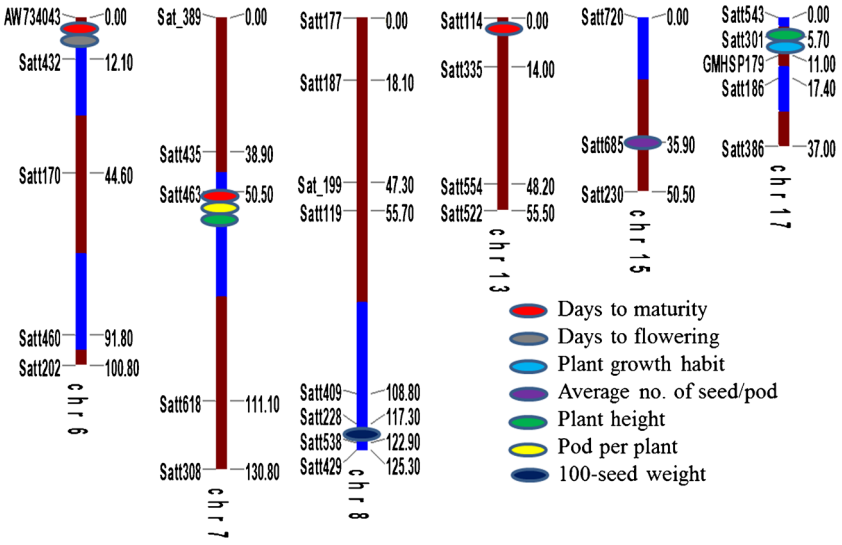

Figure 3. Localization of six putative loci, AW734043, Satt463, Satt538, Satt114, Satt685 and Satt301 on their respective chromosomes 6, 7, 8, 13, 15 and 17 which were found to be associated with seven important agronomic traits.

to exploit the genetic variation in natural populations (Zhu *et al.* 2008). However, the false discoveries in association mapping are a major concern and can be partially attributed to spurious associations caused by population structure and unequal relatedness among individuals. Two major approaches namely, GLM and MLM are used for studying marker-trait association. Generally, the number of significant marker-trait associations detected by GLM is much higher than in MLM (Neumann *et al.* 2011). The GLM-based studies on marker-trait associations consider only Q matrix generated during the study of population structure, however MLM simultaneously accounts for both population structure as well as kinship (genetic relatedness among individuals), and hence more reliable.

In this study, Satt463 was found to be significantly associated with days to maturity and plant height (both in GLM and MLM), and pods per plant (in GLM only). Similarly Jun *et al.* (2008) identified the genomic region close to Satt463 as being responsible for maturity in soybean. Since days to maturity and pods per plant are significantly correlated with seed yield per plant, hence linked marker Satt463 may be used to select seed yield per plant. Satt301 was significantly associated with plant height and plant growth habit (determinate/indeterminate) in both GLM and MLM. The plants with indeterminate growth habit are expected to be tall. Overlapping of loci was associated with different trait variants indicate the biological correlation between them. Therefore, plant height and plant growth habit might be located adjacent to each other on the same chromosome. The marker Satt538, located on chromosome 8, was found to be significantly associated with 100-seed weight in GLM studies, which in turn is significantly correlated with seed yield per plant. Therefore, Satt538 may be a marker of choice, which can be indirectly used for yield enhancement. Only 40% of the markers associated with various traits in GLM have shown the significant marker-trait association in MLM. Neumann *et al.* (2011) have also found similar types of results in wheat. Further, the phenotypic variation explained by MLM was comparatively lesser than that of GLM. It seems to be more stringent in eliminating the spurious associations than GLM.

In conclusion, it can be said that there are enough genetic diversity in the genotypes under investigation, although seven subpopulations were also present as per the origin or collection of germplasm from different locations. Long range of LD was found to exist which demonstrated the potential for genomewide association mapping studies in soybean. The false discoveries in association mapping are a major concern, therefore extra care should be taken while selecting germplasm materials as well as studying population structure. MLM method was proven to be useful in controlling false associations. This study, thus, reaffirms the usefulness of association mapping approach for mapping of loci underlines the complex traits. The markers identified in this study would be useful in MAS for important traits including yield and its components.

Acknowledgement

First author sincerely acknowledges PG School, IARI, for providing the fellowship during post-graduate study.

References

Agrama H. A. and Eizenga G. C. 2008 Molecular diversity and genome-wide linkage disequilibrium patterns in a worldwide collection of *Oryza sativa* and its wild relatives. *Euphytica* **160**, 339–355.

Agrama H. A., Eizenga G. C. and Yan W. 2007 Association mapping of yield and its components in rice. *Mol. Breed.* **19**, 341–356.

Bradbury P. J., Zhang Z., Kroon D. E., Casstevens T. M., Ramdoss Y. and Buckler E. S. 2007 TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635.

Cardon R. L. and Bell J. I. 2001 Association study designs for complex disease. *Nat. Rev. Genet.* **2**, 91–99.

Charlson D. V., Cianzio S. R. and Shoemaker R. C. 2003 Associating SSR markers with soybean resistance to iron deficiency chlorosis. *J. Plant Nutr.* **26**, 2267–2276.

Charlson D. V., Bailey T. B., Cianzio S. R. and Shoemaker R. C. 2005 Molecular marker Satt481 is associated with iron-deficiency chlorosis resistance in a soybean breeding population. *Crop Sci.* **45**, 2394–2399.

Chung J., Babka H. L., Graef G. L., Staswick P. E., Lee D. J., Cregan P. B. *et al.* 2003 The seed protein, oil, and yield QTL on soybean linkage group I. *Crop Sci.* **43**, 1053–1067.

Cober E. R. and Voldeng H. D. 1999 Developing high-protein, high-yield soybean populations and lines. *Crop Sci.* **40**, 39–42.

Cregan P. B., Jarvik T., Bush A. L., Shoemaker R. C., Lark K. G., Kahler A. L. *et al.* 1999 An integrated genetic linkage map of the soybean genome. *Crop Sci.* **39**, 1464–1490.

Doerge R. W. 2002 Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.* **3**, 43–52.

Elmore R. W., Roeth F. W., Nelson L. A., Shapiro C. A., Klein R. N., Knezevic S. Z. and Martin A. 2001 Glyphosate-resistant soybean cultivar yields compared with sister lines. *Agronomy J.* **93**, 408–412.

Falush D., Stephens M. and Pritchard J. K. 2003 Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

Farnir F., Coppieters W., Arranz J. J, Berzi P., Cambisano N., Grisart B. *et al.* 2000 Extensive genome wide linkage disequilibrium in cattle. *Genome Res.* **10**, 220–227.

Felsenstein J. 2005. PHYLIP (phylogeny inference package) version 3.6, distributed by author. Department of Genome Sciences, University of Washington, Seattle, USA.

Garris A. J., McCouch S. R. and Kresovich S. 2003 Population structure and its effects on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice *Oryza sativa* L. *Genetics* **165**, 759–769.

Gouis J. L., Bordes J., Ravel C., Heumez E., Faure S., Praud S. *et al.* 2012 Genome-wide association analysis to identify chromosomal regions determining components of earliness in wheat. *Theor. Appl. Genet.* **124**, 597–611.

Govindarao C. N. 2010. Characterization of soybean [*Glycine max* (L.) Merr.] varieties through morphological, chemical molecular markers and image analyzer (pp. 50–55). M.Sc. thesis. University of Agricultural Sciences, Dharwad, India.

Hamblin M. T., Mitchell S. E., White G. M., Gallego J., Kukatla R. and Wing R. A. 2004 Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of Sorghum bicolor. *Genetics* **167**, 471–483.

Hill W. G. and Robertson A. 1968 Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231.

Holland J. B. 2007 Genetic architecture of complex traits in plants. *Curr. Opin. Plant. Biol.* **10**, 156–161.

Hu Z., Zhang H., Kan G., Ma D., Zhang D., Shi G. *et al.* 2013 Determination of the genetic architecture of seed size and shape via linkage and association analysis in soybean (*Glycine max* L. Merr.) *Genetica* **141**, 247–254.

Jun T. H., Van K., Kim M. Y., Lee S. H. and Walker D. R. 2008 Association analysis using SSR markers to find QTL for seed protein content in soybean. *Euphytica* **162**, 179–191.

Kraakman A. T. W., Niks R. E., Van den Berg P. M. M. M., Stam P. and Van Eeuwijk F. A. 2004 Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* **168**, 435–446.

Kraft T. M., Hansen M. and Nilsson N. O. 2000 Linkage disequilibrium and fingerprinting in sugar beet. *Theor. Appl. Genet.* **101**, 323–326.

Kruger S. A., Able J. A., Chalmers K. J. and Langridge P. 2004. Linkage disequilibrium analysis of hexaploid wheat. In *Plant and animal genomes XII conference* (10–14 January). San Diego, CA, USA, P321.

Kumar B., Talukdar A., Verma A., Girmilla V., Bala I., Lal S. K. *et al.* 2014 Screening of soybean [*Glycine max* (L.) Merr.] genotypes for yellow mosaic virus (YMV) disease resistance and their molecular characterization using RGA and SSRs markers. *AJCS* **8**, 27–34.

Maccaferri M., Sanguineti M. C., Noli E. and Tuberosa R. 2005 Population structure and long-range linkage disequilibrium in a drum wheat elite collection. *Mol. Breed.* **15**, 271–289.

Malysheva-Otto L. V., Ganal M. W. and Roder M. S. 2006 Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.) *BMC Genet.* **7**, 6.

Maskri Y. A., Sajjad M. and Khan S. H. 2012 Association mapping: A step forward to discovering new alleles for crop improvement. *Int. J. Agr. Biol.* **14**, 153–160.

Mather D. E., Hayes P. M., Chalmers K., Eglinton J., Matus I., Richardson K. *et al.* 2004. Use of SSR marker data to study linkage disequilibrium and population structure in *Hordeum vulgare*: Prospects for association mapping in barley. In *Linkage disequilibrium workshop*, april 4–7, Novotel Barossa Valley Resort, South Australia.

Neumann K., Kobiljski B., Dencic S., Varshney R. K. and Borner A. 2011 Genome-wide association mapping: a case study in bread wheat (*Triticum aestivum* L.) *Mol Breed.* **27**, 37–58.

Nordborg M. and Tavare S. 2002 Linkage disequilibrium: What history has to tell us? *Trends Genet.* **18**, 83–90.

Palaisa K., Morgante M., Williams M. and Rafalski A. 2004 Long range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci. USA* **101**, 9885–9890.

Price A. H. 2006 Believe it or not, QTLs are accurate! *Trends Plant Sci.* **11**, 213–216.

Pritchard J. K., Stephens M. and Donnelly P. 2000 Inference of population structure using multi-locus genotype data. *Genetics* **155**, 945–959.

Remington D. L., Thornsberry J. M., Matsuoka Y., Wilson L. M., Whitt S. R., Doebley J. *et al.* 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. USA* **98**, 11479–11484.

Saghai-Maroof M. A., Soliman K. M., Jorgensen R. A. and Allard R. W. 1984 Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. USA* **81**, 8014–8018.

Senior M. L. and Henn M. 1993 Mapping maize microsatellites and polymerase chain reaction confirmation of the targeted repeats using a CT primer. *Genome* **36**, 884.

Singh R. K., Bhat K. V., Bhatia V. S., Mohapatra T. and Singh N. K. 2008 Association mapping for photoperiod insensitivity trait in soybean. *Natl. Acad. Sci. Lett.* **31**, 281–283.

Smith J. R. and Nelson R. L. 1986 Relationship between seed-filling period and yield among soybean breeding lines. *Crop Sci.* **26**, 469–472.

Stich B., Mohring J., Piepho H. P., Heckenberger M., Buckler E. S. and Melchinger A. E. 2008 Comparison of mixed-model approaches for association mapping. *Genetics* **178**, 1745–1754.

Sun G., Zhu C., Kramer M. H., Yang S. S., Song W., Piepho H. P. *et al.* 2010 Variation explained in mixed model association mapping. *Heredity* **105**, 333–340.

Szalma S. J., Buckler I. V. E. S., Snook M. E. and McMullen M. D. 2005 Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks. *Theor. Appl. Genet.* **110**, 1324–1333.

Tenaillon M. I., Sawkins M. C., Long A. D., Gaut R. L., Doebley J. F. and Gaut B. S. 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* L.) *Proc. Natl. Acad. Sci. USA* **98**, 9161–9166.

Thornsberry J. M., Goodman M. M., Doebley J., Kresovich S., Nielsen D. and Buckler E. S. I. V. 2001 Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.* **28**, 286–289.

Wang J., McClean P. E., Lee R., Goos R. J. and Helms T. 2008 Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. *Theor. Appl. Genet.* **116**, 777–787.

Yu J., Pressoir G., Briggs W. H., Bi I. V., Yamasaki M., Doebley J. F. *et al.* 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208.

Zhang P., Li J., Li X., Liu X., Zhao X. and Lu Y. 2011 Population structure and genetic diversity in a rice core collection (*Oryza sativa* L.) investigated with SSR markers. *PLoS One* **6**, e27565.

Zhang W. J., Niu Y., Bu S. H., Li M., Feng J. Y., Zhang J. *et al.* 2014 Epistatic association mapping for alkaline and salinity tolerance traits in the soybean germination stage. *PLoS One* **9**, e84750.

Zhao K., Tung C. W., Eizenga G. C., Wright M. H., Ali M. L., Price A. H. *et al.* 2011 Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* **2**, 1–10.

Zhu C., Gore M., Buckler E. S. and Yu J. 2008 Status and prospects of association mapping in plants. *Plant Genome* **1**, 5–20.

Received 31 December 2013, in final revised form 15 July 2014; accepted 23 July 2014
Unedited version published online: 1 August 2014
Final version published online: 18 December 2014