Theor Appl Genet (2012) 125:517–530
DOI 10.1007/s00122-012-1850-6

ORIGINAL PAPER

Population structure and linkage disequilibrium in *Lupinus albus* L. germplasm and its implication for association mapping

Muhammad Javed Iqbal · Sujan Mamidi · Rubina Ahsan · Shahryar F. Kianian · Clarice J. Coyne · Anwar A. Hamama · Satya S. Narina · Harbans L. Bhardwaj

Received: 30 August 2011 / Accepted: 10 March 2012 / Published online: 28 March 2012

Abstract White lupin (*Lupinus albus* L.) has been around since 300 B.C. and is recognized for its ability to grow on poor soils and application as green manure in addition to seed harvest. The seed has very high levels of protein (33–47 %) and oil (6–13 %). It also has many secondary metabolites that are potentially of nutraceutical value to animals and humans. Despite such a great potential, lupins role in modern agriculture began only in the twentieth century. Although a large collection of *Lupinus* germplasm accessions is available worldwide, rarely have they been genetically characterized. Additionally, scarce genomic resources in terms of recombinant populations and genome information have been generated for *L. albus*. With the advancement in association mapping methods, the natural populations have the potential to replace the recombinant populations in gene mapping and marker-trait associations. Therefore, we studied the genetic similarity, population structure and marker-trait association in a USDA germplasm collection for their current and future application in this crop improvement. A total of 122 PI (Plant Inventory) lines were screened with 18 AFLP primer pairs that generated 2,277 fragments. A subset of 892 polymorphic markers with MAF >0.05 (minor allele frequency) were used for association mapping. The cluster analysis failed to group accessions on the basis of their passport information, and a weak structure and low linkage disequilibrium (LD) were observed indicating the usefulness of the collection for association mapping. Moreover, we were also able to identify two markers (a p value of 1.53×10^{-4} and 2.3×10^{-4}) that explained 22.69 and 20.5 % of seed weight variation determined using R^2_{LR}. The implications of lack of geographic clustering, population structure, low LD and the ability of AFLP to map seed weight trait using association mapping and the usefulness of the PI collections in breeding programs are discussed.

Communicated by J. Yu.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1850-6) contains supplementary material, which is available to authorized users.

M. J. Iqbal (✉) · S. Mamidi · S. F. Kianian
Department of Plant Sciences,
North Dakota State University, Fargo, ND 58108, USA
e-mail: muhammad.iqbal@ndsu.edu

R. Ahsan
Department of Biological Sciences, North Dakota State University, Fargo, ND 58108, USA

C. J. Coyne
USDA-ARS Western Regional Plant Introduction Station, 59 Johnson Hall, WSU, Pullman, WA 99164-6402, USA

A. A. Hamama · S. S. Narina · H. L. Bhardwaj (✉)
Agricultural Research Station, Virginia State University, Petersburg, VA 23806, USA
e-mail: hbhardwj@vsu.edu

Introduction

Lupinus albus L. is an old world (Pazy et al. 1977) species of genus *Lupinus* with $2n = 50$ (Gladstones 1998) and has long been cultivated around the Mediterranean and in the Nile valley (Gladstones 1998; Zohary and Hopf 2000). Major producing countries as of 2007 include Australia, Germany, Chile, and Poland followed by some sustained lower level production in South Africa, Morocco, and France. In Europe, the reasons for lupin production contraction include Anthracnose and Fusarium wilt diseases and competition with soybean imports. In the early part of twentieth century, it has been used as a cover crop in many

parts of the USA (Cowling et al. 1998; Huyghe 1997; Wells et al. 1980). However, due to the phenomenal success of soybean in the USA agriculture system, the availability of affordable nitrogen fertilizers and the presence of some undesirable alkaloids in older varieties, it did not receive due attention needed to become a major legume crop. Recently, there is increased interest in lupins in southern USA for use as a late winter, high protein livestock feed, food, forage and cover crop (Bhardwaj et al. 1998; Bhardwaj et al. 2004; Hamama and Bhardwaj 2004; Noffsinger and van Santan 2005; Bhardwaj 2006; van Santen et al. 2006; Hill and van Santen 2006).

Germplasm collections are important source of genes for improving disease and pest resistance and tolerance to abiotic stresses in breeding programs in addition to the species conservation. According to the IPGRI Directory of Germplasm Collections Database, there are estimated 40,000 holdings of *Lupinus* germplasm accessions around the world (Wolko et al. 2011). USDA-NPGS has a collection of over 200 PI lines of *L. albus* collected/donated from various parts of the world. Genetic diversity studies have been conducted on some *Lupinus* germplasm collections, which have provided useful information for crop improvement programs. Four geographical races of *L. albus* from the Mediterranean region were characterized using multivariate statistical method (Simpson 1986). Neves-Martins (1986, 1994) evaluated 200 Portuguese *L. albus* ecotypes using a range of morphological characters and described winter, spring, and intermediate types. However, the diversity studies based on morphological traits have their limitations as they can be easily influenced by the environment and growth conditions. Moreover, the utility of germplasm collections for crop improvement rests largely on the accuracy of evaluation and passport data, and on the genetic fidelity of the materials. There is a need, therefore, to test the genetic identity of all accessions held within a collection.

Major agricultural crops such as corn, wheat, soybean and to some extent cotton have received most of the attention for the development of genomic resources including genome sequencing, identification of high density single nucleotide polymorphism (SNP) and the development of new powerful approaches to the mapping of complex traits and to the subsequent identification of causal genes. The genome size of the genus *Lupinus* is relatively small. The DNA amount of *L. albus* is 0.6 pg/1C, which is slightly larger than *Arabidopsis thaliana* (L.) Heynh. (0.30 ± 0.14 pg/1C) (http://data.kew.org/cvalues/CvalServlet?querytype=2). A National Center for Biotechnology Information search showed that there are little over 9,000 (9,325) expressed sequence tags (ESTs), 150 Genome Survey Sequence (GSS) records and 890 nucleotide sequences available for *L. albus* (http://www.ncbi.nlm.nih.gov) and also reported in Tian et al. (2009) and Rodriguez-Medina et al. (2011). There have also been some efforts to construct linkage maps for *L. albus*. Phan et al. (2007) identified 28 major linkage groups (three more than the haploid chromosome number) in an F_8 RIL population using STS and AFLP markers. In this study, they identified QTLs for anthracnose resistance, flowering time and seed alkaloid content. In another independent mapping study using F_5 RIL and STS and AFLP markers, Croxford et al. (2008) identified 25 linkage groups and mapped QTLs for flowering time, seed alkaloid content, and stem height. The two studies shared only a few markers which made the comparison between the two linkage maps difficult.

Genome-wide association studies in populations of unrelated individuals provide an efficient way to map the locations of quantitative trait loci (QTL) (Rafalski 2010; Astle and Balding 2009). Compared to linkage mapping where allele frequencies and recombination events are determined by experimental design, the association mapping faces challenges that arise from the complex history of the populations under study (Hamblin et al. 2011; Myles et al. 2009). Distinct patterns of population structure, allele frequency distribution and linkage disequilibrium arise from the domestication history, breeding history, ancestral population characteristics and mating system of the crop under investigation (Hamblin et al. 2011). Therefore, the knowledge of all these attributes of the populations under study is important to achieve maximum power and resolution with appropriate experimental designs. The association mapping approach can be exploited in the lupin breeding and genetics as large well-characterized (phenotyped) germplasm collections and breeding populations exist for the main agricultural lupin species (Wolko et al. 2011). The main limiting factor is the scarcity of genomic information for effective implementing of association mapping. However, even before the start of SNP discovery and large scale genotyping of populations, it is important to explore the population structure and the level of linkage disequilibrium (LD) in the target populations.

There are several methods available for genetic diversity and population structure analysis including random amplified polymorphic DNA (RAPD), simple sequence repeat (SSR) and amplified fragment length polymorphism (AFLP) markers among many others. Application and suitability of various marker systems in genetic diversity analysis and gene bank management have been extensively reviewed by Spooner et al. (2005). More recently, Zhang et al. (2011) has used diversity arrays technology (DArT) markers to investigate the genetic diversity and population structure of Chinese common wheat (*Triticum aestivum* L.). Although the DArT is a sequence-independent method

but requires complexity reduction steps that can be biased to the certain genomic regions. In the absence of the detailed knowledge of the molecular basis or DNA sequence of the trait of interest, the whole genome scan or genome-wide association study works better for association mapping. The whole genome scan involves genotyping densely distributed marker loci covering all the chromosomes and, therefore, testing for association of most of the markers covering the genome (Rafalski 2010). Therefore, due to the lack of sequence information for *L. albus* and the potential uniform coverage by AFLP analysis, we selected AFLP markers for exploring the genetic diversity of *L. albus* PI lines and investigated the suitability of the PI lines for association mapping and marker-trait association. We further employed various models to analyze the population structure or lack of it which can have far reaching consequences for association mapping using any germplasm collection. Using association mapping, the 122 PI lines were tested for seed weight, an important agronomic trait for association with the AFLP markers.

Materials and methods

Plant materials and DNA isolation and AFLP analysis

A total of 122 PI lines were obtained from the USDA germplasm collection (Table 1) and grown in 30-cm pots at the Agricultural Research Station of Virginia State University, Petersburg, VA, USA. Young leaves were collected and DNA was isolated using DNeasy plant mini kit (QIAgen, CA, USA) according to the manufacturer's instructions. The quality of isolated DNA was checked by running on 0.8 % agarose gel in TBE buffer and the concentration was measured by a UV–Visible spectrophotometer. AFLP analysis was carried out using 20 primer pairs (Supplementary Table 1) according to the Vos et al. (1995) with some modifications (Chang et al. 2009). The AFLP fragments were scored as present (1) and absent (0) for each amplified locus by the fragment analysis software of CEQ8800. The binary data sets were exported for further analysis.

Data analysis: imputation and marker statistics

A likelihood based imputation was used to impute missing data implemented in fastPHASE 1.3 with default settings (Scheet and Stephens 2006). Minor allele frequency (MAF) was estimated in Powermarker 3.0 (Liu and Muse 2005). Nei's gene diversity (Nei 1973) provides an estimation of the discriminatory power of each marker (Le Couviour et al. 2011) and was calculated using the software Popgene 1.32 (Yeh et al. 1999).

UPGMA tree

For further analysis, only markers with MAF >0.05 were used. A similarity matrix with Jaccard similarity coefficient as suggested by Blair et al. (2011) was estimated in SAS 9.2 using the DISTANCE procedure. Further, a UPGMA tree was built using CLUSTER and TREE procedures in SAS 9.2.

Population structure

To prevent bias in estimation of population structure, we used a subset of markers that have an LD <0.5 with every other marker estimated in TASSEL (Bradbury et al. 2007). For estimation of number of sub-populations, STRUCTURE 2.3 was used. The basic algorithm was described by Pritchard et al. (2000). Extensions to the method were published by Falush et al. (2007) and Hubisz et al. (2009). The admixture model was used with a burn in of 100,000 and 500,000 iterations for sub-populations numbers (k) ranging from 1 to 15 considering the allele frequencies to be independent. Five runs for each k value were performed, and the posterior probability of the model was determined for each run. The optimum number of subpopulations was determined using the Δk approach by Evanno et al. (2005) and Wilcoxon two sample tests as described by Rosenberg et al. (2001). For the Δk approach, we used structure harvester (http://taylor0.biology.ucla.edu/struct_harvest) to obtain the best number of subpopulations. For the Wilcoxon test, we compared the posterior probabilities of two successive sub-populations ($k1$ vs. $k2$, $k2$ vs. $k3$, $k3$ vs. $k4$, and so on) using the NPAR1WAY procedure in SAS. The smaller k value in a pairwise comparison for the first non-significant Wilcoxon test was chosen as the best number of subpopulations (Mamidi et al. 2011).

Genotypes were further divided into sub-populations based on membership coefficients estimated in STRUCTURE. We used Popgene 1.32 to estimate the genetic identity and genetic distance between the different subpopulations. Principal component analysis (PCA) which also controls for population structure was performed using the PRINCOMP procedure in SAS. A 3D plot of the first three PCs was used to visualize the dispersion of the genotypes.

Linkage disequilibrium

LD coefficients (r^2) were calculated for each of the pairwise comparison and the significance was estimated using 1,000 permutations in TASSEL. Average r^2 and percent of observations $p < 0.01$ significance level were estimated over all the pairwise comparisons (Rossi et al. 2009). Since

Table 1 List of *L. albus* accessions, their passport information, the cluster (as in Fig. 1) they fall in after genetic similarity analysis (UPGMA tree) and subpopulation number based on 8 [Sub # (8-pop)] or 5 [Sub # (5-pop)] subpopulations from STRUCTURE analysis

Country of origin	Plant-ID	Accession	DNA ID	Cluster no.	Sub # (8-pop)	Sub # (5-pop)
Algeria	LA 71	PI 457940	39	4	5	5
AR, USA	LINE NO. 10	PI 606481	118	5	1	2
AR, USA	LINE NO. 7	PI 615405	123	5	1	2
AR, USA	LINE NO. 8	PI 615406	124	5	7	2
Australia	HAMBURG	PI 469095	61	4	5	5
Australia	ULTRA	PI 469096	62	4	5	5
Brazil	MN 3	PI 244572	6	2	6	1
Bulgaria	31620	PI 368914	16	3	3	4
Bulgaria	31621	PI 368915	17	3	8	4
Bulgaria	VIR 1550	PI 481549	71	3	7	4
Bulgaria	CPI 31620	PI 487437	89	4	5	5
Czech Republic	22840	PI 368911	15	4	2	1
Egypt	TERMIS	PI 250094	7	4	6	3
Egypt	NA	PI 250572	8	2	6	3
Egypt	EGYPTICA	PI 481551	73	3	8	4
Egypt	E92-2	PI 606484	121	5	1	2
Egypt	E92-13	PI 606485	122	3	1	2
Former Soviet Union (FSU)	KIEVSKIJ MUTANT	PI 381322	18	4	3	4
France	LA 25	PI 457933	32	1	8	3
France	LA 26	PI 457934	33	1	2	3
France	LA 27	PI 457935	34	3	8	3
France	LUBLANC	PI 467348	57	3	3	4
France	LUCKY	PI 467349	58	3	3	4
France	C 9	PI 467351	59	3	8	3
France	LUCKY	PI 606482	119	3	2	3
FSU	KIEV EARLY	PI 487434	87	4	4	1
FSU	KIEV N409	PI 487435	88	4	4	1
GA, USA	TIFWHITE-78	PI 615409	125	3	7	2
Germany	PFLUG-MANSA	PI 237719	4	4	6	5
Germany	MN 48	PI 287241	12	4	4	1
Germany	SAATGUT	PI 316610	14	4	2	1
Germany	P.B. WEISLUP	PI 481554	76	1	8	3
Germany	WEISE BITTERLUPINE	PI 481555	77	4	5	5
Germany	NA	PI 502648	90	1	2	3
Germany	GR 337	PI 516625	100	1	8	4
Germany	GR 338	PI 516626	101	3	2	3
Germany	GR343	PI 516630	126	1	3	4
Greece	GR 1	PI 457921	22	4	5	5
Greece	GR 3	PI 457923	23	4	5	5
Greece	GR 5	PI 457924	24	3	8	3
Greece	GR 7	PI 457926	25	1	8	3
Greece	GR 8	PI 457927	26	3	8	2
Greece	GR 9	PI 457928	27	3	2	3
Greece	GR 10	PI 457929	28	3	5	3
Greece	GR 11	PI 457930	29	4	5	5
Greece	GR 12	PI 457931	30	4	5	5
Greece	GR 13	PI 457932	31	4	5	5

Table 1 continued

Country of origin	Plant-ID	Accession	DNA ID	Cluster no.	Sub # (8-pop)	Sub # (5-pop)
Greece	LA 56	PI 457936	35	3	2	3
Greece	LA 57	PI 457937	36	4	4	1
Greece	VIR 1437	PI 481546	68	3	7	5
Hungary	GYULATANYAI EDES	PI 232924	3	2	6	3
Hungary	KRAFTQUELL	PI 289160	13	4	5	5
Hungary	VIR 1504	PI 481547	69	4	4	1
Hungary	ME 74	PI 502652	93	1	2	3
Italy	LA 106	PI 457959	56	1	2	3
Italy	GR 334	PI 516624	99	3	4	1
Lebanon	IFLU 32	PI 483074	84	1	3	3
Morocco	9483	PI 457938	37	4	4	1
Morocco	VIR 2005	PI 481556	78	4	5	5
Morocco	IFLU 31	PI 483073	83	3	2	3
Morocco	GR 333	PI 516623	98	4	4	1
NA	no. 22	PI 543013	116	5	2	3
NA	no. 563	PI 543024	117	5	3	2
The Netherlands	NA	PI 168891	1	2	2	3
New Zealand	KALI	PI 434855	20	4	3	4
New Zealand	ULTRA	PI 434856	21	4	5	1
Poland	KALI	PI 386098	19	4	3	4
Poland	WTD 180	PI 468129	60	4	5	5
Poland	KALINA	PI 476374	66	1	3	4
Poland	KALI	PI 476375	67	4	4	4
Poland	BIALY POZNY	PI 481548	70	4	5	5
Poland	KALI	PI 502650	91	3	3	4
Poland	BIALY 1	PI 505844	94	1	8	3
Russian Federation	VIR 1423	PI 457956	55	4	5	5
South Africa	NA	PI 243335	5	4	6	1
Spain	LA 70	PI 457939	38	4	5	5
Spain	1082	PI 457941	40	1	2	3
Spain	1107	PI 457942	41	1	8	3
Spain	1186	PI 457943	42	3	3	4
Spain	1134	PI 457944	43	3	2	3
Spain	1190	PI 457945	44	4	5	5
Spain	1585	PI 457946	45	4	5	5
Spain	1586	PI 457947	46	4	5	5
Spain	1587	PI 457948	47	4	5	5
Spain	1588	PI 457949	48	3	8	3
Spain	1589	PI 457950	49	1	2	3
Spain	1590	PI 457951	50	1	2	3
Spain	1591	PI 457952	51	3	8	3
Spain	1592	PI 457953	52	2	2	3
Spain	1593	PI 457954	53	4	5	5
Spain	1594	PI 457955	54	4	5	5
Spain	VIR 2362	PI 481559	80	4	5	5
Spain	No. 267	PI 533694	102	3	3	4
Spain	No. 269	PI 533695	103	3	3	4
Spain	No. 530	PI 533697	105	5	1	1

Table 1 continued

Country of origin	Plant-ID	Accession	DNA ID	Cluster no.	Sub # (8-pop)	Sub # (5-pop)
Spain	No. 544	PI 533698	106	5	1	2
Spain	No. 558	PI 533700	107	5	1	2
Spain	No. 571	PI 533701	108	3	1	2
Spain	No. 576	PI 533702	109	1	2	3
Spain	No. 584	PI 533703	110	5	1	2
Spain	R-6002, NORTo 486	PI 533704	111	5	2	3
Spain	R-6019, NORTo 484	PI 533705	112	5	1	3
Spain	No. 47	PI 533706	113	5	1	2
Spain	No.175	PI 533707	114	5	1	1
Spain	870529-02	PI 533714	115	5	2	3
Spain	MULTULUPA	PI 606483	120	3	1	2
Sudan	ME 51	PI 476370	63	4	5	5
Sudan	VIR 1644	PI 481552	74	1	2	3
Syria	VIR 2229	PI 481558	79	4	5	5
Syria	IFLU 29	PI 483072	82	1	3	4
Syria	IFLU 33	PI 483075	85	1	8	3
Turkey	NA	PI 179361	2	2	2	3
Ukraine	KIEVSKIJ SKOROSPELYJ	PI 476372	64	3	3	4
Ukraine	GORIZONT	PI 476373	65	1	3	4
Ukraine	NOSOVSKIJ-3	PI 505845	95	3	8	3
Ukraine	KIEVSKIJ MUTANT	PI 505846	96	4	4	1
Ukraine	LOTOS	PI 505847	97	4	4	1
Ukraine	VIR 2603	PI 533696	104	5	2	1
Yugoslavia	MN 181	PI 251559	9	1	8	2
Yugoslavia	NA	PI 255375	10	3	8	2
Yugoslavia	NA	PI 255471	11	3	2	3
Yugoslavia	VIR 2374	PI 481560	81	3	2	3

no mapping data were available, we assumed that 10 %, or 15 % or 20 % or 25 % of the total pairwise comparisons could be intra-chromosomal comparisons and the rest are inter-chromosomal comparisons. We created 100,000 random permutation datasets for each of the four levels assumed in SAS 9.2. For each of the permuted dataset, we calculated the average r^2 and percent of observations <0.01 significance level.

Association mapping

To test the usefulness of this population for association mapping, we used seed weight, an important agronomic trait. The phenotypic data were obtained through weighing 100 seeds (g) of harvested seed from regeneration plots at Washington State University's Whitlow farm (46°43′28″N 117°08′07″W), Pullman, WA, USA. The number of principal components (eigenvectors) which collectively explain 25 % of the variation was selected for the association analysis (Stich and Melchinger 2009); in addition to structure matrix that has membership coefficients of an individual in a sub population. A pairwise Loiselle kinship coefficient matrix (**K** matrix) (Loiselle et al. 1995) was estimated using SPAGeDi 1.2 (Hardy and Vekemans 2002). Negative values for the kinship matrix were set to zero as described by Yu et al. (2006). Another allele similarity matrix $\mathbf{K}^*$ (Zhao et al. 2007), representing the proportion of shared alleles for all pairwise comparisons in each population, was estimated in SAS 9.2.

Twelve different linear regression models were tested for marker-trait association using the MIXED procedure in SAS 9.2 (Table 2). The underlying equation for the 12 models is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Q}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{v} + \boldsymbol{\varepsilon}$$

In this model, $\mathbf{y}$ is a vector for phenotypic observations, $\boldsymbol{\alpha}$ is the fixed effects related to the AFLP marker, $\boldsymbol{\beta}$ is a vector of the fixed effects related to the population structure, $\boldsymbol{v}$ is a vector of the random effects related to the relatedness among the individuals, and $\boldsymbol{\varepsilon}$ is a vector of

Table 2 Summary of the statistical models used to test the data for marker-trait associations

Model	Statistical model	Information captured in the model
Naive	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$	**y** is related to **X**
K	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{K}\mathbf{v} + \boldsymbol{\varepsilon}$	**y** is related to **X**, with Kinship (Loiselle coefficient)
K*	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{K}^{*}\mathbf{v} + \boldsymbol{\varepsilon}$	**y** is related to **X**, with allele similarity matrix
Q	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Q}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	**y** is related to **X**, with **Q**
PCA	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{P}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	**y** is related to **X**, with PCA
Q + **K**	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Q}\boldsymbol{\beta} + \mathbf{K}\mathbf{v} + \boldsymbol{\varepsilon}$	**y** is related to **X**, with **Q** and (Loiselle coefficient)
Q + **K***	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Q}\boldsymbol{\beta} + \mathbf{K}^{*}\mathbf{v} + \boldsymbol{\varepsilon}$	**y** is related to **X**, with **Q** and allele similarity matrix
PCA + **K**	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{P}\boldsymbol{\beta} + \mathbf{K}\mathbf{v} + \boldsymbol{\varepsilon}$	**y** is related to **X**, along with PCA and kinship (Loiselle coefficient)
PCA + **K***	$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{P}\boldsymbol{\beta} + \mathbf{K}^{*}\mathbf{v} + \boldsymbol{\varepsilon}$	**y** is related to **X**, along with PCA and allele similarity matrix

All the three models with **Q** were run for five and eight subpopulations

the residual effects. **X** is a matrix of alleles of the markers, **P** is the matrix of the principal components (in place of **Q** matrix), **K** is the Loiselle kinship coefficient matrix, and **K*** is the allele similarity matrix. The variances of the random effects were estimated as Var $(u) = 2\mathbf{K}V_{\mathrm{g}}$ and Var $(e) = \mathbf{I}V_{\mathrm{R}}$, where **K** is a kinship matrix, **I** is an identity matrix with the off-diagonal elements as 0 and diagonal elements is the reciprocal of the number of the observations for which the phenotypic data were obtained, V_{g} is the genetic variance, and V_{R} is the residual variance. For each model, the positive false discovery rate (pFDR) was estimated for all markers using the MULTTEST procedure in SAS 9.2 to correct for multiple marker-trait association.

For the selection of best model, mean square difference (MSD) was calculated as:

$$\mathrm{MSD} = \frac{\sum_{i=1}^{n}\left(p_i - \frac{i}{n}\right)^2}{n}$$

where i is the rank number, p_i is the probability of the ith ranked p value, and n is the number of markers (Mamidi et al. 2011). Best model is defined as the one with lowest MSD value. The multiple R^2_{LR} values for the significant loci were calculated using MIXED procedure in SAS as described in Sun et al. (2010).

Results

A total of 20 *Eco*RI + *Mse*I primer combinations were used for the AFLP analysis of the 122 *L. albus* PI lines. However, two primer pairs did not produce consistent amplification profiles among majority of the varieties and, therefore, were dropped from the analysis. The 18 primer pairs amplified a total of 2,277 fragments that were detected by the CEQ 8800 genetic analysis system. The scored fragments ranged in size from ≥53 to ≤650 bp in length. The average number of DNA fragments amplified by each primer pair was 126. This relatively high number of amplified fragments was due to the high sensitivity of D4-dye detection by the CEQ 8800 system and the two additional selective nucleotide used in the EcoRI primer (*Eco*RI + 2) in the selective amplification. The *Mse*I primer contained 3 additional nucleotides (*Mse*I + 3) in the selective amplification.

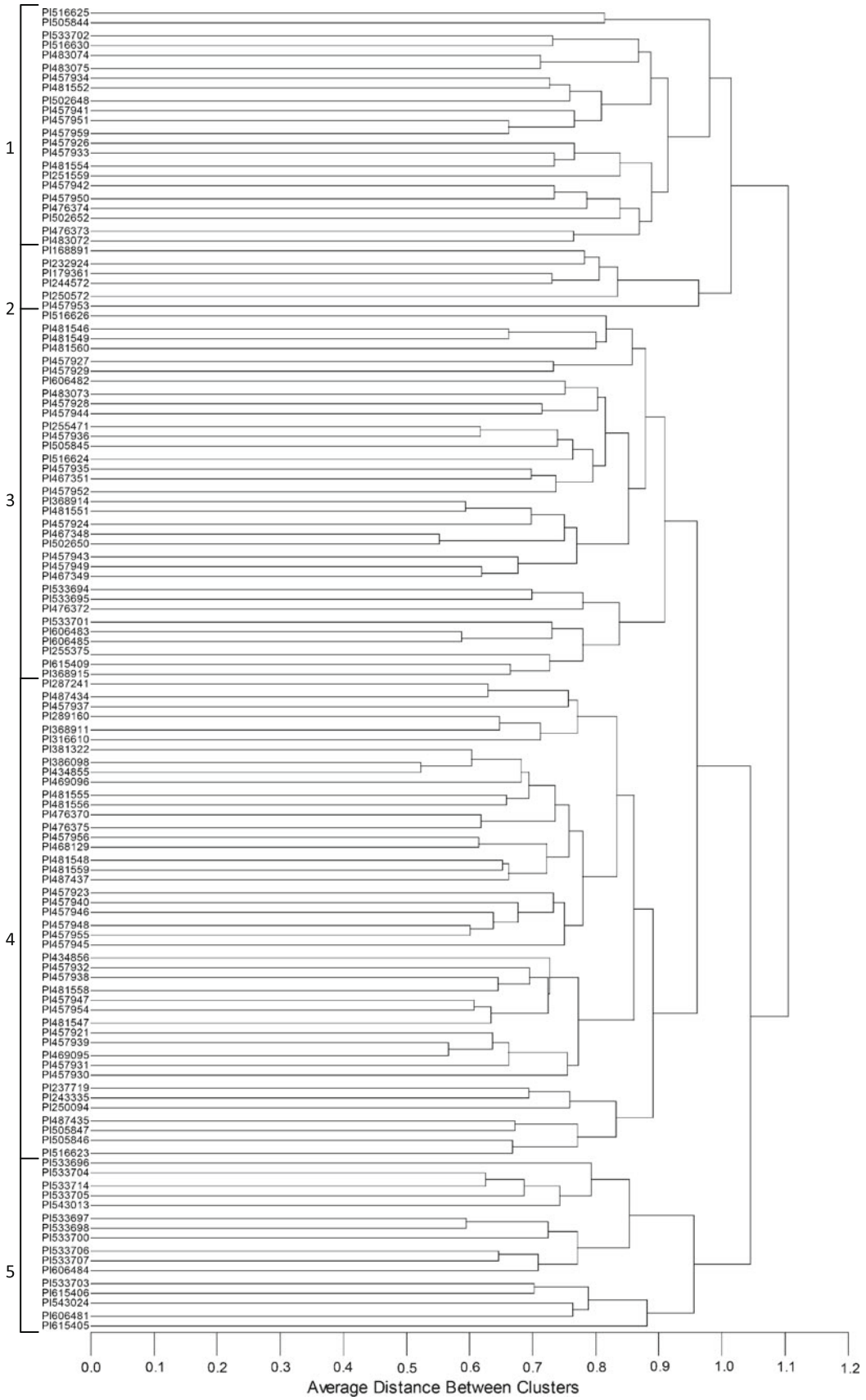
A total of 2,277 AFLP loci were used for the analyses. Missing loci which contribute about 9.55 % were imputed. Of the 2,277 loci, 892 loci have a MAF >0.05, and only these were used for further analyses. For these loci that have a MAF >0.05, the mean Nei's gene diversity was 0.2985.

Genetic diversity

The UPGMA analysis of the AFLP fingerprints for the 122 lines resulted into five major clusters (Fig. 1). The fewest number of accessions (six) was grouped in second cluster and the largest numbers of accessions (44) were grouped in the fourth cluster. The other three clusters contained 16 (cluster number 5), 22 (cluster number 1) and 34 (cluster number 3) accessions. However, the cluster analysis did not group accessions according to their country of origin. The second cluster grouped accessions originating from Brazil, Egypt, Hungary, The Netherlands, Spain and Turkey (Table 1). The most closely clustered accessions PI434855 and PI386098 (cluster 4) were collected/donated from New Zealand and Poland, respectively. Similarly, the second most closely related accessions pair, PI467348 and PI502650 (cluster 3) originated from France and Poland, respectively. The similarity (Jaccard similarity coefficient) of individuals ranges from 0.276 to 0.662. The highest is between the accessions PI 502650 and PI 467348, and the lowest is between the accessions PI 457950 and PI 615406.

Population structure

Only 625 loci that have an LD <0.5 with any other loci were used for population structure analysis. The Bayesian-based clustering approach implemented in STRUCTURE


1
2
3
4
5
PI516625
PI505844
PI533702
PI516630
PI483074
PI483075
PI457934
PI481552
PI502648
PI457941
PI457951
PI457959
PI457926
PI457933
PI481554
PI251559
PI457942
PI457950
PI476374
PI502652
PI476373
PI483072
PI168891
PI232924
PI179361
PI244572
PI250572
PI457953
PI516626
PI481546
PI481549
PI481560
PI457927
PI457929
PI606482
PI483073
PI457928
PI457944
PI255471
PI457936
PI505845
PI516624
PI457935
PI467351
PI457952
PI368914
PI481551
PI457924
PI467348
PI502650
PI457943
PI457949
PI467349
PI533694
PI533695
PI476372
PI533701
PI606483
PI606485
PI255375
PI615409
PI368915
PI287241
PI487434
PI457937
PI289160
PI368911
PI316610
PI381322
PI386098
PI434855
PI469096
PI481555
PI481556
PI476370
PI476375
PI457956
PI468129
PI481548
PI481559
PI487437
PI457923
PI457940
PI457946
PI457948
PI457955
PI457945
PI434856
PI457932
PI457938
PI481558
PI457947
PI457954
PI481547
PI457921
PI457939
PI469095
PI457931
PI457930
PI237719
PI243335
PI250094
PI487435
PI505847
PI505846
PI516623
PI533696
PI533704
PI533714
PI533705
PI543013
PI533697
PI533698
PI533700
PI533706
PI533707
PI606484
PI533703
PI615406
PI543024
PI606481
PI615405
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
Average Distance Between Clusters

◀**Fig. 1** UPGMA tree showing the genetic relationship of individual accessions analyzed using AFLP markers. The dendrogram was generated with SAS 9.2 using Jaccard's similarity coefficient. Five clusters were identified and the passport information of the accessions belonging to each cluster is listed in Table 1

reveals the presence of five or eight subpopulations by Δk approach used for selecting the best number of subpopulations (Fig. 2a). Alternatively, the Wilcoxon test revealed the presence of eight subpopulations (Fig. 2b). Majority of the individuals have a membership coefficient (qi) <0.7 (111 of 122 individuals) to be assigned to a subpopulation revealing a weak population structure among individuals and/or an admixed sample (Rossi et al. 2009; Mamidi et al. unpublished).

We divided the population into five and eight subpopulations based on the estimated membership in the STRUCTURE matrix, to look at the genetic identity and genetic distance between the clusters (Table 3). When the population is divided into eight subpopulations, the genetic distance between the groups is within the range of 0.035–0.193 with a mean value of 0.096. The genetic identity within a cluster is in the range of 0.824–0.965 with a mean of 0.9085. When the population is divided into five subpopulations, the genetic distance between the groups is within the range of 0.0274–0.0504 with a mean value of 0.04116. The genetic identity within a cluster is in the range of 0.9508–0.973 with a mean of 0.9597. This also indicates a weak population structure similar to the results obtained above. Further, a plot of the three principal components that explain 13.7 % variation reveals no clear clustering pattern of the eight subpopulations and supports the idea of a weak population structure (Fig. 3). Eight principal components that explain 25.5 % variation were included for association mapping analyses. The allele similarity between individuals is in the range of 0.53–0.8528.

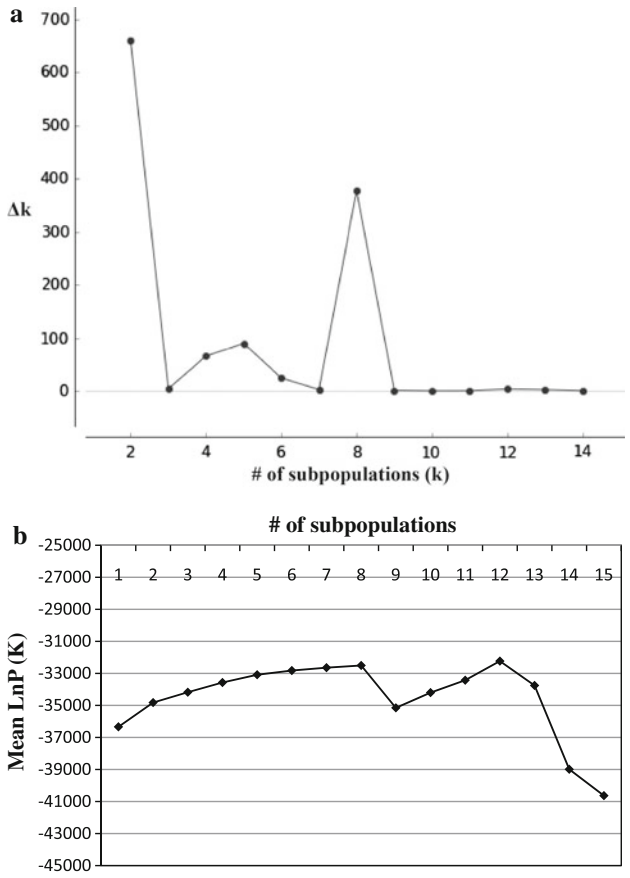


Fig. 2 **a** A graph with Δk and number of subpopulations to determine the number of subpopulations (Evanno et al. 2005). The *peak* represents the appropriate number of subpopulations. **b** A *graph* generated from Wilcoxon test with mean LnP (k) on y axis and number of sub-populations on X axis

Linkage disequilibrium

The overall average r^2 for all pairwise comparisons is 0.0168 (95 % CI 0.0165–0.0171) and the percent of observations with a $p < 0.01$ are 4.75 %. The distribution of average r^2 for 100,000 permutations (10 % pairwise comparisons) is within the range of 0.0166–0.0175 with a peak distribution at 0.01686. The proportion of r^2 values that are significant ($p < 0.01$) (10 % pairwise comparisons) is within the range of 4.375–5.125 with a peak distribution at 4.725. Similar distributions were obtained for other three levels of intra-chromosomal pairwise comparisons (15, 20, and 25 %).

Association mapping

Of the 12 models tested, model with PCA and **K** performed best (MSD = 0.00069). All other models have the MSD within the range of 0.0006–0.002. Two markers that meet the criteria of significance ($p < 0.05$ and pFDR < 0.1) were $E_{CAG}M_{CGC}76$ and $E_{CAC}M_{CGC}105$. These two markers have a p value of 1.53×10^{-4} and 2.3×10^{-4}, respectively (Table 4). These two markers explain 22.69 and 20.5 % of seed weight variation, respectively.

Discussion

Plant breeders use genetic resources to create novel gene combinations and to select crop varieties more suited for diverse agriculture systems and rapidly changing climatic conditions. There are over 1,400 gene banks containing a wealth of over six million accessions available and accessible for crop improvement (Hammer et al. 2003). Still, these resources are barely used (Upadhyaya et al. 2006) by breeders, may be due to the scarcity of information about these collections other than their geographic origin and

Table 3 Subpopulation differentiation and genetic identity for five and eight subpopulations.

Five subpopulations					
Subpop.	1	2	3	4	5
1	****	0.9569	0.9635	0.9548	0.9688
2	0.0441	****	0.9617	0.9584	0.9527
3	0.0371	0.0391	****	0.973	0.9508
4	0.0463	0.0425	0.0274	****	0.9564
5	0.0317	0.0484	0.0504	0.0446	****

Eight subpopulations								
Subpop.	1	2	3	4	5	6	7	8
1	****	0.9259	0.9166	0.9034	0.9015	0.8667	0.8923	0.9228
2	0.0770	****	0.9491	0.9351	0.9116	0.9072	0.8857	0.9616
3	0.0871	0.0523	****	0.9254	0.9217	0.8817	0.9057	0.9655
4	0.1016	0.0671	0.0775	****	0.9404	0.8954	0.8818	0.9185
5	0.1036	0.0925	0.0816	0.0615	****	0.8832	0.8913	0.9198
6	0.1431	0.0974	0.1259	0.1105	0.1242	****	0.8245	0.8921
7	0.1140	0.1214	0.0990	0.1258	0.1151	0.1930	****	0.9126
8	0.0803	0.0392	0.0351	0.0850	0.0837	0.1142	0.0915	****

Nei's genetic identity (above diagonal) and genetic distance (below diagonal)

taxonomic status. Genetic structure or the genetic diversity often reflects biologically meaningful processes. Understanding the patterns in genetic diversity and physical addresses of genes in genetic maps both of which are the result of natural processes, the characteristic of the species and historical events, can provide a stronger scientific basis for the faster and better use of germplasm collections in plant improvement. Molecular characterization has become the favored means to access variation within large germplasm samples. For non-model organisms, AFLPs are a valuable tool when large numbers of marker are required for genomic scans and subsequent hypothesis testing (Meudt and Clarke 2007). The anonymous AFLP markers consist largely of non-coding DNA (Shirasawa et al. 2004; Wong et al. 2001), are widely distributed throughout the genome and allow the assessment of genome-wide variation (Meudt and Clarke 2007). In this study, the 18 AFLP primer combinations amplified 2,277 fragments each potentially a unique locus giving an overall good coverage of the *L. albus* genome which is slightly bigger than *Arabidopsis*. Although the di-allelic loci such as AFLP are individually less informative, their sheer number gives the statistical power to outperform microsatellite loci for discriminating taxa and populations (Woodhead et al. 2005; Campbell et al. 2003).

Our results indicate that the clusters based on the genetic distances did not group accessions on the basis of their geographic origin. Gilbert et al. (1999) used ISSR-PCR based DNA fingerprints to study genetic variability among 37 *L. albus* accessions from University of Reading, UK collection. They scored 137 DNA bands and upon UPGMA analysis observed some evidence of clustering. However, they also failed to relate clustering to the geographical origin and suspected that the cause may be the poor documentation and widespread transportation of stocks between the countries. Our inability to correlate the clustering with the geographic origin of various accessions even with the 16 times higher (2,277) number of loci is also suspected to be due to extensive movement of germplasm across various countries before arrival at the USDA germplasm collection. The genetic similarity information of the germplasm collection generated in this study is more useful than the passport information in the optimal use for crop improvement. Moreover, the genetic information from the AFLP markers was also used to infer the population structure and their suitability for association mapping.

The *L. albus* germplasm studied showed weak signs of population structure. Although the genotypes are classified into five or eight sub-populations based on STRUCTURE, the genetic identity between the clusters is high. With this Bayesian model-based STRUCTURE, the estimated memberships for each individual to be assigned to a subpopulation are very low. In addition to these, the MSD values calculated are approximately similar, for example the naïve model has an MSD of 0.0014, the **Q**-model has an MSD of 0.0017 and the PCA model has 0.0007. This means that addition of a structure matrix does not add much to the model supporting a weak population structure.

This was not an unexpected result as the ancestral history of this species is different than other major crops such as rice, corn, beans, etc. First of all, the wild ancestor is only semi-domesticated (Wolko et al. 2011), which makes the species remain similar to the wild types without much selection in the semi-domesticated/cultivated forms. Second, the wild ancestor and the derived forms have the same Mediterranean distribution, which leads to no adaptation or selection differences. Third, due to the lack of domestication bottleneck, there was less impact upon the allele frequencies, and genetic variation that segregate within populations of cultivated plants (Hamblin et al. 2011). The weak population structure makes this genotype collection ideal as a starting point for association mapping because structure leads to spurious associations (Abdurakhmonov and Abdukarimov 2008; Astle and Balding 2009; Myles et al. 2009; Hamblin et al. 2011; Mamidi et al. 2011).

The power of an association mapping study depends on the strength of the LD (Hamblin et al. 2011; Myles et al. 2009; Abdurakhmonov and Abdukarimov 2008). LD is usually measured using r^2 and is said to perform well in small sample sizes (Flint-Garcia et al. 2003). The LD coefficient, r^2, summarizes both recombination and mutation history (Flint-Garcia et al. 2003). The overall LD measured in our sample was very low in terms of r^2.

With the ancestral history of *L. albus* and the weak population structure this result was expected. The low LD values can be explained by the lack of bottlenecks, and selection which generally reduce the diversity and change allele frequencies either to fixation or intermediate frequencies (Hamblin et al. 2011). In addition to these, there is no presence of genetic isolation, population structure or admixture that makes LD lower (Abdurakhmonov and Abdukarimov 2008). The other important factor that can

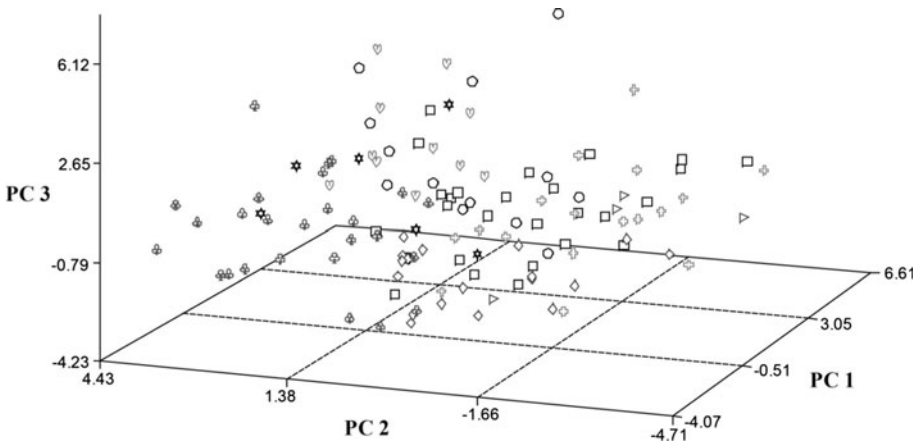


Fig. 3 A 3D plot of the first three principal components that together explain 13.7 % variation. Each *symbol* represents a subpopulation (8 subpopulations) estimated from Bayesian-based clustering software STRUCTURE 2.3

Table 4 Significant AFLP markers associated with seed weight trait in *L. albus*

Marker	MAF	Minor allele mean	Major allele mean	*p* value	pFDR	R^2
$E_{CAG}M_{CGC}76$	0.1475	24.65	30.65	1.53E−04	22.69	8.05
$E_{CAC}M_{CGC}105$	0.459	27.99	31.30	2.3 E−04	20.5	4.86

MAF minor allele frequency, *p* value obtained in a PCA + **K** mixed model, *pFDR* positive false discovery rate, R^2 % variation explained by the marker

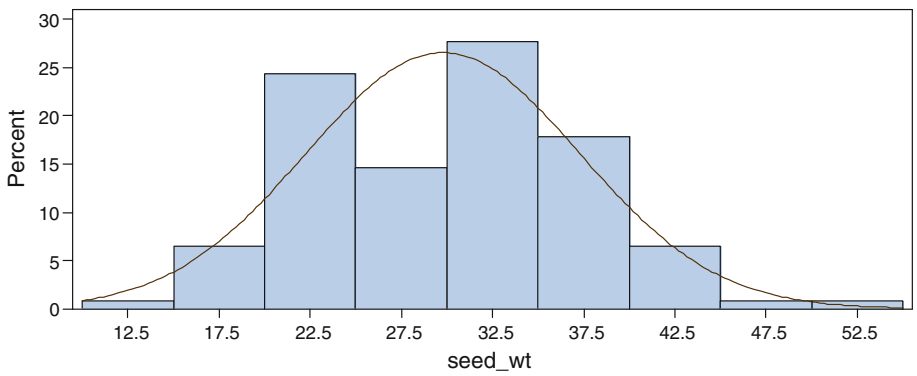


Fig. 4 Distribution of phenotypic values of seed trait for the 122 genotypes. Seeds were harvested from regeneration plots and random sample of 100 seeds were weighed (g)

lead to low LD is the pollinating pattern of the species. A self-pollinating crop has a higher LD because no opportunities for new recombinants can be generated (Nordborg et al. 2002). On the other hand, out-crossing leads to decreased LD due to creation of new recombination. *L. albus* even though a self-pollinating crop has an out-crossing rate of 8–10 % (Green et al. 1980) which may have lead to a lower LD. Additionally, low coverage can also lead to low LD.

Since allele frequencies have a large effect on the LD and can lead to inaccurate estimates, we imputed the missing values and removed the markers/loci that have a MAF <5 % based on suggestions of Abdurakhmonov and Abdukarimov (2008). LD varies among species, populations within a species and even the marker system used to capture the diversity information (Abdurakhmonov and Abdukarimov 2008). Differences in the extent of LD have an important implication for association mapping studies. With the low LD for this population, it can be inferred that a higher number of markers are needed to identify the QTL responsible for the traits. However, this can avoid the spurious associations which are possible due to the long stretched LD and or loci on different chromosomes (Abdurakhmonov and Abdukarimov 2008; Hamblin et al. 2011).

Population structure can be the result of common ancestry of large groups of individuals leads to spurious associations and can be controlled by using a structure matrix (Pritchard et al. 2000) and/or PCA (Patterson et al. 2006). Cryptic relatedness which is due to recent common ancestry among smaller groups of individuals should also be controlled as this can have a confounding effect similar to that of population structure (Astle and Balding 2009; Myles et al. 2009). With this, we used a mixed model proposed by Yu et al. (2006) and which was successfully implemented for many traits in many crops (Weber et al. 2007, 2008; Casa et al. 2008; Ghavami et al. 2011; Gurung et al. 2011; Mamidi et al. 2011). In a mixed model, **Q** takes only a few axes of variation into account, while the **K** matrix captures the relatedness between each possible pair of individuals in a sample (Astle and Balding 2009). In many cases, a combination of structure and kinship approach has been successful in interpreting the results (Ghavami et al. 2011; Mamidi et al. 2011). Given the weak population structure of the genotypes, lack of significant bottleneck effects and selection, association mapping is feasible for this population of *L. albus*. We selected seed weight, an important trait that is used to breed new varieties. Seed weight for the 122 accessions varied between 11.0 and 52.0 with a mean of 29.77 (±7.5) (Fig. 4). The two markers identified explained 22.69 and 20.5 % variation in the seed weight trait.

Since the power of a mixed model is dependent on phenotype, markers, population structure and relatedness, we tested multiple models that perform better than other models (Flint and Mackay 2009; Atwell et al. 2010; Mamidi et al. 2011). Ideally, the *p* values obtained from a mixed model follow a uniform distribution in a p-p plot (Yu et al. 2006; Mamidi et al. 2011). So we tested 12 different linear regression models and the distribution of the *p* values with that of a uniform distribution and selected the model with the lowest MSD.

It can be argued that with such a low LD, higher number of markers may have been required for association mapping. However, it is observed that the genome-wide estimate of LD might not adequately reflect LD patterns of specific regions (Abdurakhmonov and Abdukarimov 2008). Secondly, we used 892 AFLP markers (MAF >0.05) from 2277 markers that are supposed to be randomly distributed across the 25 pairs of chromosomes on a genome slightly larger in size to *A. thaliana* (Hajdera et al. 2003).

Lupinus albus or white lupin is recognized for its wide adaptation, high protein (33–47 %) and oil (6–13 %) depending on the varieties and genotypes (Huyghe 1997; Petterson et al. 1997). However, it is still a relatively unadopted crop. AFLPs are valuable for such non-model species as the large number of loci required for whole genome scans can be easily produced and are very powerful for intra-species genetic diversity and population structure analysis. Efficiently accessing genetic diversity in the germplasm collections could enhance their use in crop improvement programs. The overall genetic diversity and lack of any structure are indicative of the suitability of this *Lupin* germplasm collection for association mapping and developing gene based SNP markers for study of associations without waiting for the development of purpose-created populations.

Acknowledgments This project was supported by USDA-NIFA-1890 Institutions Capacity Building Grants Program funds allocated to Virginia State University (H. L. Bhardwaj as PI/PD). The AFLP lab work was done at Institute for Advanced Learning and Research, Danville, VA by MJI and RA. Contribution of Virginia State University, Agricultural Research Station Journal Article Series Number 284. The use of any trade names or vendors does not imply approval to the exclusion of other products or vendors that may also be suitable.

References

Abdurakhmonov IY, Abdukarimov A (2008) Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics 2008:574927

Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci 24:451–471

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631

Bhardwaj HL (2006) Muskmelon and sweet corn production with legume cover crops. HortScience 41:1222–1225

Bhardwaj HL, Hamama AA, Merrick LC (1998) Genotypic and environmental effects on lupin seed composition. Plant Foods Hum Nutr 53:1–13

Bhardwaj HL, Hamama AA, van Santen E (2004) White lupin performance and nutritional value as affected by planting date and row spacing. Agronomy J 96:580–583

Blair MW, Chaves A, Tofiño A, Calderón JF, Palacio JD (2011) Extensive diversity and inter-gene pool introgression in a world-wide collection of indeterminate snap bean accessions. Theor Appl Genet 120:1381–1391

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633

Campbell D, Duchesne P, Bernatchez L (2003) AFLP utility for population assignment studies: analytical investigation and empirical comparison with microsatellites. Mol Ecol 12:1979–1991

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Sci 48:30–40

Chang YK, Veilleux RE, Iqbal MJ (2009) Analysis of genetic variability among Phalaenopsis species and hybrids using amplified fragment length polymorphism. J Am Soc Hort Sci 134:58–66

Cowling WA, Huyghe C, Swiecicki W, Gladstones JS, Atkins CA, Hamblin J (1998) Lupin breeding. In: Gladstones JS, Atkins C, Hamblin J (eds) Lupins as crop plants: biology, production and utilization. Cambridge University Press, Cambridge, pp 93–120

Croxford AE, Rogers T, Caligari PDS, Wilkinson MJ (2008) High resolution melt analysis to identify and map sequence tagged site anchor points onto linkage maps: a white lupin (*Lupinus albus*) map as an exemplar. New Phytol 180:594–607

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7:574–578

Flint J, Mackay TFC (2009) Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res 19:723–733

Flint-Garcia SA, Thornsberry JM, Iv B (2003) Structure of linkage disequilibrium in plants. Ann Rev Plant Biol 54:357–374

Ghavami F, Elias EM, Mamidi S, Ansari O, Sargolzaei M, Adhikari T, Mergoum M, Kianian SF (2011) Mixed model association mapping for Fusarium head blight resistance in Tunisian-derived durum wheat populations. G3-Genes Genomes Genet 1:209–218

Gilbert JE, Lewis RV, Wilkinson MJ, Caligari PDS (1999) Developing an appropriate strategy to assess genetic variability in plant germplasm collections. Theor Appl Genet 98:1125–1131

Gladstones JS (1998) Distribution, origin, taxonomy, history and importance. In: Gladstones JS, Atkins C, Hamblin J (eds) Lupins as crop plants—biology, production and utilization. Cambridge University Press, Cambridge, pp 1–40

Green AG, Brown A, Oram R (1980) Determination of outcrossing rate in a breeding population of *Lupinus albus* L. (White Lupin). Z Pflanzenzeucht 84:181–191

Gurung S, Mamidi S, Bonman JM, Jackson EW, Del Río LE, Acevedo M, Mergoum M, Adhikari TB (2011) Identification of novel genomic regions associated with resistance to *Pyrenophoratritici-repentis* races 1 and 5 in spring wheat landraces using association analysis. Theor Appl Genet. doi:10.1007/s00122-011-1645-1 (Online First)

Hajdera I, Siwinska D, Hasterok R, Maluszynska J (2003) Molecular cytogenetic analysis of genome structure in *Lupinus angustifolius* and *Lupinus cosentinii*. Theor Appl Genet 107:988–996

Hamama AA, Bhardwaj HL (2004) Phytosterols, triterpinealcohols, and phospholipids in seed oil from white lupin. J Am Oil Chem Soc 81:1039–1044

Hamblin MT, Buckler ES, Jannink JL (2011) Population genetics of genomics-based crop improvement methods. Trends Genet 27:98–106

Hammer K, Arrowsmith N, Gladis T (2003) Agrobiodiversity with emphasis on plant genetic resources. Naturwissenschaften 90:241–250

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Hill GD, van Santen E (2006) 267. International Lupin Association, Poland, pp 267–279

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9:1322–1332

Huyghe C (1997) White lupin (*Lupinus albus* L.). Field Crops Res 53:147–160

Le Couviour F, Faure S, Poupard B, Flodrops Y, Dubreuil P, Praud S (2011) Analysis of genetic structure in a panel of elite wheat varieties and relevance for association mapping. Theor Appl Genet 123:715–727

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, Psychotriaofficinalis (Rubiaceae). Am J Bot 82:1420–1425

Mamidi S, Chikara S, Goos RJ, Hyten DL, Annam D, Moghaddam SM, Lee RK, Cregan PB, McClean PE (2011) Genome-wide association analysis identifies candidate genes associated with iron deficiency chlorosis in soybean. Plant Genome 4:1–11

Meudt HM, Clarke AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. Trends Plant Sci 12:106–117

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202

Nei M (1973) Analysis of gene diversity in subdivided populations. PNAS 70:3321–3323

Neves-Martins JM (1986) Pattern types of *L. albus* populations from Portugal after multivariate analysis. In: Department of Agriculture Western Australia, South Perth. Proceedings of the 4th international lupin conference, 15–22 August 1986, Geraldton, Australia, p 282

Neves-Martins JM (1994) Characterization in *Lupinus albus* and *Lupinus mutabilis* populations types. In: Neves-Martins JM, Beirao da Costa ML (eds) Advances in lupin research. Proceedings of the 7th international lupin conference, 18–23 April 1994. Technical University of Lisbon, Evora, Portugal, pp 65–69

Noffsinger SL, van Santan E (2005) Evaluation of *Lupinus albus* L. germplasm for the southeastern USA. Crop Sci 45:1941–1950

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:190–193

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190

Pazy P, Heyn CC, Herrnstadt I, Plitmann U (1977) Studies in populations of the Old World *Lupinus* species. I. Chromosomes of the east-mediterranean lupines. Israel J Bot 26:115–127

Petterson DS, Sipsas S, Mackintosh JB (1997) The chemical composition and nutritive value of Australian grain legumes, 2nd edn. Grains Research and Development Corporation, Canberra

Phan HTT, Ellwood SR, Adhikari K, Nelson MN, Oliver RP (2007) The first genetic and comparative map of white lupin (*Lupinus albus* L.): identification of QTLs for anthracnose resistance and flowering time, and a locus for alkaloid content. DNA Res 14:59

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945

Rafalski JA (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174–180

Rodriguez-Medina C, Atkins CA, Mann AJ, Jordan ME, Smith PMC (2011) Macromolecular composition of phloem exudate from white lupin (*Lupinus albus* L.). BMC Plant Biol 11:36

Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MAM, Hillel J, Mäki-Tanila A, Tixier-Boichard M, Vignal A (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics 159:699

Rossi M, Bitocchi E, Bellucci E, Nanni L, Rau D, Attene G, Papa R (2009) Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. Evol Appl 2:504–522

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644

Shirasawa K, Monna L, Kishitani S, Nishio T (2004) Single nucleotide polymorphisms in randomly selected genes among japonica rice (*Oryza sativa* L.) varieties identified by PCR-RF-SSCP. DNA Res 11:275–283

Simpson MJA (1986) Geographical variation in *Lupinus albus* L. II: northwest Spain, the Nile Valley, the Balkans and Turkey. Zeitschriftfür Pflanzenzüchtung 96:241–251

Spooner D, van Treuren R, de Vicente MC (2005) Molecular markers for genebank management. IPGRI Technical Bulletin No. 10. International Plant Genetic Resources Institute, Rome

Stich B, Melchinger AE (2009) Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. BMC Genomics 10:94

Sun G, Zhu C, Kramer MH, Yang S-S, Song W, Piepho H-P, Yu J (2010) Variation explained in mixed-model association mapping. Heredity 105:333–340

Tian L, Peel GJ, Lei Z, Aziz N, Dai X, He J, Watson B, Zhao PX, Sumner LW, Dixon RA (2009) Transcript and proteomic analysis of developing white lupin (*Lupinus albus* L.) roots. BMC Plant Biol 9:1

Upadhyaya HD, Furman BJ, Dwivedi SL, Udupa SM, Gowda CLL, Baum M, Crouch JH, Buhariwalla HK, Singh S (2006) Development of a composite collection for mining germplasm possessing allelic variation for beneficial traits in chickpea. Plant Genet Resour 4:13–19

van Santen C, Noffsinger SL, van Santen E (2006) Low-tech breeding approach to develop low-return cultivars. In: van Santen E, Hill GD (eds) Mexico, where old and new world lupins meet. Proceedings of the 11th international lupin conference, 4–9 May 2005, Guadalajara, Jalisco, Mexico, Canterbury, New Zealand, pp 80–83

Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP—a new technique for DNA-fingerprinting. Nucleic Acids Res 23:4407–4414

Weber A, Clark RM, Vaughn L, de Jesus Sánchez-Gonzalez J, Yu J, Yandell BS, Bradbury P, Doebley J (2007) Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp. *parviglumis*). Genetics 177:2349

Weber AL, Briggs WH, Rucker J, Baltazar BM, de Jesus Sánchez-Gonzalez J, Feng P, Buckler ES, Doebley J (2008) The genetic architecture of complex traits in teosinte (*Zea mays* ssp. *parviglumis*): new evidence from association mapping. Genetics 180:1221–1232

Wells HD, Forbes I, Burns R, Miller JD, Dobson J (1980) Registration of Tifwhite-78 white lupine (Reg. No. 6). Crop Sci 20:824

Wolko B, Clements JC, Naganowska B, Nelson MN, Yang H (2011) *Lupinus*. In: Kole C (ed) Wild crop relatives: genomic and breeding resources, legume crops and forages. Springer, Berlin, pp 153–206

Wong A, Forbes MR, Smith ML (2001) Characterization of AFLP markers in damselflies: prevalence of codominant markers and implications for population genetic applications. Genome 44:677–684

Woodhead M, Russell J, Squirrell J, Hollingsworth PM, Mackenzie K, Gibby M, Powell W (2005) Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. Mol Ecol 14:1681–1695

Yeh FC, Yang RC, Boyle T (1999) POPGENE version 1.32.Microsoft Window-based freeware for population genetic analysis. http://www.ualberta.ca/~fyeh/index.html

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhang LY, Liu DC, Guo XL, Yang WL, Sun JZ, Wang DW, Sourdille P, Zhang AM (2011) Investigation of genetic diversity and population structure of common wheat cultivars in northern China using DArT markers. BMC Genet 12:42. doi: 10.1186/1471-2156-12-42

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4

Zohary D, Hopf M (2000) Domestication of plants in the Old World: the origin and spread of cultivated plants in West Asia, Europe, and the Nile Valley. Oxford University Press, USA