

Genetic diversity, seed size associations and population structure of a core collection of common beans (*Phaseolus vulgaris* L.)

Matthew W. Blair · Lucy M. Díaz · Hector F. Buendía · Myriam C. Duque

Received: 18 October 2008 / Accepted: 11 May 2009 / Published online: 18 August 2009
© Springer-Verlag 2009

Abstract Cultivated common bean germplasm is especially diverse due to the parallel domestication of two gene-pools in the Mesoamerican and Andean centers of diversity and introgression between these gene pools. Classification into morphological races has helped to provide a framework for utilization of this cultivated germplasm. Meanwhile, core collections along with molecular markers are useful tools for organizing and analyzing representative sets of these genotypes. In this study, we evaluated 604 accessions from the CIAT core germplasm collection representing wide genetic variability from both primary and secondary centers of diversity with a newly developed, fluorescent microsatellite marker set of 36 genomic and gene-based SSRs to determine molecular diversity and with seed protein analysis to determine phaseolin alleles. The entire collection could be divided into two gene-pools and five predominant races with the division between the Mesoamerica race and the Durango–Jalisco group showing strong support within the Mesoamerican gene-pool and the Nueva Granada and Peru races showing less diversity overall and some between-group admixture within the Andean

gene-pool. The Chile race could not be distinguished within the Andean gene-pool but there was support for the Guatemala race within the Mesoamerican gene-pool and this race was unique in its high level of diversity and distance from other Mesoamerican races. Based on this population structure, significant associations were found between SSR loci and seed size characteristics, some on the same linkage group as the phaseolin locus, which previously had been associated with seed size, or in other regions of the genome. In conclusion, this study has shown that common bean has very significant population structure that can help guide the construction of genetic crosses that maximize diversity as well as serving as a basis for additional association studies.

Introduction

Common bean (*Phaseolus vulgaris*) is the third most important grain legume in the world and is the most important food crop among these, produced over an area of 18 million hectares with large regions of production in Latin America and Eastern and Southern Africa, where the crop is critical to nutritional security and income generation (Broughton et al. 2003). Cultivated common beans originated in two centers of diversity giving rise to two gene-pools: the Mesoamerican gene-pool from Central America and Mexico and the Andean gene-pool from the Andes mountains of South America (Singh et al. 1991a). The differences between Mesoamerican and Andean gene pools include seed size, plant morphology (Singh et al. 1991b), phaseolin (seed storage protein) patterns (Gepts et al. 1986) and alleles of various types of molecular markers including microsatellites (Blair et al. 2006a, 2007; Díaz and Blair 2006). Cultivated bean gene pools have further been divided into races according to morphological criteria and

Communicated by D. Hoisington.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-009-1064-8) contains supplementary material, which is available to authorized users.

M. W. Blair · L. M. Díaz · H. F. Buendía · M. C. Duque
Centro Internacional de Agricultura Tropical (CIAT),
Apartado Aéreo 6713, Cali, Colombia, South America

M. W. Blair (✉)
International Center for Tropical Agriculture (CIAT),
1380 N.W. 78th Ave, Miami, FL 33126, USA
e-mail: m.blair@cgiar.org

agro-ecological adaptation (Singh et al. 1991a, b) where the term ‘race’ is used to denote a group of related genotypes and where members of each race have distinctive and specific physiological, agronomic, biochemical and molecular characteristics and differ from other races in the allelic frequencies at specific isozyme or microsatellite loci (Singh et al. 1991c; Blair et al. 2007; Díaz and Blair 2006). The original six races identified by Singh et al. (1991a) were Durango, Jalisco and Mesoamerica in the Mesoamerican gene pool; and Chile, Nueva Granada and Peru in the Andean gene pool. The race concept for Mesoamerican common beans was further refined by Beebe et al. (2001) who provided molecular evidence for a seventh race, termed Guatemala and by Díaz and Blair (2006) and Blair et al. (2006a) who found the Durango and Jalisco races to be grouped together rather than as separate races.

Microsatellite markers based on simple sequence repeat loci have been particularly useful in studying population structure given their higher polymorphism level compared to other markers and better resolution in diversity studies (Akkaya et al. 1992; Pejic et al. 1998; Cho et al. 2000; Díaz and Blair 2006; Blair et al. 2007). Large-scale studies, with more than 200 genotypes at a time, initially were conducted principally in US or European institutions for cereal crops such as barley, *Hordeum vulgare* L. (Malysheva-Otto et al. 2006), maize, *Zea mays* L. (Liu et al. 2003), pearl millet, *Pennisetum glaucum* (L.) R. Br. (Oumar et al. 2008) and rice, *Oryza sativa* L. (Garris et al. 2005). More recently, the germplasm collections held in the CG centers for international agricultural research have been analyzed, mainly as part of the Generation challenge program on “Unlocking genetic diversity” (<http://www.generationcnp.org>) and these large-scale studies are shedding light on population structure of the most important germplasm for the developing world (Thomson et al. 2007; Warburton et al. 2008). Microsatellite methods based on fluorescent labeling and automated band calling with precise, software-based allele detection are considered the most accurate way of genotyping with this type of marker (Diwan and Cregan 1997; Blair et al. 2002; McCouch et al. 1997). In addition, fluorescent labeling of microsatellites and organization of these into various dye color panels has advantages in allowing multiplexing during band separation, with advantages for high throughput genotyping and simultaneous analysis of multiple loci (Reed et al. 1994; Schwengel et al. 1994; Coburn et al. 2002; Tommasini et al. 2003). Multiplexing can also be carried out in the PCR amplification step with mixtures of the appropriate primers (Hayden et al. 2008). Current technology allows for capillary-based separation of microsatellite bands with four color panels of four individual markers labeled with different dyes evaluated through a single capillary with a separate size standard, occupying a fifth detection wavelength range, allowing for precise band size estimates (Coburn et al. 2002).

Association of phenotypic traits with microsatellite alleles has been found to be practical in several crop species, although SNP markers, sequence haplotypes or high density marker sets such as DART or AFLPs are usually preferred for association mapping (Flint-Garcia et al. 2005; Rostoks et al. 2006; Stick et al. 2006; Crossa et al. 2007). The number of markers used for association genetics varies although there is general agreement that more are required in out-crossing species or widely diverse germplasm samples where linkage disequilibrium is low when compared with the number of markers required for less diverse germplasm sets in inbreeding species where linkage disequilibrium is high (Buckler and Thornsberry 2002). Estimates of linkage disequilibrium are usually precursors to actual association analysis, which requires phenotypic trait evaluation in addition to a large genotypic data set. However, the usual process of constituting a germplasm panel for genotyping that is later phenotyped is inefficient. An alternative strategy is to genotype previously phenotyped collections such as those held in germplasm banks or assembled into core collections (Brown 1989). In any case, the phenotypic traits evaluated should be of high heritability to find associations that are not influenced by environmental variability (Flint-Garcia et al. 2005).

In common bean, the evaluation of microsatellite marker diversity has been useful for analysis of race structure as described in Blair et al. (2006a, 2007) and Díaz and Blair (2006); however, in these cases gel-based assays were used and this limited the number of genotypes that could be evaluated. Therefore, there has been a need for the analysis of larger and more comprehensive sets of germplasm such as core collections. In the case of common beans, two core collections exist one for 1,441 individuals at CIAT that has been extensively phenotyped (Beebe et al. 1997; Islam et al. 2002, 2004) and another prepared by the USDA, which has been partially phenotyped (Miklas et al. 1999). Correspondingly, there has been the need for the development of a high-throughput method for microsatellite diversity assessments using fluorescently labeled markers in common bean. To date, only a single study has been conducted with this type of marker in this crop species, evaluating only a limited number of genotypes with a small set of cDNA-based microsatellites (Masi et al. 2003). Meanwhile, no attempts have been made to perform association studies with microsatellites, although one study did attempt to associate specific RAPD markers with various traits in part of the CIAT core collection (Islam et al. 2004).

In this study, our overall goal was to determine population structure in a large-scale international collection of common bean accessions from the CIAT core collection using fluorescent microsatellites and conduct association analysis with seed phenotypic characteristics. Our specific objectives were (1) to assess genetic diversity in a group of

604 genotypes from the CIAT germplasm and core collections with microsatellite markers and phaseolin evaluation; (2) to test a set of fluorescent microsatellite markers for capillary-based genotyping based on previous polymorphism evaluation of a large set of cDNA (gene-based or coding) microsatellites and genomic (non-coding) microsatellites, (3) to determine if race structure was well defined in this core collection when compared with our previous studies and if diversity was similar in primary and secondary centers of diversity and (4) to conduct a test for association of seed size, length, width and height with the markers evaluated, comparing associations with phaseolin and with individual microsatellite loci in each subpopulation.

Materials and methods

Plant material

A total of 604 genotypes were evaluated in this study, including totals of 304 Mesoamerican gene pool accessions and 300 Andean gene pool accessions, all selected from the CIAT germplasm collection. Of these, all the Andean genotypes were from the core collection while the Mesoamerican genotypes included 23 genotypes from the general collection that had been previously evaluated by Díaz and Blair (2006). The majority of the accessions were from the primary center of diversity especially the countries of Mexico (179 genotypes) and Peru (180 genotypes), while the remainder were from Argentina (10), Brazil (26), Bolivia (11), Chile (8), Colombia (32), Costa Rica (16), Cuba (2), Dominican Republic (2), Ecuador (36), El Salvador (6), Germany (2), Guatemala (60), Haiti (9), Honduras (9), Nicaragua (6) and the United States (3). One accession each was from Australia, Burundi, France, Jamaica, Malawi, Rwanda and Uganda. The accessions had the following phaseolin alleles: S, Sb, Sd, B and M for the Mesoamerican genotypes; T, C, H and A for the Andean genotypes. Control genotypes included the Andean genotypes, Calima/G4494 and Chaucha Chuga/G19833, as well as the Mesoamerican genotypes, ICA Pijao/G5773 and Dorado/DOR364 (with common name and germplasm entry or advanced line name listed in each case). Seed samples for the core collection are maintained by the bean project at CIAT and were originally from the Genetic Resource Unit (<http://isa.ciat.cgiar.org/urg/main.do>). Selection from the CIAT core collection was based on the random sampling of genotypes evaluated by Beebe et al. (2000, 2001) while the core collection itself was carefully chosen to represent the broadest possible diversity including modern varieties, landraces and a few breeding lines which represent all agro-ecological zones where the crop is grown (Tohme et al. 1995; Beebe et al.

1997). Of the 1,441 genotypes in the core collection, 41% were evaluated in our study.

DNA extraction and phaseolin evaluation

DNA extraction involved germinating 10 seeds selected at random from each accession. The seed was scarified to ensure uniform germination in the laboratory and pre-germinated in darkness on germination paper. The first trifoliate leaves of 8-day old seedlings were collected and ground in liquid nitrogen for DNA extraction with the method of Afanador et al. (1993). DNA was re-suspended in TE buffer and DNA quality was evaluated on 0.8% agarose gels followed by quantification with Hoescht H 33258 dye on a Hoefer DyNA fluorometer (DNA Quant™ 200). DNA was diluted to 10 ng/ml for further experiments. Common bean is a predominantly self-pollinating species and the genotypes in the core collection are maintained as single seed types; therefore, this project did not expect to find high levels of observable heterozygosity. Phaseolin evaluations were as performed by Durán et al. (2005) with a method developed by the Genetic Resource Unit of CIAT for total seed protein.

Fluorescent microsatellite analysis

Microsatellite amplification used a fluorescent marker kit that we developed as part of this project and included microsatellites selected from those of Yu et al. (2000), Gaitán et al. (2002) and Blair et al. (2003). The selection of markers was based on the evaluation of 150 microsatellites previously tested for polymorphism and PCR conditions (Blair et al. 2006a, b). The kit included a total of nine-four-color marker panels for the analysis of a total of 36 individual microsatellite markers. These microsatellites were distributed across the genome and covered 9 of the 11 chromosomes of common bean genome according to map positions from Blair et al. (2003). One of the markers, BM188 detected two loci in separate size ranges. To sample both fast-evolving genomic sequences and slower evolving gene sequences, microsatellites were selected that had been designed and tested from non-gene coding as well as gene coding regions of the genome, termed genomic and gene-based microsatellites, respectively. The 20 genomic markers included AG, BM or GATS series markers developed by Gaitán et al. (2002) plus BMd56 from Blair et al. (2003). The 16 gene-based microsatellites used were mostly BMd series markers from Blair et al. (2003) or PV series markers from Yu et al. (2000). Forward primers for each of the microsatellites were 5'-end labeled with 6-FAM, NED, PET or TET fluorescent labels as indicated in the supplementary table and used with unlabeled reverse primers in PCR amplification. Amplification conditions

were as recommended in Blair et al. (2006a) using a standard thermocycle of 95°C hot start for 3 min, followed by 28 cycles of 95°C denaturation for 40 s, 55°C annealing for 40 s and 72°C extension for 1 min. A 1-h extension at 72°C was used post-thermocycling. PCR reactions included 1× PCR buffer (10 mM of Tris–HCl pH 8.8, 50 mM of KCl, 0.1% of TritonX-100), 1.5 mM MgCl₂, 0.2 mM dNTPs, 0.3 nM of each primer and 1 U of Taq polymerase (Promega). A total of 2 µl from each of the fluorescent PCR products corresponding to the appropriate markers were mixed into multiplex panels as listed. This mixture was diluted 1:1 with dH₂O and 0.5 µl of the solution was then prepared in 9 µl of formamide with 0.06 µl of Genescan LIZ500 size standard and 0.44 µl of dH₂O, which was then run on an ABI PRISM 3730 fragment analysis system (Applied Biosystems) at the Cornell University Biotechnology Resource Center.

Data analysis

Allele sizes (in base pairs) were estimated by comparing the fragment peaks with the internal size standard calculated with GeneMapper v. 3.7 software using the default method for band calling with microsatellites (cut-off value, plus A ratio/distance, stutter ratio/distance, signal level, heterozygote balance, peak morphology and pull up/cross talk peaks), the second order least squares size calling method and the expected repeat size. Electropherograms were also reviewed visually in Genemapper software to ensure that clear peaks were found for the expected marker sizes, that stutter peaks were not called and to assess that a proper selection of multiplex markers had been made.

Raw allele size calls were then binned to assign a whole integer allele value using the software program AlleloBin (<http://www.icrisat.org/gt-bt/biometrics.htm>) based on a least-square minimization algorithm of Idury and Cardon (1997), which transforms fragment sizes based on migration into a binned values with sizes in base pairs. Binned allele information was databased at <http://gcpcr.grinfo.net/index.php> and used for a principal coordinate analysis (PCoA) with the goal of determining sub-populations. This was based on the proportion of shared alleles estimated with the IML module of the software program Statistical Analysis Systems v. 9.1.3 (SAS Institute 1996) and on the formula $PS = \sum PS_i / n$, where i each locus, n total number of loci and PS_i proportion of shared alleles at locus i . PowerMarker v. 3.25 (Liu and Muse 2005) was then used to evaluate allele frequency, number of alleles, percentage of polymorphic alleles, expected heterozygosity (H_e /Nei 87), observed heterozygosity (H_o) and polymorphism information content (PIC). Relationships between sub-populations were graphed in three dimensions using software program NTSYSpc 2.1 (Rohlf 2002).

In addition, Wright statistics and molecular analysis of variance (AMOVA) were determined with Arlequin v. 3.11 (Excoffier 2007) and clusters were analyzed for phylogenetic relationships with PowerMarker using Nei's (1973) coefficient. A consensus tree was drawn based on 1,000 bootstraps with the subprogram CONSENSE in Phylip (Felsenstein 1993). Finally, the number of populations (K) was evaluated with both STRUCTURE (Pritchard et al. 2000), which determines a Q matrix of population relatedness and with the similar program InStruct (Gao et al. 2007), which is thought to provide a better estimate of population number in an inbreeding species such as common bean. Population structure was determined assuming an admixture model with $K = 2$ to $K = 15$ for the first of these programs, which was run locally and from $K = 2$ to $K = 10$ for the second program, which was run remotely at <http://cbsuapps.tc.cornell.edu/InStruct.aspx>. A correlation model for allele frequency and 100,000 burn-ins with 200,000 iterations was used for InStruct. The results of both programs were visualized with the software DISTRUCT (Rosenberg 2002).

Phenotypic data and association analysis

The following seed characteristics were measured on ten seeds of each genotype with a caliper with millimeter (mm) adjustment: seed length from end to end, and seed width and height from side to side at the hilum when the seed lays flat on a horizontal surface. In addition the one-hundred seed weight (100 SW) was measured in grams with an analytical balance. Seed color had already been scored on all accessions according to a CIAT developed scale based on commercial class color classifications (CIAT 1987) and photographs of the genotypes are available at <http://isa.ciat.cgiar.org/urg/main.do>. Phenotypic data was used in a structured association analysis with a mixed linear model, taking into account population substructure and races as determined by this study, with the software program Tassel v. 2.0 (<http://www.maizegenetics.net/>; Buckler et al. 2007). Significance of probability (P) values for association tests were determined with 10,000 permutations using the same software program.

Results

Allele number and genetic diversity

The total number of alleles identified in the entire study was 679 with an average of 18.4 alleles per locus with all the markers analyzed being polymorphic having from 2 to 76 alleles each (Table 1). The marker PV-at001 (confirmed through re-evaluation and visual inspection of electropherograms)

Table 1 Genetic diversity values for microsatellites evaluated across the 604 genotypes of common bean

Marker	Andean		Mesoamerican		All genotypes			
	No. alleles	Exp. Het.	No. alleles	Exp. Het.	No. alleles	Exp. Het.	Obs. Het.	PIC
Gene-based ($n = 16$)								
BMd01	16	0.876	6	0.476	16	0.799	0	0.782
BMd02	3	0.062	4	0.370	5	0.494	0.042	0.384
BMd08	4	0.434	7	0.594	7	0.545	0.019	0.454
BMd15	7	0.109	10	0.607	11	0.52	0.297	0.458
BMd16	4	0.056	6	0.179	6	0.523	0.04	0.412
BMd17	3	0.102	5	0.599	5	0.614	0.075	0.548
BMd18	3	0.126	5	0.263	5	0.198	0.063	0.189
BMd20	7	0.137	6	0.715	8	0.605	0.048	0.558
BMd46	2	0.046	3	0.033	3	0.505	0	0.382
BMd47	4	0.146	4	0.135	6	0.515	0.007	0.399
BMd51	2	0.007	2	0.007	2	0.007	0	0.007
PV-ctt001	7	0.495	9	0.743	10	0.753	0.091	0.715
PV-ag003	3	0.302	5	0.470	5	0.547	0	0.446
PV-at003	8	0.393	5	0.383	9	0.541	0.147	0.436
PV-at001	63	0.961	42	0.952	76	0.973	0.047	0.972
PV-cct001	3	0.189	6	0.272	6	0.234	0.005	0.214
Total	139		125		180			
Average	9	0.278	8	0.425	11	0.523	0.055	0.46
Genomic ($n = 21$)								
AG01	6	0.552	8	0.291	10	0.545	0.033	0.48
BM137	43	0.950	34	0.949	45	0.952	0	0.95
BM139	19	0.914	10	0.321	20	0.763	0.051	0.751
BM140	25	0.893	11	0.293	28	0.677	0.071	0.667
BM141	25	0.919	12	0.532	28	0.832	0.121	0.816
BM143	36	0.943	24	0.826	41	0.929	0.077	0.925
BM149	4	0.490	2	0.036	4	0.426	0.02	0.337
BM156	36	0.932	12	0.569	40	0.862	0.087	0.851
BM160	25	0.933	11	0.349	25	0.82	0.053	0.812
BM172	21	0.916	13	0.489	21	0.795	0.054	0.784
BM175	13	0.800	17	0.598	22	0.842	0.048	0.828
BM183	14	0.494	17	0.707	21	0.779	0.04	0.751
BM187	50	0.957	26	0.861	54	0.945	0.071	0.942
BM188A	9	0.191	4	0.492	9	0.5	0.016	0.421
BM188B	15	0.797	19	0.807	19	0.807	0	0.791
BM200	14	0.742	51	0.961	53	0.905	0.042	0.899
BM201	7	0.082	12	0.742	13	0.685	0.035	0.65
BM205	7	0.649	11	0.635	11	0.751	0.03	0.712
BMd56	3	0.024	3	0.426	4	0.264	0.018	0.231
GATs54	3	0.098	3	0.499	4	0.42	0.005	0.336
GATs91	20	0.852	23	0.910	27	0.92	0.078	0.914
Total	395		323		499			
Average	19	0.673	15	0.585	24	0.734	0.045	0.707
Overall total	534		448		679			
Overall average	14	0.502	12	0.516	18	0.643	0.049	0.6

showed the highest number of alleles while the next highest allele numbers were found for BM187 and BM200 with 54 and 53 alleles, respectively. Meanwhile, the next most polymorphic markers were BM137, BM143 and GATs91, and all of the above were genomic microsatellites. In contrast, the gene-based microsatellites were correspondingly less polymorphic, especially BMd51 (bi-allelic), BMd46 (tri-allelic) and BMd02, BMd17, BMd18 and PV-ag003 (each with 5 alleles). Among the least polymorphic genomic markers were GATs54 (bi-allelic), BM149 (tri-allelic) and BMd56 (4 alleles), followed by BM188A (9 alleles), AG01 (10 alleles) and BM205 (11 alleles). A total of 12 markers produced under ten alleles in the case of the gene-based markers while only 4 did so in the case of the genomic markers. Correspondingly, the average number of alleles per marker was higher for the genomic microsatellites (23.8) when compared with the gene-based microsatellites (11.3). It was notable that 12 microsatellites produced over 20 alleles each in the case of the genomic microsatellites while in the case of the gene-based microsatellites only one did. All the markers detected single loci except for one marker (BM188), which detected two loci distinguished by the pattern of stutter band amplification and size range with BM188A having alleles between 130 and 153 bp and BM188B having alleles between 160 and 200 bp. Genetic mapping has confirmed that the two loci are independent and on separate linkage groups (CIAT, unpublished results).

Individual markers detected from 0.264 to 0.952 expected heterozygosity for genomic microsatellites and 0.007 to 0.973 for gene-based microsatellites. Observed heterozygosity was low, ranging from 0.000 to 0.297 and averaging 0.049 across all markers (0.055 for gene based and 0.045 for genomic markers). The markers with more observed heterozygosity were BMd15, PV-at003 (gene-based) and BM141 and BM156 (genomic) and observed heterozygosity could be explained by out-crossing and heterozygous individuals, by residual heterozygosity in breeding lines or by heterogeneity in the genebank accession, although most accessions were selected to be of a single seed type.

PCoA and race structure

The PCoA for the core collection genotypes is shown in Fig. 1 based on genetic similarity values from the proportion of shared alleles. The three-dimensional graphical representation allows the evaluation of population structure and geometric distances between all the genotypes in the study. The most evident subdivision was that of the two major gene pools of common bean, the Andean and the Mesoamerican, which are clearly distinguished as two separate clusters (Fig. 1a). Within each genepool, various

subgroups were found, some of them corresponding to morphological races as described by Singh et al. (1991a). Five of the subgroups corresponded to divisions within the Mesoamerican genepool while four of the subgroups were found in the Andean genepool. The PCoA explained 85.7% of the variability for the germplasm set, with separation of the genepools in the first axis (64.3% of variability explained), separation of Mesoamerican races in the second axis (11.6%) and separation of the Andean races in both the third axis (8.3%) as well as the second axis. One advantage of the PCoA was that the relationships between each of the genotypes could be evaluated and this helps to visualize possible introgression between subgroups, races or genepools. The subgroups within each genepool and their relationship to morphological races are discussed separately below.

Subgroups or clusters from the 3D graph within the Mesoamerican genepool (Fig. 1b) represented the Mesoamerica race (subgroups M1 and M2), the Durango–Jalisco complex (subgroups D–J1 and D–J2) and a third group representing the Guatemala race (subgroup G). These results agree with those of Díaz and Blair (2006) where the Durango and Jalisco races are not substantially distinct from each other but do cluster together separately from the Mesoamerica race. As in this previous study, the Mesoamerica race was made up of small-seeded genotypes of type II, III and IV growth habits, while the Durango–Jalisco complex was made up of medium-seeded genotypes mostly with type III or IV growth habits (Table 2). Similar phaseolin alleles were found distributed among both races (S, Sb, Sd and B) and it was notable that while members of the Mesoamerica race were from many countries in the primary and secondary centers of diversity, the Durango–Jalisco genotypes were predominantly from Mexico with a few additional genotypes from Guatemala.

Within race Mesoamerica, the subgroups M1 and M2 were closely related. Subgroup M1 was a compact group predominantly of small-seeded, black beans and some cream and red beans mostly from Central America (Costa Rica, Guatemala, Honduras, Nicaragua, El Salvador) but also Mexico and Brazil. Growth habits ranged from type II to a few type IV and the representative genotypes in this subrace were Rio Tibagi, Mulatillo and the control genotypes DOR364 and ICA Pijao. Subgroup M2 was similar to M1 in geographic origin and in having small seed size but most of its genotypes had type III growth habit and a greater proportion were red seeded when compared with M1. One genotype in this group had a wild form of phaseolin (M) while all others within race Mesoamerica were cultivated forms (S, Sd and Sb).

The two groups identified within the Durango–Jalisco complex were similar in being predominantly type III or IV growth habit, although D–J2 had slightly more type III

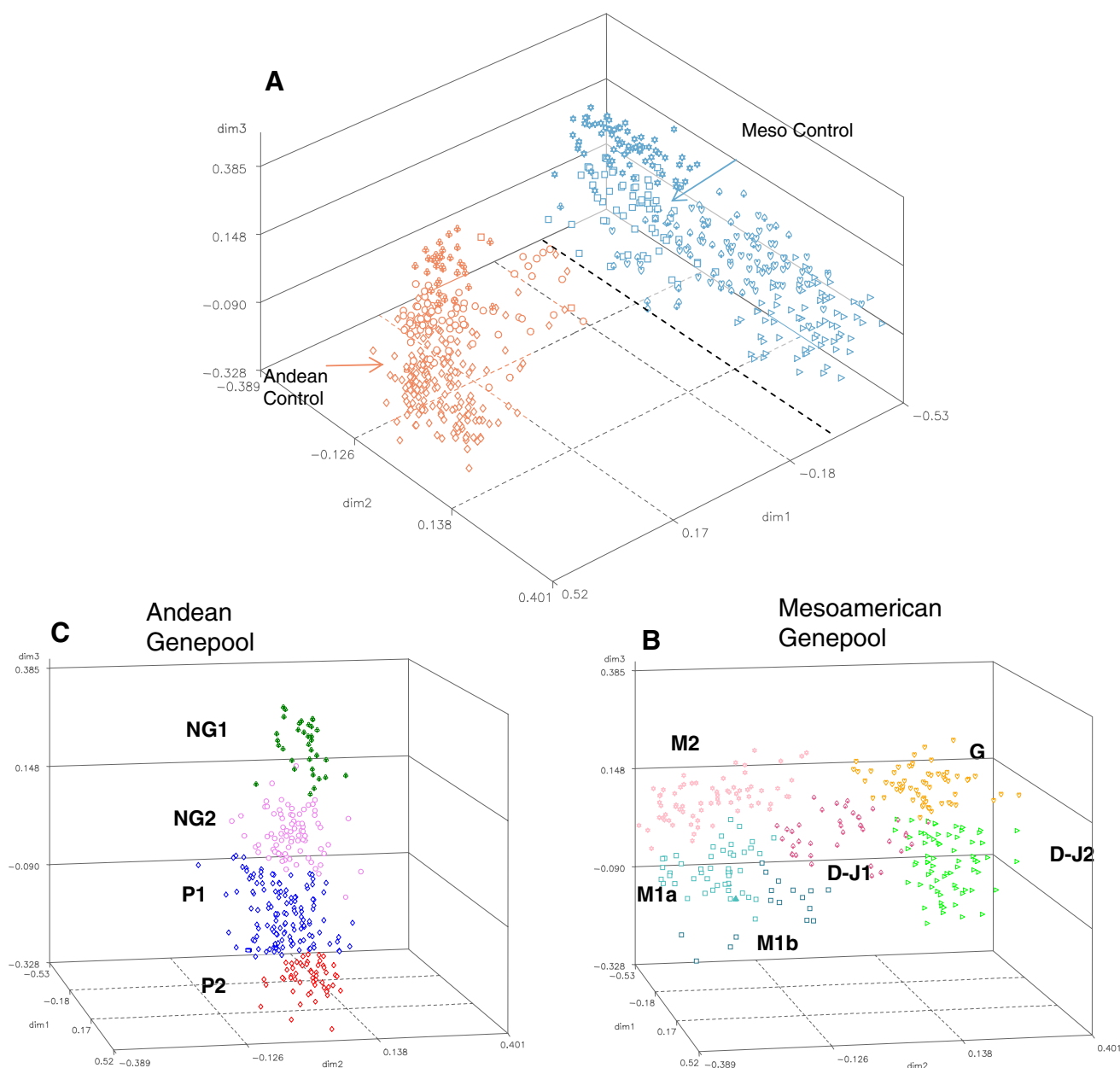


Fig. 1 Three-dimensional principal coordinate analysis based on the analysis of: **a** 604 genotypes of common bean from the core germplasm collection explaining 83.89% of the variation and showing the position of control genotypes for each gene pool, **b** Andean gene pool genotypes divided into subgroups NG1 and NG2 (race Nueva Granada) and sub-

groups P1 and P2 (race Peru) and **c** Mesoamerican gene pool genotypes divided into subgroups D-J1 and D and N-J2 (Durango–Jalisco complex), subgroup G (race Guatemala) and subgroups M1a, M1b and M2 (race Mesoamerica)

genotypes than D-J1, which had more type II genotypes. The largest difference between the two groups was in seed size with D-J1 having more medium and large-seeded genotypes than D-J2. Seed colors were similar in both subgroups with a predominance of lighter shades (creams and yellows) in D-J1. As expected, the majority of the genotypes in both groups were from Mexico, however, in the case of D-J2 some were found outside of Mexico and in the case of D-J1, one genotype was from Africa. In addition to

S type phaseolins, both subgroups contained wild type phaseolins M, M1 and M15.

The final subgroup within the Mesoamerican gene pool consisted of 61 genotypes representing race Guatemala that were predominantly type IV climbing or type III semi-climbing beans and mostly from Guatemala or the Mexican state of Chiapas, which are the two places where the race was predicted to originate according to Beebe et al. (2000). It was notable that while members of the Mesoamerica race

Table 2 Distribution among subgroups of seed size, seed color, seed pattern, growth habit and phaseolin in 604 common bean genotypes

Subgroups ^a	D–J1 (<i>n</i> = 69)	D–J2 (<i>n</i> = 39)	G (<i>n</i> = 61)	M1 (<i>n</i> = 65)	M2 (<i>n</i> = 69)	NG1 (<i>n</i> = 35)	NG2 (<i>n</i> = 80)	P1 (<i>n</i> = 133)	P2 (<i>n</i> = 53)	Total (<i>n</i> = 604)
Growth habit ^b										
I	0	4	1	1	1	20	18	8	0	53
II	18	5	2	31	2	10	13	9	0	89
III	40	21	17	28	49	4	23	19	2	203
IV	11	9	40	5	17	0	20	91	51	244
V	0	0	1	0	0	1	6	6	0	14
Phaseolin ^c										
A	0	0	0	0	0	0	0	0	1	1
B	0	4	9	11	5	0	1	0	0	30
C	0	0	0	0	0	0	1	5	2	8
Ca	0	0	0	0	0	3	2	9	2	16
H	0	1	0	1	0	1	7	23	12	45
KO	0	0	0	0	0	0	0	1	2	3
M	16	8	0	0	1	0	0	0	0	25
M1	0	2	1	0	0	0	0	0	0	3
M15	0	1	0	0	0	0	1	0	0	2
Pa	0	0	0	0	0	0	3	7	2	12
S	25	10	32	34	25	0	1	2	0	129
Sb	19	9	15	8	18	1	2	0	0	72
Sd	8	2	3	10	17	0	0	0	0	40
T	0	2	1	0	2	27	59	84	28	203
Tm	0	0	0	0	0	3	0	0	0	3
To	0	0	0	0	0	0	2	1	4	7
To2	0	0	0	0	0	0	1	1	0	2
NA	1	0	0	1	1	0	0	0	0	3
Seed color ^d										
1	3	4	4	2	0	0	9	27	15	64
2	29	11	12	8	2	6	17	38	15	138
3	22	6	4	5	4	5	14	21	6	87
4	5	2	2	2	4		3	9	4	31
5	2	0	4	4	3	7	7	4	1	32
6	0	3	5	8	22	10	8	11	0	67
7	2	3	3	3	5	2	15	13	9	55
8	0	9	24	32	28	3	2	4	1	103
9	6	1	3	1	1	2	5	6	2	27
Seed pattern ^e										
Bi	0	0	2	0	0	0	5	11	5	23
J	13	1	5	3	7	2	8	17	13	69
M	4	2	3	1	3	16	20	45	14	108
R	7	3	9	3	1	1	7	17	1	49
P	15	12	3	10	7	4	11	5	3	70
Un	30	21	39	48	51	12	29	38	17	285
Seed size ^f										
L	28	1	17	4	1	29	58	111	50	299
M	37	19	36	11	16	5	19	20	3	166
S	4	19	8	50	52	1	3	2	0	139

NA not available

^a Subgroups: D–J Durango–Jalisco, G Guatemala, M Mesoamerica, NG Nueva Granada, P Peru^b Growth habit: I determinate bush, II indeterminate bush, III prostrate bush, IV climbing according to CIAT (1987)^c Phaseolin: analyzed based on known phaseolin standards and the method of Brown et al. (1981) in the Genetic Resource Unit, CIAT^d Seed color: 1 white, 2 cream, 3 yellow, 4 brown, 5 pink, 6 red, 7 purple, 8 black, 9 other^e Seed pattern: Bi bicolor, P pinto or heavy stripes, J stippled, M mottled, Nu Nuña, R striped, Un unpatterned^f Seed size: L large (40 g/100 seed or more), M medium (between 25 and 40 g/100 seed), S small (below 25 g/100 seed) according to CIAT (1987)

were from many countries in the primary and secondary centers of diversity and the Durango–Jalisco genotypes were predominantly from central Mexico, the Guatemala race genotypes were predominantly from this specific region. Seed types for this cluster were variable with most being medium-sized seed with black or cream seed colors. Phaseolin types included S, Sb and finally B, which is an allele that is associated with northern South American genotypes.

The Andean genepool showed four subgroups in the principal correspondence analysis corresponding to two subgroups in race Nueva Granada (NG1 and NG2) and two subgroups in race Peru (P1 and P2). The races and subgroups were distinguished by the third axis of the PCoA (Fig. 1c) with NG2 and P1 intermediate between NG1 and P2 with these two subgroups at the extremes of the distribution of Andean beans and representing the most unique genotypes of each race. In addition, for NG1, NG2 and P1 a certain amount of between-genepool introgression was evident with genotypes that were intermediate between these Andean subgroups and the Mesoamerican genepool. Among the Nueva Granada subgroups, both clusters had medium to large-seeded genotypes of diverse color classes, however, NG1 was predominantly red or pink and mottled, while in the NG2 group cream mottled, yellow and large red or purple beans were more common. The NG1 subgroup was mostly bush beans with the majority having determinate type I growth habit and smaller numbers of type II or type III beans with few type IV climbing beans. The NG2 subgroup meanwhile had more type III and IV growth habit climbing or semi-climbing beans. In terms of geographic origin, subgroup NG1 individuals were from a range of countries from the primary center of diversity (Argentina, Colombia, Ecuador and Peru) and from secondary centers of diversity (Brazil, Mexico and the Caribbean countries of Dominican Republic and Haiti); while subgroup NG2 individuals were mostly from the Andean region.

Meanwhile, among the race Peru subgroups, P1 contained a mix of climbing beans and bush beans of type I, II and III growth habit that were from Colombia, Ecuador and Peru including the Andean control genotype Chaucha Chuga (G19833), while P2 was almost exclusively made up of climbing beans from Peru alone. Seed size and color in race Peru was variable with more white, cream and yellow beans in this race when compared with Nueva Granada. The race also included all of the popping beans from our study (a.k.a ñuñas as they are referred to in the Andes), most of which were clustered with subgroup P2. Inter-genepool introgression was higher for the NG2 subgroup as evidenced by more Mesoamerican phaseolin alleles than among the NG1 or the Peru race individuals. Genotypes showing introgression between the genepools tended to

have smaller seed than the other Andean genotypes and in the PCoA were located between the major genepool clusters. Given our interest in classifying the genotypes into races and subgroups we did not analyze the genotypes showing introgression separately.

Population structure and analyses of molecular variance

Population structure analysis using InStruct (Gao et al. 2007) confirmed the basic divisions found in the PCoA. Figure 2 shows subfigures for various K values up to the number of populations detected in PCoA since each level of subdivision was instructive. For example, $K = 2$ showed the basic division between Andean and Mesoamerican genepool with some Andean admixture in all the Mesoamerican subgroups and some Mesoamerican admixture in all the Andean subgroups except for P2. $K = 3$ showed an additional division within the Mesoamerican genepool for the Durango–Jalisco group and the Mesoamerica race but with admixture occurring in subgroup D–J2. $K = 4$ showed the separation of Nueva Granada and Peru race genotypes but with admixture in the NG2 and P1 subgroups. At $K = 5$, race Guatemala appears as a separate subpopulation but with substantial admixture from race Mesoamerica and the Durango–Jalisco complex; while subgroup D–J2 has admixture from race G and race M. The evaluation at $K = 9$ showed some separation of subgroups although admixture was prevalent between subgroups within each individual race especially within the Andean genepool. It was notable that subgroup NG2 separated from NG1 but was associated with the first part of subgroup P1.

AMOVAs were conducted to determine the variation explained by genepools, races and subgroups as well as to compare variability in the primary versus secondary centers of diversity. Table 3 shows that more variability could be assigned to genepools and races when compared with subgroups but that substantial variability remained within each subpopulation. The comparison between races within genepools were significant in all cases (data not shown), however, more variability was explained by the contrast of races Guatemala, Mesoamerica and the Durango–Jalisco complex (17.2%) than by the contrast of races Nueva Granada and Peru (8.9%). In addition, the comparison between the primary and secondary centers of diversity was also significant ($P < 0.001$) but only explained 6.6% of variability (equivalent to F_{ST} of 0.066) with 87.6% of variability remaining among individuals within the primary and secondary centers and 5.7% within individuals. As expected for an inbreeding species such as common bean, the subpopulations showed a deficit in heterozygotes and were therefore in Hardy–Weinberg disequilibrium. The within population fixation indices (F_{IS})

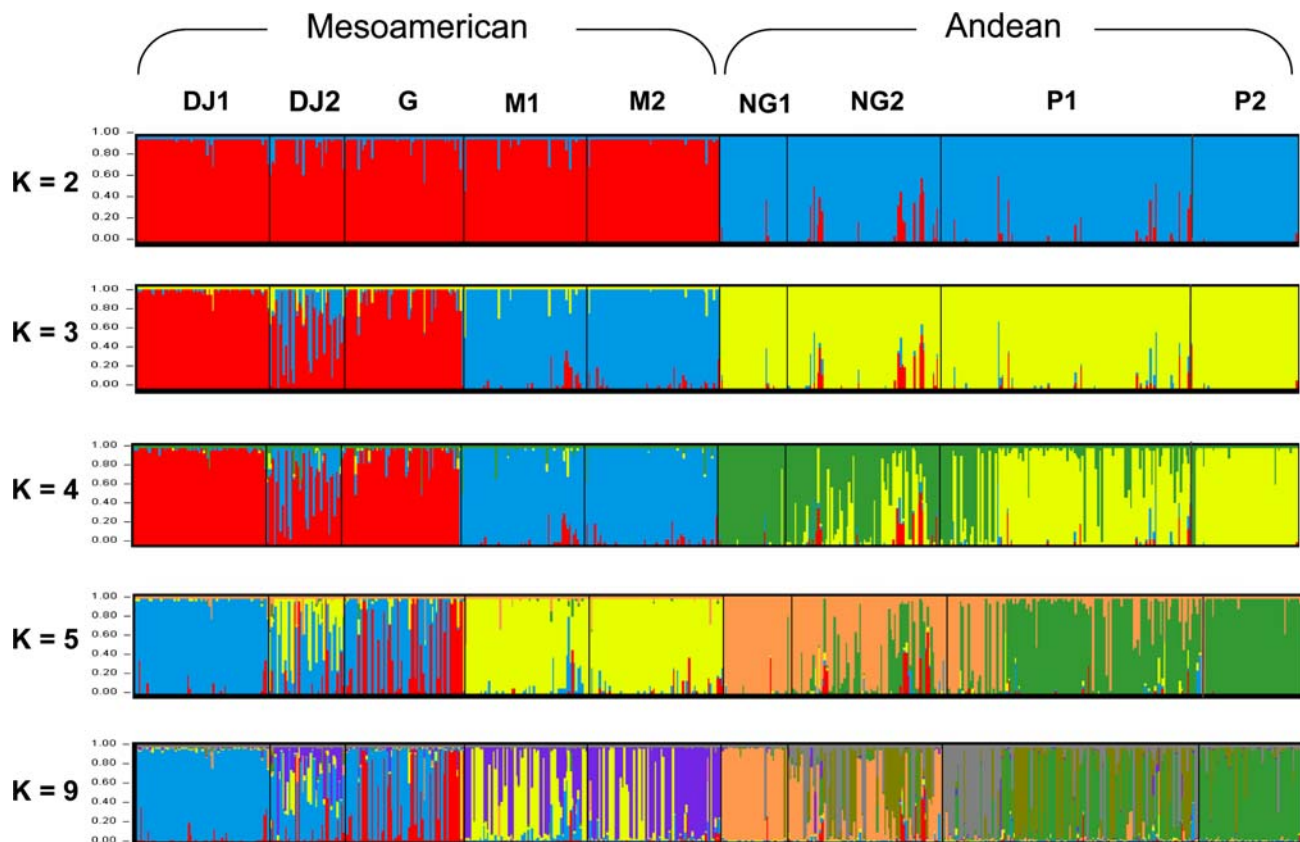


Fig. 2 Structure analysis with $K = 2$ to $K = 5$ and at $K = 9$ populations for the core collection of common bean showing the sub-divisions within each race. Sub-group abbreviations are D–J1 and D–J2 (Durango–Jalisco complex), G (race Guatemala) and M1 and M2 (race

Mesoamerica) within the Mesoamerican genepool; and subgroups NG1 and NG2 (race Nueva Granada), and P1 and P2 (race Peru) within the Andean genepool

were 0.915 to 0.939 across subpopulations but were higher in the Andean genepool (0.946) than in the Mesoamerican genepool (0.839).

Genetic differentiation and diversity parameters for subgroups

Genetic differentiation (F_{ST}) values between subgroups within genepools were generally low ranging from 0.028 to 0.241 among the Andean subgroup comparisons and from 0.044 to 0.268 among the Mesoamerican subgroup comparisons (Table 4). Meanwhile, genetic differentiation was high when comparing subgroups between genepools. The lower genetic differentiation between the Guatemala race and D–J2 may indicate shared ancestry for these two groups; however, this subgroup also was closely related to the two Mesoamerica race subgroups. In the Andean genepool, subgroup NG2 was closely related to both Peru race subgroups. The software program PowerMarker was used to carry out a phylogenetic analysis with the neighbor-joining tree represented as part of Table 4 and showing placement of the Guatemala race relative to the other

Mesoamerican and Andean races. In this dendrogram, race Guatemala is very distinct while the subgroups within each race cluster together (e.g. M1 with M2, NG1 with NG2 and P1 with P2).

Genetic diversity parameters for the subgroups found within the principal correspondence analysis are shown in Table 5. The gene diversity values (H_e) were slightly higher for the Mesoamerican genepool subgroups compared to the Andean genepool subgroups; while within the Andean subgroups, NG2 and P1 were the most diverse and within the Mesoamerican group, subgroups D–J2 and race G were the most diverse. Diversity values were not correlated with the number of genotypes in the subgroup with D–J2 having fewer genotypes but the highest diversity values. Observed heterozygosity within the subgroups averaged 0.049 being very low for M1, M2, NG1, NG2, P1 and P2 (all below 0.050) and slightly higher for race D–J1 (0.058), D–J2 (0.0102) and race G (0.096), which may result from the climbing beans prevalent in these last three subgroups as climbing beans often have higher out-crossing rates (this laboratory, unpublished data). Upon comparing the allele number and expected heterozygosity values for

Table 3 Molecular analyses of variance (AMOVA) for the comparison of primary versus secondary centers of diversity and for separation of genepools, races and subgroups

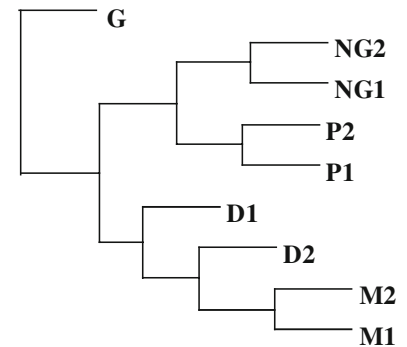
Source of variation	<i>df</i>	Sum of squares	Components of variance	Percentage variation	<i>P</i> value
Primary versus secondary centers ^a					
Among centers	1	84.9	0.39765	6.64	<0.001
Among individuals within centers	602	6,423.2	5.16684	87.56	<0.001
Genepools ^b					
Among genepools	1	1,522.069	2.5063	36.77	<0.001
Among individuals within genepools	602	4,986.069	3.9732	58.3	<0.001
Races ^b					
Among races	4	1,865.356	1.9530	32.57	<0.001
Among individuals within races	599	4,642.782	3.7074	61.83	<0.001
Subgroups ^b					
Among subgroups	8	2,018.9	1.86179	32.09	<0.001
Among individuals within subgroups	595	4,489.2	3.60437	62.12	<0.001
Within individuals	604	203.0	0.33609	5.79	<0.001
Total	1,207	6,711.1	5.80225		

^a Based on country identification and regions in which the wild common bean exists and domestication could have occurred constituting primary centers, while the regions of introduction in the post-Columbian age, such as Africa, Asia or Europe are secondary centers (Islam et al. 2002)

^b Mesoamerican genepool included races Guatemala, Mesoamerica and Durango–Jalisco; while Andean genepool included races Nueva Granada and Peru. Subgroups were identified as described in the text

Table 4 Genetic differentiation based on F_{ST} values and phylogenetic analysis for subgroups identified among 604 genotypes of common bean

	D–J1	D–J2	G	M1	M2	NG1	NG2	P1	P2
D–J1	–								
D–J2	0.133	–							
G	0.138	0.044	–						
M1	0.238	0.077	0.17	–					
M2	0.268	0.079	0.167	0.037	–				
NG1	0.563	0.474	0.453	0.536	0.514	–			
NG2	0.469	0.381	0.365	0.447	0.429	0.063	–		
P1	0.44	0.364	0.349	0.424	0.412	0.129	0.028	–	
P2	0.446	0.364	0.345	0.432	0.433	0.241	0.118	0.058	–



Phylogenetic analysis represented by Neighbor-joining with Nei73 coefficient and subgroup abbreviations: *D–J* Durango–Jalisco, *G* Guatemala, *M* Mesoamerica, *NG* Nueva Granada, *P* Peru

each microsatellite in each subgroup and genepool, we found that genomic microsatellites tended to be more variable in Andean subgroups while gene-based microsatellites were more variable in Mesoamerican subgroups, although this was partly related to the number of genotypes in each subgroup. For example, the subgroup NG1 was low in number of genotypes and in allele number. Among the genomic microsatellites, some markers were more powerful than others in defined subgroups (e.g. BM141, BM156 and BM187 for Andean beans and BM175 and BM200 for Mesoamerican beans) while other markers functioned

equally across all subgroups and both genepools (e.g. BM137 and BM143). The same can be said for specific gene-based markers, so for example BMd01 was more useful in the Andean genepool than in the Mesoamerican genepool and PV-at001 was more useful across all subgroups. Another interesting observation was that phaseolin was approximately as diverse in number of alleles/patterns as the average for all of the microsatellites considering both genomic and gene-based markers for each of the subgroups, the one exception being higher average number of phaseolin alleles than microsatellite alleles in the D–J2 group.

Table 5 Overall observed heterozygosity per subgroup and number of alleles and expected heterozygosity found for genomic and gene-based microsatellite markers and for phaseolin patterns across each subgroup found within the collection of 604 common bean genotypes

Subgroups ^a	D-J1 (n = 69) ^b	D-J2 (n = 39)	G (n = 61)	M1 (n = 65)	M2 (n = 69)	NG1 (n = 35)	NG2 (n = 80)	P1 (n = 133)	P2 (n = 53)
Overall Obs. Het.	0.058	0.102	0.096	0.050	0.035	0.024	0.044	0.028	0.034
Marker	Allele No	Exp. Het.	Allele No	Exp. Het.	Allele No	Exp. Het.	Allele No	Exp. Het.	Allele No
Mean for genomic	7.2	0.440	8.5	0.575	8.0	0.581	9.0	0.490	8.2
Mean for gene-based	4.9	0.371	4.9	0.429	5.4	0.370	4.8	0.366	4.5
Mean for all SSR markers	6.2	0.411	6.9	0.512	6.9	0.490	7.1	0.436	6.6
Phaseolin	4	0.718	9	0.819	6	0.639	5	0.648	6

^a Subgroups: D–J Durango–Jalisco, G Guatemala, M Mesoamerica, NG Nueva Granada, P Peru^b Number of individuals per subgroup indicated as n = total genotypes in that subgroup

Expected heterozygosity was also higher for the phaseolin locus than for the average of the microsatellite loci in each subgroup.

Association of seed phenotypes and marker alleles

Association of marker loci with seed characteristics was evaluated across the genome with the justification that phaseolin and other regions of the genomes have been associated with seed size in several previous quantitative trait loci (QTL) studies. Our interest was in testing associations of both the phaseolin locus (*Phs*) and microsatellite loci with seed characteristics to test whether markers on linkage group b07, where *Phs* is located, were more likely to be associated than markers on other linkage groups and to confirm associations identified in previous studies of QTL for these characteristics. To control for population structure, we performed association analyses independently for each race. As seen in Table 6, the most highly significant ($P < 0.001$) associations found were BMd8 (linkage group b04) for seed weight in races of both gene pools (Mesoamerica, Nueva Granada and Peru; BM183 (b07) for seed weight, width and length also in both gene pools; BM200 (b01) for seed weight in the Durango–Jalisco race; BM156 (linkage group b02), AG01 (b03), BMd16 (b04), *Phs* and BM205 (b07) for seed characteristics in races of the Mesoamerican gene pool and BM143 and PV-cct001 (b02), PV-at001 (b04) and BM187 (b06) for seed characteristics in races of the Andean gene pool. It was notable that Durango–Jalisco presented only one highly significant association while race Guatemala had associations only at $P < 0.05$, but not at the more significant level.

The average seed weight and size were smallest in the Mesoamerica race, intermediate in the Durango–Jalisco complex, and highest in the Andean races as shown at the bottom of Table 6. Interestingly, while accessions from race Peru had heavier seeds on average, these seeds were shorter on average than race Nueva Granada accessions. The difference in weight, therefore, was due to the wider and higher seed dimensions that are found in the race Peru accessions, which typically were rounded while Nueva Granada races were elongated or kidney shaped. To statistically evaluate the differences in variability for seed characteristics in each of the races a Levene's test was used to determine if variances were equal or not in these different subpopulations. Variances were unequal ($P < 0.0001$) for seed weight, length and height but were equal for seed width ($P < 0.05$). The most diversity for seed length was found in races Nueva Granada and Peru, while seed weight was very diverse for race Peru. Mesoamerican races had less diversity for seed weight, height and length than Andean races as reflected in the lower standard deviations.

Table 6 Associations between microsatellite or phaseolin alleles and seed size traits in common bean races according to population structure analysis

Linkage group ^a	Durango–Jalisco				Mesoamerica				Nueva Granada				Peru			
	100S ^b	W	H	L	100S	W	H	L	100S	W	H	L	100S	W	H	L
B01																
BM200	*** ^c			*	*						*		*	*		
PV-ag003								*								
B02																
BM139					*				*						***	*
BM143								*		*	***		***	***		
BM156					***					*	*			*		
BMd02									*		*		*			
BMd17					*					*	*		*			
BMd18							*				*					*
BMd47										*	*					
PV-cct001							*		***						***	*
B03																
AG01					***											
BM172					*									*		
BMd01			*		*											*
GATs54					*											
B04																
BM140					*						*		*	*		
BMd08					***			*	***						***	*
BMd15					*						*					
BMd16					***										*	
PV-ctt001					*	*	*	*			*				***	
PV-at003					*						*	*			***	*
PV-at001										***	*	*		*		
B05																
BM175											*					
BMd20													*	*		
B06																
BM137		*						*								
BM187								*			***				***	***
B07																
BM160					*					*	*				***	
BM183					***	*		***			*		***	***		
BM201								*								
BM205	*				***						*	*				
<i>Phs</i>		*			***					*		*				*
B09																
BM141										*	*				*	
BM188A									*							
B011																
BM188B										*	*					
Mean	32.40	7.61	5.24	11.16	21.60	6.85	4.77	10.69	48.40	8.24	5.97	13.35	57.92	8.93	6.91	13.12
SD	10.69	1.19	0.76	1.51	8.91	1.15	0.66	1.67	14.66	0.96	0.95	2.08	16.25	0.99	1.01	2.18

^a Trait abbreviations: *100S* weight of 100 seeds (in g), *W* width (in mm), *H* height (in mm), *L* length (in mm)^b Microsatellite loci organized according to linkage group assignment from Blair et al. (2003)^c Probability for each significant association (* $P \leq 0.05$, ** $P < 0.01$, *** $P < 0.001$) whit Bonferroni correction

Discussion

This study describes one of the first large-scale genotyping exercises for common bean using a core collection and fluorescent microsatellite marker technology. Core collections are considered valuable ways of analyzing phenotypic diversity (Brown 1989) but generally have not been genotyped before due to a lack of high throughput marker technologies. Therefore, in this project our initial step was to develop the fluorescent marker set needed for large-scale genotyping. Before starting this project, fluorescent microsatellites had not been widely used for common bean as only one publication had attempted this technology for this crop (Masi et al. 2002). We decided to use fluorescent microsatellites and automated detection on a capillary system since this technology allows many samples to be evaluated in a uniform manner and gives more accurate allele calls than gel-based readings of microsatellites.

Fluorescent microsatellite evaluation uncovered a larger number of alleles than in our previous studies of common bean genetic diversity with microsatellites (Gomez et al. 2004; Blair et al. 2007; Díaz and Blair 2006; Zhang et al. 2008), however, a direct comparison is difficult due to the different number and source of genotypes included in each study. Allele richness was also higher than in the previous studies, with a Nei's index for the entire set of genotypes of 0.643, which may reflect the large number of unique alleles. These values were even higher for the genomic (non-coding) microsatellites analyzed alone (0.734) when compared with the gene-based microsatellites analyzed alone (0.523), which agrees with our previous study comparing a large number of silver-stained markers from both of these categories (Blair et al. 2006a). It was notable that the AT motif-based marker was more polymorphic than GA- and GAT-based microsatellites and that overall fluorescent marker technology was highly accurate in calling alleles with small size differences (Diwan and Cregan 1997).

For the diversity assessment carried out in the PCoA, the results from fluorescent microsatellite evaluation agreed well with previous evaluations of common bean diversity with smaller sets of germplasm and non-fluorescent markers (Blair et al. 2006a, 2007; Díaz and Blair 2006). In turn, these results augment the analysis done by Beebe et al. (2001, 2002) with anonymous, dominant RAPD markers used to characterize the Mesoamerican section of this same core collection. We found the proportion of shared alleles to be a useful measure of genetic similarity, which was preferable to Dice similarity coefficient for a large dataset such as ours where non-amplification occurred for 4.1% of the total genotype \times marker combinations. The reproducibility of fluorescent microsatellites as precisely read, single copy, co-dominant markers is advantageous over other types of markers, but some non-amplification occurred,

perhaps, due to the small reaction volumes used in PCR or to the threshold sensitivity of the detection system.

As in our previous studies, using gel-based assays of microsatellites (Blair et al. 2007; Díaz and Blair 2006), race structure of common bean was doubly dichotomous with two large groups in each of two gene pools, namely the Durango–Jalisco complex and the Mesoamerica race in the Mesoamerican gene pool and the Nueva Granada and Peru races in the Andean gene pool. However, in addition, approximately 10% of the genotypes clustered into a very distinct group of climbing beans all from Guatemala or southern Mexico that we defined as race Guatemala. These results confirm the existence of race Guatemala as an additional race within the Mesoamerican gene pool validating the results with RAPD markers from Beebe et al. (2000) and expanding on the microsatellite analysis from Díaz and Blair (2006). Furthermore, these results are useful in better defining race Guatemala, since the initial differentiation of the race by Beebe et al. (2000) was based on a small number of genotypes (12) while here we identified 61 race Guatemala genotypes. Meanwhile, race Guatemala may have been difficult to detect in RFLP or RAPD analyses such as those by Becerra-Velazquez and Gepts (1994) or Johns et al. (1997) because it is less frequent in random germplasm collections and therefore may not have been sampled in these previous studies. Within the Andean gene pool, the Chile race could not be identified here, perhaps, due to reduced sampling of germplasm from the southern Andes of South America or due to the fact that this race is not highly distinct from race Peru as was found in our previous study of a smaller group of genotypes (Blair et al. 2007). Specific studies with Chilean genotypes are needed to define this race further.

Another observation was that secondary center germplasm was found to represent part of the diversity found in the primary centers. As a result, genetic diversity in this study was higher in the primary centers than in the secondary centers of diversity. This could be explained, perhaps, due to the limited gene flow out of the primary center for some of the races or due to more limited sampling of secondary center diversity. The races that appear to be most extensive outside the center of diversity are races Mesoamerica and Nueva Granada, which agrees with observations made by Singh et al. (1991b). Further analysis and integration of SSR datasets from this study and from previous evaluations of dry beans with primary center germplasm or race standards Díaz and Blair 2006; Blair et al. 2007) or with secondary center germplasm from Central America (Gomez et al. 2004), China (Zhang et al. 2008) or Southern Europe (Lioi et al. 2005; Maras et al. 2006; Sicard et al. 2005) will be useful for an even better understanding of common bean race structure and genetic diversity. Another observation from our study was that the gene pools

had equivalent amounts of SSR diversity, while in a previous study we predicted more diversity in the Andean genepool than in the Mesoamerican genepool (Blair et al. 2006a).

This study also helped validate subgrouping within each of the races, which can be considered as subraces. For example, the subgroups within the Mesoamerica race largely corresponded to the subraces observed by Beebe et al. (2000) who predicted the existence of groups M1 and M2; however, the subdivisions of the Durango–Jalisco group did not correspond directly to the subgroups identified by Beebe et al. (2000) but did correspond well to those of Díaz and Blair (2006). Meanwhile, we did not identify any subdivisions within race Guatemala in contrast to the results of Beebe et al. (2000) even though we included four of the same genotypes as this previous study with two from each of the subdivisions identified by them. Subgrouping within the Andean genepool agreed with those of Blair et al. (2007) but much greater admixture was found among the groups from races Nueva Granada and Peru. Growth habit was found to not be a defining characteristic of each of the races or subraces. As was found in Blair et al. (2007), genotypes with climbing growth habit were found within the Nueva Granada subgroups and genotypes with bush bean growth habit were found within the Peru race subgroups, contrasting with the original race morphological description from Singh et al. (1991b).

Meanwhile, phaseolin pattern that was useful for genepool identification was found to be only partially predictive of race and subrace/subgroup divisions and mostly in the Andean genepool, where a diverse set of phaseolin patterns was observed, although phaseolin ‘T’ and then to lesser extent ‘H’ were most common. Interestingly, the proportion of these two patterns varied in each of the races, with phaseolin ‘T’ more prevalent in race Nueva Granada (74.8%) than in race Peru (60.2%) and ‘H’ higher in race Peru (18.8%) than in race Nueva Granada (6.9%). Meanwhile in the Mesoamerican geneopool, all the races and subgroups had the same phaseolin alleles (S, Sb, Sd and B) so microsatellite allele frequency would be a better method of separating races within this genepool.

The finding of similar levels of diversity in both geneopools but varying diversity within subgroups contrast with our previous studies with non-fluorescently labeled microsatellite markers (Blair et al. 2006a, 2007; Díaz and Blair 2006) where Andean diversity was higher than Mesoamerican diversity and where diversity was equivalent in most subgroups. This may have been due to the evaluation of more diverse genotypes from race Guatemala and the Durango–Jalisco complex than in these previous studies. The similarity in genetic diversity between geneopools could also be explained by the discovery of higher diversity for genomic microsatellites in Andean subgroups and

contrastingly the higher diversity of cDNA-based microsatellites in Mesoamerican subgroups. In our previous analysis with microsatellite markers, we used fewer gene-based markers when compared with genomic markers (Blair et al. 2006a) so this may also explain the difference between studies. Higher diversity within the Andean genepool for genomic microsatellites may also be due to ascertainment bias since an Andean genotype was used to develop the genomic markers (Gaitán et al. 2002). However, overall the results indicate the value of using microsatellites from several sources to evaluate genetic diversity in the crop and by extension the utility of both types of markers for association analysis, as discussed below.

Association analysis for microsatellite and phaseolin alleles with seed size or shape was performed to compare the effectiveness of association mapping with previous reports of linkage mapping and QTL analysis for these traits in bi-parental populations. Some of the most highly significant marker \times seed weight associations ($P < 0.0001$) agreed with previously reported QTL on linkage groups b02, b03, b07 and b11 in wild \times cultivated backcross and recombinant inbred line (RIL) populations (Blair et al. 2006b; Koinange et al. 1996) as well as in inter and intra-genepool RIL populations (Park et al. 2000; Tar’an et al. 2002). The phaseolin locus, *Phs*, was found to be associated with seed weight in the Mesoamerica race agreeing with results from the first of these studies and appears to be especially important when considering wild bean introgression. Additional associations at $P < 0.01$ were found for phaseolin alleles and seed width in race Durango, for seed width or length in race Nueva Granada and for seed length in race Peru but none were found for race Guatemala. The association of BM143 and several other markers on linkage group b02 with seed weight agrees with results for a seed size QTL from Blair et al. (2006b) and may highlight this region of the genome as important in determining this characteristic. We decided to evaluate seed size characteristics because these are among the most important traits of common bean and a major determinant of commercial acceptability of traditional or improved varieties (Park et al. 2000; Voysest et al. 1994) as well as being highly heritable (Motto et al. 1978). Furthermore, bean seed weight has also been correlated with yield potential and with various yield components (Terán and Singh 2002) and varieties of dry beans that do not meet minimum seed weight criteria can be rejected by the marketplace while in snap beans, varieties with elongated seed shape or smaller seed size are preferred as they are correlated with pod quality (Blair et al. 2006b).

This study found that within each molecularly defined race of dry bean, there was variability for seed size and shape and that the range between the smallest and largest cultivated varieties for seed weight within the Mesoamerica

race represented part of the sixfold difference found for all cultivated beans. When comparing wild and cultivated beans this difference is amplified to over 20–25-fold difference in seed weight (Acosta-Gallegos et al. 2007). Furthermore, seed size can influence and be influenced by the concentration of macronutrients such as protein, phytate and starch levels; and the proportion of seed coat to endosperm is directly proportional in most cases to seed size and shape with resulting influences on tannin levels and mineral concentration, since these are differentially found in seed coat tissues (Ariza-Nieto et al. 2007). Therefore, the markers we identified as being associated with seed weight, length, width and height can be important for the understanding or marker-assisted selection of this trait, especially in validating results from QTL studies, which only use a few genotypes as parents. In addition, the comprehensive germplasm set that was evaluated here and that included all races of the crop could be useful for discovering signatures of domestication bottlenecks for seed size traits where seed weight was a major characteristic emphasized by human selection (Papa and Gepts 2003).

In conclusion, the results of this large scale genotyping of common bean using fluorescent microsatellite markers will be useful as a baseline for further genotypic characterization of this and other collections, especially since this is the first integrated analysis of genotypes from primary and secondary centers of diversity and from both major gene pools of common bean using SSR markers. Overall, race structure was similar to previous studies but the current analysis was more definitive due to the larger number of genotypes analyzed and highlighted the importance of races Nueva Granada and Peru in the Andean gene pool and Mesoamerica, Durango–Jalisco and Guatemala in the Mesoamerican gene pool. The core collection analyzed here is unique in representing all major agro-ecologies where common bean is grown based on a geographic information system analysis that took into account climatic parameters and origin and the microsatellites we used cover most of the genome. As such, these results can be used to complement diversity assessments of national collections by providing the global population structure and whole-genome genotype in which to integrate specific collections and markers. In addition, a selection of the genotypes from the core collection with adaptation to tropical environments and lack of photoperiod sensitivity has been constituted into a reference collection that will be used to assay diversity at individual loci and for further phenotyping and association mapping.

Acknowledgments The authors wish to thank A. V. Gonzales, M. Muñoz and F. Pedraza for DNA extractions and the personnel of the CIAT Genetic Resource Units and CIAT bean program for seed characterization and multiplication. We also acknowledge S. Kresovich, S. Mitchell and T. Fulton at the Institute for Genomic Diversity as well as the Cornell Biotechnology Center for support in genotyping. We are

also grateful to D. Debouck and S. Beebe from CIAT for germplasm advice. This research was supported by the Generation Challenge Program and CIAT.

References

- Acosta-Gallegos JA, Kelly JD, Gepts P (2007) Prebreeding in common bean and use of genetic diversity from wild germplasm. *Crop Sci* 47:44–59
- Afanador L, Hadley S, Kelly JD (1993) Adoption of a mini-prep DNA extraction method for RAPD marker analysis in common bean (*Phaseolus vulgaris* L.). *Bean Improv Coop* 36:10–11
- Akkaya MS, Bhagwat AA, Cregan PB (1992) Length polymorphism of simple sequence repeat DNA in soybean. *Genetics* 132:1131–1139
- Ariza-Nieto M, Blair MW, Welch RM, Glahn RP (2007) Screening of bioavailability patterns in eight bean (*Phaseolus vulgaris* L.) genotypes using the Caco-2 cell in vitro model. *J Agr Food Sci* 55:7950–7956
- Becerra-Velazquez L, Gepts P (1994) RFLP diversity of common bean (*Phaseolus vulgaris* L.) in its centres of origin. *Genome* 37:256–263
- Beebe S, Lynch J, Galwey N, Tohme J, Ochoa I (1997) A geographical approach to identify phosphorus-efficient genotypes among landraces and wild ancestors of common bean. *Euphytica* 95:325–336
- Beebe S, Skroch P, Tohme J, Duque MC, Pedraza F, Nienhuis J (2000) Structure of genetic diversity among common bean landraces of middle American origin based on correspondence analysis of RAPD. *Crop Sci* 40:264–273
- Beebe S, Renjifo J, Gaitán-Solís E, Duque MC, Tohme J (2001) Diversity and origin of Andean landraces of common bean. *Crop Sci* 41:854–862
- Blair MW, Hedetale V, McCouch SR (2002) Fluorescent-labeled microsatellite panels useful for detecting allelic diversity in cultivated rice (*Oryza sativa* L.). *Theor Appl Genet* 105:449–457
- Blair MW, Pedraza F, Buendia H, Gaitán-Solís E, Beebe S, Gepts P, Tohme J (2003) Development of a genome-wide anchored microsatellite map for common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 107:1362–1374
- Blair MW, Giraldo MC, Buendia HF, Tovar E, Duque MC, Beebe S (2006a) Microsatellite marker diversity in common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 113:100–109
- Blair MW, Iriarte G, Beebe S (2006b) QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean x wild common bean (*Phaseolus vulgaris* L.) cross. *Theor Appl Genet* 112:1149–1163
- Blair MW, Diaz JM, Hidalgo R, Diaz LM, Duque MC (2007) Microsatellite characterization of Andean races of common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 116:29–43
- Broughton WJ, Hernandez G, Blair MW, Beebe SE, Gepts P, Vanderleyden J (2003) Beans (*Phaseolus* spp.)—model food legumes. *Plant Soil* 252:55–128
- Brown AHD (1989) Core collections: a practical approach to genetic resources management. *Genome* 31:818–824
- Brown JWS, Ma Y, Bliss FA, Hall TC (1981) Genetic variation in the subunits of Globulin-1 storage protein of French bean. *Theor Appl Genet* 59:83–88
- Buckler ES, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. *Curr Opin Plant Biol* 5:107–111
- Buckler E, Bradbury P, Kroon D, Ramdoss Y, Casstevens T, Zhang Z (2007) TASSEL v. 2 (Trait analysis by association, evolution and linkage). <http://www.maizegenetics.net/tassel>
- Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhou S (2000) Diversity of microsatellites

- derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *Theor Appl Genet* 100:713–722
- Coburn JR, Temnykh SV, Paul EM, McCouch SR (2002) Design and application of microsatellite marker panels for semiautomated genotyping of rice (*Oryza sativa* L.). *Crop Sci* 42:2092–2099
- Crossa J, Burgueño J, Dreisigacker S, Vargas M, Herrera-Foessel SA, Morten L, Singh RP, Trethowan R, Warburton M, Franco J, Reynolds M, Crouch JH, Ortiz R (2007) Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics* 177:1889–1913
- Díaz LM, Blair MW (2006) Race structure within the Mesoamerican gene pool of common bean (*Phaseolus vulgaris* L.) as determined by microsatellite markers. *Theor Appl Genet* 114:143–154
- Diwan N, Cregan PB (1997) Automated sizing of fluorescent-labeled simple sequence repeat (SSR) markers to assay genetic variation in soybean. *Theor Appl Genet* 95:723–733
- Durán LA, Blair MW, Giraldo MC, Machiavelli R, Prophete E, Nin JC, Beaver JS (2005) Morphological and molecular characterization of common bean (*Phaseolus vulgaris* L.) landraces from the Caribbean. *Crop Sci* 45:1320–1328
- Excoffier L (2007) Arlequin ver 3.11 Computational and molecular population genetics lab CPMG. Zoological Institute, University of Berne. <http://cmpg.unibe.ch/software/arlequin3/>
- Felsenstein J (1993) PHYLIP—phylogeny inference package, Version 3.67. Department of Genome Sciences and Department of Biology, University of Washington, Seattle, WA, USA. <http://cmgm.stanford.edu/phylip/>
- Flint-Garcia SA, Thuitte A-C, Yu J, Pressoir G, Romero SM, Mitchel SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* 44:1054–1064
- Gaitán E, Duque MC, Edwards K, Tohme J (2002) Microsatellite repeats in common bean (*Phaseolus vulgaris* L.): isolation, characterization, and cross-species amplification in *Phaseolus* spp. *Crop Sci* 42:2128–2136
- Gao H, Williamson S, Bustamante CD (2007) An MCMC approach for joint inference of population structure and inbreeding rates from multi-locus genotype data. *Genetics* (online). Cornell Computational Biology Service Unit (CBSU). <http://cbsuapps.tc.cornell.edu/InStruct.aspx>
- Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. *Genetics* 169:1631–1638
- Gepts P, Osborn T, Rashka K, Bliss F (1986) Phaseolin—protein variability in wild forms and landraces of the common bean (*Phaseolus vulgaris* L.): evidence for multiple centers of domestication. *Econ Bot* 40:451–468
- Gomez O, Blair MW, Frankow-Lindberg B, Gullberg U (2004) Molecular and phenotypic diversity of common bean landraces from Nicaragua. *Crop Sci* 44:1412–1418
- Hayden MJ, Nguyen TM, Waterman A, McMichael GL, Chalmers KJ (2008) Application of multiplex-ready PCR for fluorescence-based SSR genotyping in barley and wheat. *Mol Breeding* 21:271–281
- Idury RM, Cardon LR (1997) A simple method for automated allele binning in microsatellite markers. *Genome Res* 11:1104–1109
- Islam FM, Basford KE, Redden RJ, Gonzalez AV, Kroonenberg PM, Beebe SE (2002) Genetic variability in cultivated common bean beyond the two major gene pools. *Genet Resour Crop Evol* 49:271–283
- Islam FM, Beebe S, Muñoz M, Tohme J, Redden RJ, Basford KE (2004) Using molecular markers to assess the effect of introgression on quantitative attributes of common bean in the Andean gene pool. *Theor Appl Genet* 108:243–252
- Johns MA, Skroch PW, Nienhuis J, Hinrichsen P, Bascur G, Muñoz-Schick C (1997) Gene pool classification of common bean landraces from Chile based on RAPD and morphological data. *Crop Sci* 37:605–613
- Koinange EMK, Singh SP, Gepts P (1996) Genetic control of the domestication syndrome in common-bean. *Crop Sci* 36:1037–1145
- Lioi L, Piergiovanni AR, Pignone D, Puglisi S, Santantonio M, Sonnante G (2005) Genetic diversity of some surviving on-farm Italian common bean (*Phaseolus vulgaris* L.) landraces. *Plant Breed* 124:576–581
- Liu K, Muse S (2005) PowerMarker: integrated analysis environment for genetic marker data. *Bioinformatics* 21:2128–2129. <http://www.powerMarker.net>
- Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165:2117–2128
- Malysheva-Otto LV, Ganai MW, Röder Marion S (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). *BMC Genetics* 7:6
- Maras M, Susnik S, Sustar-Vozlic J, Meglic V (2006) Temporal changes in genetic diversity of common bean (*Phaseolus vulgaris* L.) accessions cultivated between 1800 and 2000. *Russ J Genet* 42:775–782
- Masi P, Spagnoletti ZP, Donini P (2003) Development and analysis of multiplex microsatellite markers sets in common bean (*Phaseolus vulgaris* L.). *Mol Breed* 11:303–313
- McCouch SR, Chen X, Panaud O, Temnykh S, Xu Y, Gu Cho Y, Huang N, Ishii T, Blair M (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. *Plant Mol Biol* 35:89–99
- Miklas PN, Delorme R, Hannan R, Dickson MH (1999) Using a subsample of the core collection to identify new sources of resistance to white mold in common bean. *Crop Sci* 39:569–573
- Motto M, Soressi GP, Salamini F (1978) Seed size inheritance in a cross between wild and cultivated common bean (*Phaseolus vulgaris* L.). *Genetica* (The Hague) 49:31–36
- Oumar I, Mariac C, Pham JL, Vigouroux Y (2008) Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br.) as revealed by microsatellite loci. *Theor Appl Genet* 117:489–497
- Papa R, Gepts P (2003) Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. *Theor Appl Genet* 106:239–250
- Park SO, Coyne DP, Jung G, Skroch PW, Arnaud-Santana E, Steadman JR, Ariyaratne HM, Nienhuis J (2000) Mapping of QTL for seed size and shape traits in common bean. *J Am Soc Hort Sci* 125:466–475
- Pejic I, Ajmore-Marsan P, Morgante M, Kozumplick V, Castiglioni P, Taramino G, Motto M (1998) Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. *Theor Appl Genet* 97:1248–1255
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Reed PW, Davies JL, Copeman JB, Bennett ST, Palmer SM, Pritchard LE, Gough SCL, Kawaguchi Y, Cordell HJ, Balfour KM, Jenkins SC, Powell EE, Vignal A, Todd JA (1994) Chromosome specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature Genet* 7:390–395
- Rohlf F (2002) NTSYS pc. Numerical taxonomy system. Exeter publishing, Setauket
- Rosenberg NA (2002) Distruct: a program for the graphical display of structure results. <http://www.cmb.usc.edu/~noahr/distruct>
- Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing

- facilitates whole-genome association mapping in elite inbred crop varieties. *Proc Natl Acad Sci USA* 103:18656–18661
- SAS-Institute (1996) SAS/STAT Users guide 6.11. SAS Inst, Cary
- Schwengel DA, Jedlicka AE, Nanthakumar EJ, Weber JL, Levitt RC (1994) Comparison of Fluorescence-based semi-automated genotyping of multiple microsatellite loci with autoradiographic techniques. *Genomics* 22:46–54
- Sicard D, Nanni L, Porfiri O, Bulfon D, Papa R (2005) Genetic diversity of *Phaseolus vulgaris* L. and *P. coccineus* L. landraces in central Italy. *Plant Breed* 124:464–472
- Singh S, Gepts P, Debouck D (1991a) Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Econ Bot* 45:379–396
- Singh S, Gutierrez A, Molina A, Urrea C, Gepts P (1991b) Genetic diversity in cultivated common bean. II. Marker-based analysis on morphological and agronomic traits. *Crop Sci* 31:23–29
- Singh S, Nodari R, Gepts P (1991c) Genetic diversity in cultivated common bean. I. Allozymes. *Crop Sci* 31:19–23
- Tar'an B, Michaels TE, Pauls KP (2002) Genetic mapping of agronomic traits in common bean (*Phaseolus vulgaris* L.). *Crop Sci* 42:544–556
- Terán H, Singh SP (2002) Comparison of sources and lines selected for drought resistance in common bean. *Crop Sci* 42:64–70
- Thomson MJ, Septiningsih EM, Suwardjo F, Santoso TJ, Silitonga TS, McCouch SR (2007) Genetic diversity analysis of traditional and improved Indonesian rice (*Oryza sativa* L.) germplasm using microsatellite markers. *Theor Appl Genet* 114:559–568
- Tohme J, Jones P, Beebe SE, Iwanaga M (1995) The combined use of agroecological and characterisation data to establish the CIAT *Phaseolus vulgaris* core collection. In: Hodgkin T, Brown AHD, van Hintum ThJL, Morales EAV (eds) Core collections of plant genetic resources. International Plant Genetic Resources Institute (IPGRI). Wiley-Sayce Publication, London, pp 95–107
- Tommasini L, Batley J, Arnold GM, Cooke RJ, Donini P, Lee D, Law JR, Lowe C, Moule C, Trick M, Edwards KJ (2003) The development of multiplex simple sequence repeat (SSR) markers to complement distinctness, uniformity and stability testing of rape (*Brassica napus* L.) varieties. *Theor Appl Genet* 106:1091–1101
- CIAT (Centro Internacional de Agricultura Tropical) (1987) Standard system for the evaluation of bean germplasm. In: Schoonhoven AV, Pastor-Corrales MA (eds) Cali, Colombia
- Voysest O, Valencia M, Amezcuita M (1994) Genetic diversity among Latin American Andean and Mesoamerican common bean cultivars. *Crop Sci* 34:1100–1110
- Warburton ML, Reif JC, Frisch M, Bohn M, Bedoya C, Xia XC, Crossa J, Franco J, Hoisington D, Pixley K, Taba S, Melchinger AE (2008) Genetic diversity in CIMMYT nontemperate maize germplasm: landraces, open pollinated varieties and inbred lines. *Crop Sci* 48:617–624
- Yu K, Park J, Poysa V, Gepts P (2000) Integration of simple sequence repeat (SSR) markers into a molecular linkage map of common bean (*Phaseolus vulgaris* L.). *J Hered* 91:429–434
- Zhang X, Blair MW, Wang S (2008) Genetic diversity of Chinese common bean (*Phaseolus vulgaris* L.) landraces assessed with simple sequence repeat markers. *Theor Appl Genet* 117:629–640