Theor Appl Genet (2013) 126:535–548
DOI 10.1007/s00122-012-1999-z

ORIGINAL PAPER

A high-throughput SNP marker system for parental polymorphism screening, and diversity analysis in common bean (*Phaseolus vulgaris* L.)

Matthew W. Blair · Andrés J. Cortés · R. Varma Penmetsa · Andrew Farmer · Noelia Carrasquilla-Garcia · Doug R. Cook

Received: 14 June 2012 / Accepted: 6 October 2012 / Published online: 3 November 2012

Abstract Single nucleotide polymorphism (SNP) detection has become a marker system of choice, because of the high abundance of source polymorphisms and the ease with which allele calls are automated. Various technologies exist for the evaluation of SNP loci and previously we validated two medium throughput technologies. In this study, our goal was to utilize a 768 feature, Illumina GoldenGate assay for common bean (*Phaseolus vulgaris* L.) developed from conserved legume gene sequences and to use the new technology for (1) the evaluation of parental polymorphisms in a mini-core set of common bean accessions and (2) the analysis of genetic diversity in the crop. A total of 736 SNPs were scored on 236 diverse common bean genotypes with the GoldenGate array. Missing data and heterozygosity levels were low and 94 % of the SNPs were scorable. With the evaluation of the parental polymorphism genotypes, we estimated the utility of the SNP markers in mapping for inter-genepool and intra-genepool populations, the latter being of lower polymorphism than the former. When we performed the diversity analysis with the diverse genotypes, we found Illumina GoldenGate SNPs to provide equivalent evaluations as previous gene-based SNP markers, but less fine-distinctions than with previous microsatellite marker analysis. We did find, however, that the gene-based SNPs in the GoldenGate array had some utility in race structure analysis despite the low polymorphism. Furthermore the SNPs detected high heterozygosity in wild accessions which was probably a reflection of ascertainment bias. The Illumina SNPs were shown to be effective in distinguishing between the genepools, and therefore were most useful in saturation of inter-genepool genetic maps. The implications of these results for breeding in common bean are discussed as well as the advantages and disadvantages of the GoldenGate system for SNP detection.

Communicated by C. Gebhardt.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1999-z) contains supplementary material, which is available to authorized users.

M. W. Blair (✉)
Department of Plant Breeding and Genetics,
Cornell University, Ithaca, NY 14853, USA
e-mail: mwbeans@gmail.com; mwb1@cornell.edu

M. W. Blair
Universidad Nacional de Colombia, Sedes Palmira
and Bogotá, Colombia

A. J. Cortés
University of Uppsala, Uppsala, Sweden

R. V. Penmetsa · N. Carrasquilla-Garcia · D. R. Cook
University of California, Davis, CA 95616, USA

A. Farmer
National Center for Genomic Research, Albuquerque, NM, USA

Introduction

Single nucleotide polymorphism (SNP) markers are considered ideal for genetic mapping and diversity assessment in crop plants due to their high abundance and relatively even distribution across the genome (Chagné et al. 2007). In addition, various technologies exist for the evaluation of SNP loci and many of these are easy to automate for allele calling and data collection. Among these techniques, the Illumina GoldenGate assay has proven to be high throughput and useful for genetic fingerprinting (Hyten et al. 2008; Yan et al. 2010; Zhao et al. 2010). The GoldenGate detection system is based on allele-specific

primers and fluorescent dye signal detection that are standardized for a specific set of SNPs. The multiplexing capacity of the assay comes from the use of locus-specific oligonucleotides combined with IllumiCode beads used for recognition of each SNP (Oliphant et al. 2002). Sequence variants are discovered through sequencing project and then SNP markers are developed for ideal single nucleotide polymorphisms that are located in regions where the sequence allows for a functional assay. The GoldenGate assay has been developed for several legume crops including large sets for soybean (Hyten et al. 2008, 2010) and cowpea (Muchero et al. 2009) and smaller assays for other species.

Common bean (*Phaseolus vulgaris* L.) is an important crop that is native to the New World, and has become the most widely consumed legume for direct human consumption (Broughton et al. 2003). The species is a true diploid ($n = 11$) with a rapid growth cycle (60–120 days generally) and a small genome (650 Mb) that has been used in many studies of nitrogen fixation, low soil fertility adaptation and nutritional quality. The genome structure of common bean serves as a model for the more complicated structure of the soybean genome (Galeano et al. 2009b; McConnell et al. 2010) and should be closely related to cowpea and other tropical legume species genomes. Polymorphism in common bean is high due to the two separate genepool (Andean and Mesoamerican) subdivisions it has as a crop based on its multiple centers of origin (Gepts et al. 2008).

Current SNP sets in common bean include 94 from variable sources tested with the Kaspar assay by Cortés et al. (2011), approximately 300 based on genes sequenced from BAT93 and JaloEEP558 by McConnell et al. (2010) evaluated by cleaved amplified product assays and 827 based on genomic fragments of the same two genotypes obtained by reduced representation in Hyten et al. (2010). Only the last of these has been converted to a GoldenGate assay but the full set of sequences and array are not yet published. Meanwhile the SNPs from Cortés et al. (2011) are available in a flexible single-locus assay, but have not been converted to a multi-locus assay. Aside from this, we also have developed the SSCP detection system and a Ecotilling based assay for SNP detection with the enzyme CELI (Galeano et al. 2009a, b), but neither of the SNP sets tested in those studies are available as GoldenGate assays.

The objective of this research was to develop an Illumina GoldenGate assay based on tentative orthologous gene (TOG) sequences from the legumes and to use the technology for fingerprinting in common bean. The TOG markers were developed as part of a cross-legume marker project using amplicons of BAT93 and Jalo EEP558, and the resulting GoldenGate SNP set that was developed consisted of 768 individual gene-based markers. For the genotyping and diversity analysis, we evaluated both a mini-core panel of common bean genotypes used by Cortés et al. (2011) to determine polymorphism levels in the crop and wild relatives as well as a validation set from Blair et al. (2009) of diverse common beans with known cultivar race assignments. We then compared and contrasted the various SNP detection systems available for common bean.

Materials and methods

Plant material

We used two sets of genotypes for this study. The first set consisted in a mini-core panel of 50 genotypes representing parents of genetic mapping populations as described in Blair et al. (2006a), and genotypes tested for Kaspar SNP discovery in Cortés et al. (2011) as listed in Table 1. The trait characteristics for these genotypes included disease resistance to common bacterial blight caused by *Xanthomonas anopodis* pv. *phaseoli* (VAX series), to angular leaf spot caused by *Phaeoisariopsis griseola* (MAR series), anthracnose caused by *Colletotrichum lindemuthianum* and bean golden yellow mosaic virus (DOR series), insect resistance to *Apion godmani* and *Thrips palmi*, abiotic stress tolerance to aluminum toxic soils (MAM series), high heat (G122 and IJR), drought conditions (SEA series) and low phosphorous soils, extremes of seed micronutrient content, variation in growth habit or good architecture (A series), as well as yield and its components (released varieties such as AFR298, CALIMA and G series landraces).

Included in the mini-core germplasm set for diversity evaluation were G19833 and BAT93 that have been chosen for full genome sequencing as well as a range of commercial seed classes that vary greatly in seed color and size. Finally, three wild bean accessions from Argentina (G19892), Colombia (G24404) and Mexico (G24390) were included in the study. The first two wild accessions represented Andean accessions while the last wild accession represented the Mesoamerican genepool, although the Colombian wild bean was fairly diverse (Blair et al. 2006b). In total, 20 genotypes were from the Andean genepool and 30 were from the Mesoamerican genepool.

In addition to the mini-core collection, a second set of genotypes were selected from those evaluated by Blair et al. (2009). These were used as a validation set and included a total of 186 genotypes, of which 80 were from the Andean genepool and 106 were from the Mesoamerican genepool. The identities of these genotypes are listed in Supplemental Table 1, while the genepool identities of the mini-core, parental combinations and distribution between Andean and Mesoamerican genepools have been described

Table 1 Common bean genotypes used for assessment of SNP diversity and their accession number, phaseolin status, race and gene pool identity, origin and growth habit

Genotype	Ph	Genepool	Race	Status	Origin	GH
Cultivated Andean						
AFR298	NA	Andean	NA	Cultiv	CIAT	I
BRB191	T	Andean	NG	Cultiv	CIAT	II
CAL96	NA	Andean	NA	Cultiv	CIAT	I
CAL143	NA	Andean	NA	Cultiv	CIAT	I
G122	NA	Andean	NA	Cultiv	India	I
G4494 (Calima)	T	Andean	P	Cultiv	Colombia	I
G4523	NA	Andean	NA	Cultiv	Colombia	I
G5273	T	Andean	NG	Cultiv	Mexico	II
G19833	H	Andean	P	Cultiv	Peru	III
G19839	T	Andean	P	Cultiv	Peru	III
G21078	T	Andean	P	Cultiv	Argentina	IV
G21242	C	Andean	NA	Cultiv	Colombia	IV
G21657	C	Andean	P	Cultiv	Bulgaria	III
IJR	T	Andean	NG	Cultiv	Jamaica	I
JaloEEP558	T	Andean	NG	Cultiv	Brazil	III
Montcalm	NA	Andean	NA	Cultiv	USA	I
Radical Cerinza	T	Andean	P	Cultiv	Colombia	I
SEQ1027	T	Andean	NG	Cultiv	CIAT	III
Cultivated Meso						
A55	NA	Mesoamerican	NA	Cultiv	CIAT	II
BAT93	S	Mesoamerican	M	Cultiv	CIAT	II
BAT477	S	Mesoamerican	M	Cultiv	CIAT	II
BAT881	S	Mesoamerican	M	Cultiv	CIAT	II
DOR364	S	Mesoamerican	M	Cultiv	El Salvador	II
DOR390	S	Mesoamerican	M	Cultiv	CIAT	II
DOR476	S	Mesoamerican	M	Cultiv	CIAT	II
G685	Sb	Mesoamerican	G	Cultiv	Guatemala	IV
G855	Sb	Mesoamerican	J	Cultiv	Mexico	IV
G2333	S	Mesoamerican	G	Cultiv	Mexico	IV
G3513	S	Mesoamerican	M	Cultiv	Mexico	II
G4825	B	Mesoamerican	M	Cultiv	Brazil	III
G5773	S	Mesoamerican	M	Cultiv	Colombia	II
G11350	S	Mesoamerican	M	Cultiv	Mexico	III
G11360	S	Mesoamerican	J	Cultiv	Mexico	IV
G14519	S	Mesoamerican	M	Cultiv	USA	IV
G21212	B	Mesoamerican	M	Cultiv	Colombia	II
ICA Pijao	B	Mesoamerican	M	Cultiv	Colombia	II
JAMAPA	S	Mesoamerican	M	Cultiv	Mexico	II
MAM38	S	Mesoamerican	D	Cultiv	CIAT	III
MAR1	S	Mesoamerican	M	Cultiv	CIAT	II
SEA5	S	Mesoamerican	D	Cultiv	CIAT	II
SEA15	S	Mesoamerican	D	Cultiv	CIAT	II
SEA21	S	Mesoamerican	M	Cultiv	CIAT	II
SEL1309	S	Mesoamerican	M	Cultiv	CIAT	II
TioCanela	S	Mesoamerican	M	Cultiv	EAP	II
VAX1	NA	Mesoamerican	NA	Cultiv	CIAT	II
VAX3	NA	Mesoamerican	NA	Cultiv	CIAT	II
VAX6	S	Mesoamerican	M	Cultiv	CIAT	II
Wild accessions						
G19892	T	Andean	NA	Wild	Argentina	IV
G24390	M	Mesoamerican	NA	Wild	Mexico	IV
G24404	H	Mesoamerican	NA	Wild	Colombia	IV

Ph Phaseolin type, Races: *D–J* Durango–Jalisco, *G* Guatemala, *NG* Nueva Granada, *P* Peru, *M* Mesoamerica, *GH* Growth habitat as described in "Materials and methods", *NA* not applicable

in Blair et al. (2006a) and Cortés et al. (2011). Each genotype was grown from seed in a greenhouse to obtain young trifoliate leaf tissue for DNA extraction.

Illumina SNP assays

The development of the 768-feature GoldenGate assay was performed for common bean as part of a project for comparative crop legume genomics based on tentative orthologous groups (TOGs) (Penmetsa et al. in preparation). This assay monitors bi-allelic single nucleotide polymorphisms in low copy conserved orthologous loci based on these TOG genes identified through the databases for legume indices at the Dana Farber Cancer Institute. SNPs for the common bean makers were discovered by Sanger re-sequencing and alignment comparisons of a total of 1,440 TOG amplicons form one Andean genotype (JaloEEP558) and one Mesoamerican genotype (BAT93). Polymorphisms were identified by sequence alignment of the two genotype sequences. The target sequences were a set of primarily single copy orthologous genes, whose orthology was inferred initially from legume EST data (i.e., the transcriptomes of *Medicago truncatula*, *Lotus japonicus* and *Glycine max*), and subsequently based on conserved genome location in a multi-species comparative genetic analysis (Penmetsa et al. in preparation).

SNPs that were used in the common bean assay are listed in Supplemental Table 2, and were selected based on the default design criteria found in the software program Assay Design Tool from Illumina and were converted to the 768 Illumina GoldenGate genotyping assay useful for common bean. For genotyping, total genomic DNA was extracted for all the genotypes discussed above with a CTAB method from Afanador et al. (1993). DNA quantification was with a Hoefer DyNA Quant 2000 fluorometer and 200 ng/µl concentration of DNA was provided to the UC-Davis Genomic Center where the Illumina assays were carried out. Protocols for the GoldenGate SNP analysis

assay are available at the DNA Technologies Core facility (http://dnatech.genomecenter.ucdavis.edu/).

Data analysis

Allele calls were curated using the Illumina Beadstudio software package (Illumina, San Diego, CA, USA). To minimize the confounding effects of technical error, all SNP calls with monomorphism between the parents, unexpected parental alleles or high levels of missing data (20 % or more) were excluded from further analysis. Data from a total of 32 features were eliminated due to monomorphism in the parents, high levels of missing data, or excessive number of heterozygotes that would not be expected in pure lines of an inbreeding crop such as common bean. This resulted in a final data set of 736 SNPs for the genotypes evaluated. We included features from the array that had lower than 10 % missing data in some genotypes as they were still informative for other genotypes. Heterozygous calls were considered to be neither of the alleles for the sake of data analysis. We realize that in some bean genotypes heterozygosity may still exist especially in the breeding lines, but could not use this data effectively in the diversity analysis.

Allele assignments and frequencies for the common bean accessions were used to calculate the polymorphic information content (PIC) for each SNP marker using PowerMarker v. 3.25 (Liu and Muse 2005). In addition, Nei's Genetic Diversity (Nei 1978), observed heterozygosity (H_o) and maximum allele frequency were calculated for each SNP marker with the same program. Minimum allele frequencies for each SNP were calculated from maximum allele frequencies for the marker given that the SNPs were bi-allelic. Finally, as part of the diversity analysis, STRUCTURE v. 2 software (Pritchard et al. 2000) was used to determine genotype assignments to subpopulations. This involved varying the K (population number) values from 1 to 10 and running an "admixture model", with 50,000 burn-ins and 100,000 iterations, optimal K value was taken to be the one with the highest Ln P(D) values as described by Evano et al. (2005). In addition, neighbor joining trees were constructed for the mini-core and validation germplasm sets with the software program Darwin v. 5.0 software (Perrier et al. 2003).

Results

SNP marker discovery and detection

Out of the full GoldenGate set of 768 SNPs, a total of 736 SNPs gave high quality reads when the genomic DNA from the mini-core panel was tested as shown by the bitmap in Fig. 1. This amounted to 95.8 % success rate on a per marker basis. Several SNPs (3.33 % of total) still detected heterozygotes, which were not to be expected for the inbred genotypes used in the study. A total of 1,129 data points were heterozygous in the selected SNPs out of the full set of 33,856 data points. Missing data (or null alleles) were prevalent in some markers more than others and in wild genotypes more than cultivate genotypes, but overall consisted in only 908 marker × genotype combinations (2.68 % of total). Missing data or heterozygotes were far less common when evaluating the parental genotypes used in each assay plate (BAT93 and Jalo EEP558) as these were control genotypes used in each assay plate.

Characterization of SNP markers

A large number of SNP markers were found to be informative and observed heterozygosity was generally low. Figure 2 shows the distribution of observed heterozygosity in the 768 SNP markers, showing that the vast majority ranged from 0 to 0.1 in frequency with an average frequency of only 0.04 %. In addition, only 25 SNPs had higher frequencies than 0.2 while among these 12 SNPs had higher frequencies than 0.35. The low heterozygosity in the diversity panel would be expected for an inbreeding species such as common bean. High heterozygosity for a SNP of over 0.35 often represents poor allele calling in the GoldenGate assay, so these 12 SNP markers were eliminated for further analysis. Monomorphic markers detected with the genetic diversity and PIC analysis were found for 11 out of the 768 SNPs, amounting to only 1.4 % and these were also eliminated along with 9 markers having over 50 % missing data.

Figure 3 shows the distribution of genetic diversity (Nei's) and polymorphism information content (PIC) values for the SNP markers. The maximum genetic diversity value was 0.5 found for 9 out of the 768 SNP markers, while the maximum PIC value was 0.375 found for the same number of SNP markers. However, a set of 39 and 44 SNPs had genetic diversity or PIC values lower than 0.2, respectively. Finally, the average genetic diversity across all 768 SNPs was 0.423, while the average PIC value for the same set was 0.328. Although we worked mainly with genetic diversity and PIC values, because these are comparable between marker types we also report minor allele frequency (MAF) values. We found that in general the major allele represented the Mesoamerican genepool as associated with the control genotype BAT93, while the minor allele represented the Andean genepool as associated with the control genotype Jalo EEP558. Figure 4 shows the allele frequencies chart.

Correspondence between MAF and PIC values was high with a correlation coefficient of $r = 0.924$ ($P = 0.000$).

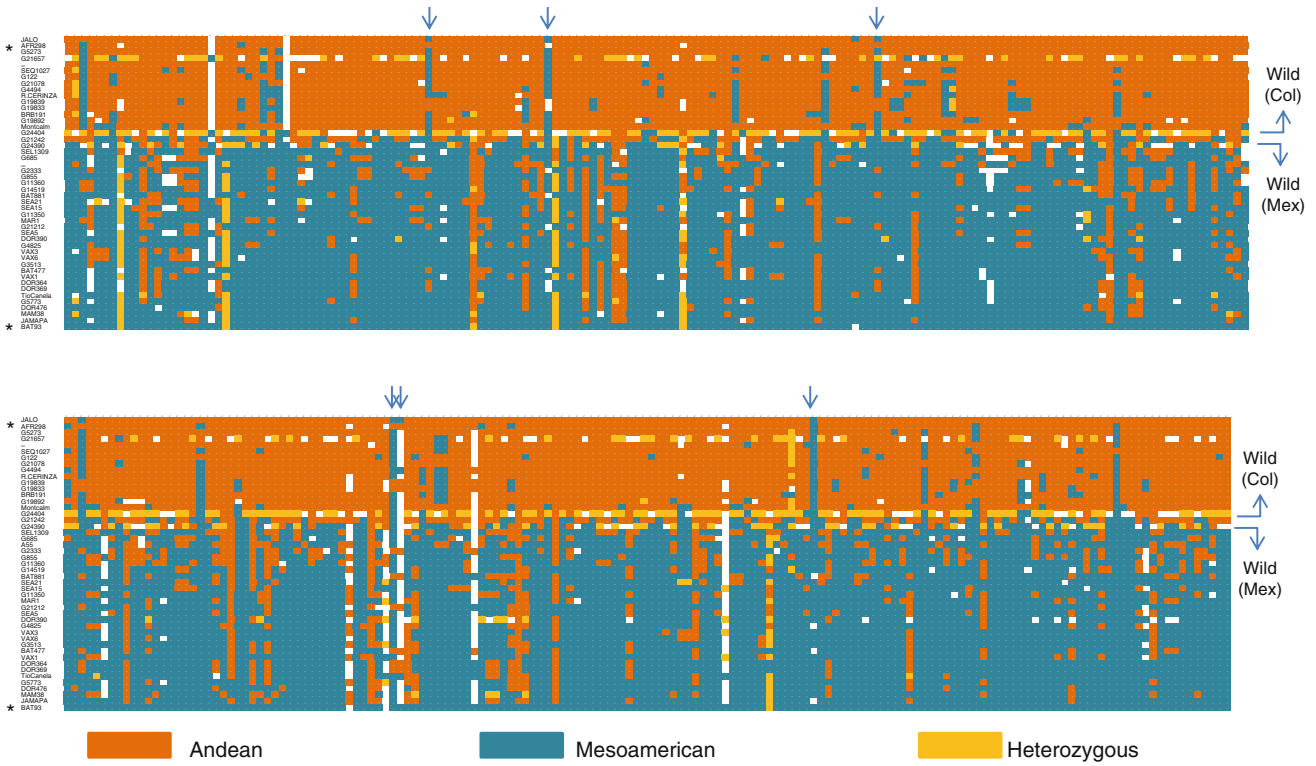

Fig. 1 Bitmap of polymorphism for common bean accessions (*rows*) and single nucleotide polymorphism (SNP) markers (*columns*) used in this study. Names to the *left* of the chart are the genotypes tested with *asterisks* indicating the Andean and Mesoamerican source genotypes used to create the Illumina SNP set, at the *top* and *bottom* of the chart, respectively, *arrow* indicates the wild genotypes from Colombia (Col) and Mexico (Mex), respectively

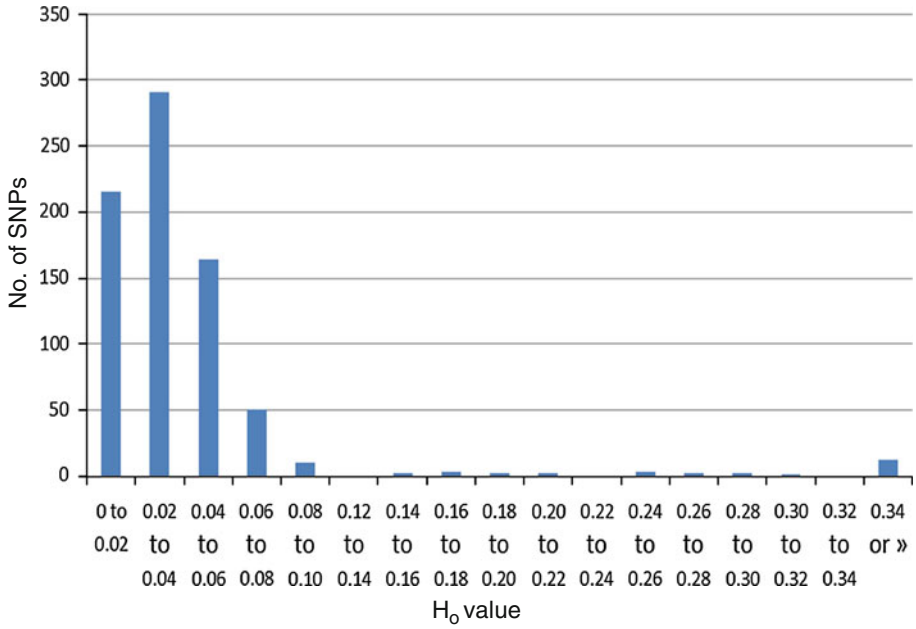

Fig. 2 Distribution of observed heterozygosity (H_o) for 768 single nucleotide polymorphism (SNP) markers used in the study

Likewise, the correlations were highly significant ($P = 0.000$) between genetic diversity and MAF ($r = 0.924$) or PIC ($r = 0.995$) values. In general, there were no inconsistencies between the rankings according to genetic diversity, MAF or PIC values. For example, the SNPs with the MAF values of zero (TOG894080_1, TOG894192_1, TOG894755_2, TOG897670_2, TOG898533_1, TOG899640_1, TOG901050_1 TOG902834_1, TOG906490_2,

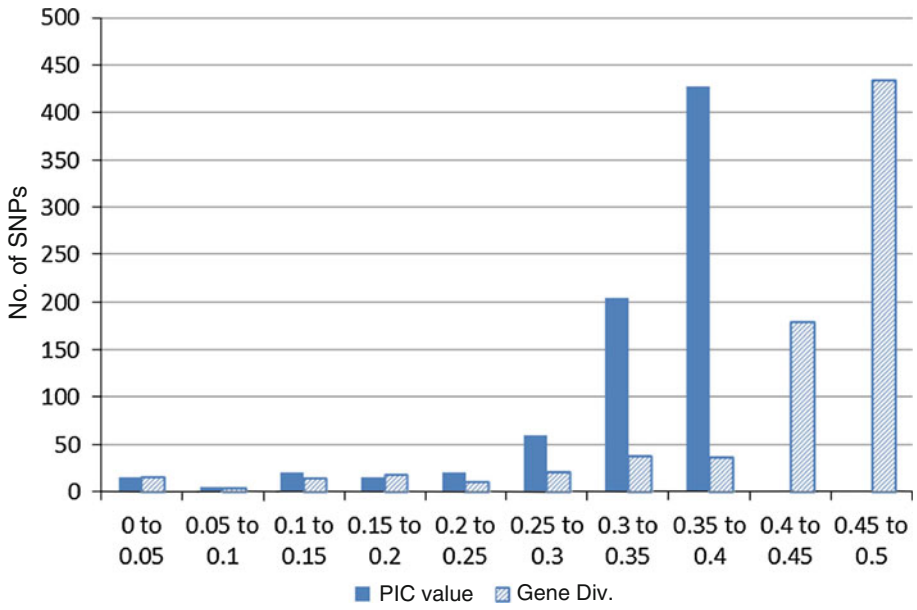

Fig. 3 Distribution of polymorphism information content (PIC) values for 768 single nucleotide polymorphism (SNP) markers used in this study

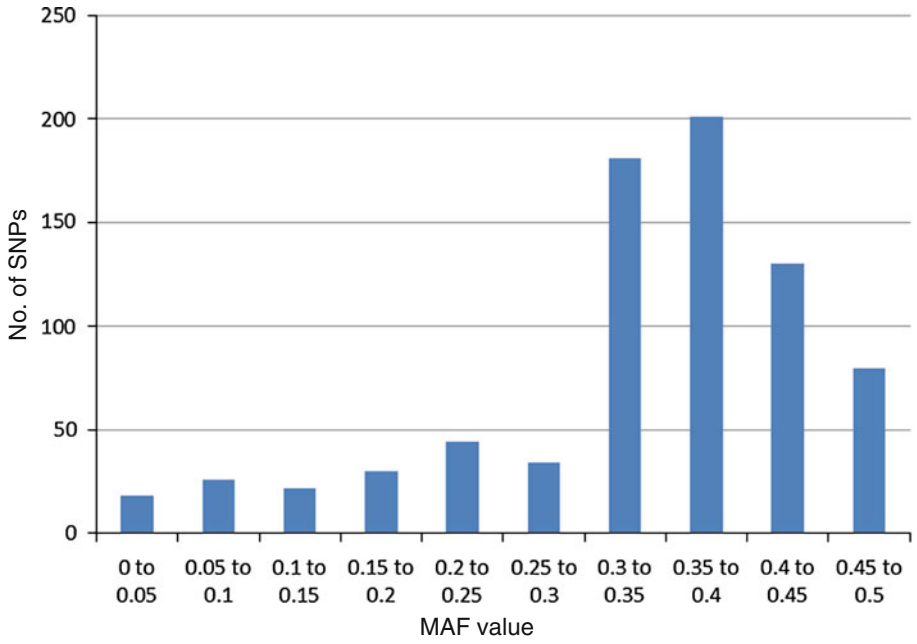

Fig. 4 Distribution of minor allele frequency (MAF) for 768 single nucleotide polymorphism (SNP) markers used in this study

TOG906490_2, and TOG917884_1) also had PIC values of zero.

As mentioned above, a total of 9 SNPs (TOG897017_1, TOG898007_1, TOG899751_1, TOG900987_2, TOG902611_1, TOG903813_1, TOG905303_1, TOG906575_1 and TOG929402_1) had the maximum PIC value of 0.375, although the theoretical maximum according to the formula of Anderson et al. (1993) would be 0.5 for a bi-allelic marker. A total of 127 other SNPs had close to the highest observed PIC values of more than 0.370. Meanwhile, the TOGs with the highest MAF values were the same as listed above for the PIC values with the highest theoretical value of 0.5 for bi-allelic markers reached for all 9 SNPs.

A repeatability test was undertaken with the two control genotypes discussed above which were used on every OPA plate that was analyzed. In total, BAT93 and Jalo EEP558 were repeated four times across the experiments, twice on each plate. In general, differences were between plates rather than between duplicates on a plate. Switched alleles, missing data or heterozygous loci were infrequent. BAT93 had 7 switched alleles (0.9 %), 10 missing data point differences (1.3 %) and 16 heterozygous versus

Table 2 Genotypes evaluated in duplicate Illumina GoldenGate OPA assays and the number of A and B alleles, missing data, successful calls and percentage successful SNPs

Genotype and genepool[a]	A (Meso)	B (Andean)	Missing data	Missing B	Missing A	Total count	No. of successful SNPs	% successful SNP
BAT477 (M)	1,246	187	25	1	0	1,433	716	98.22
BAT881 (M)	1,180	262	16	0	0	1,442	721	98.9
DOR364 (M)	1,259	177	22	1	1	1,436	716	98.22
DOR476 (M)	1,261	178	19	1	0	1,439	719	98.63
G2333 (M)	1,098	317	43	2	1	1,415	705	96.71
VAX1 (M)	1,239	197	22	5	1	1,436	712	97.67
AFR298 (A)	48	1,408	2	0	0	1,456	728	99.86
CALIMA (A)	74	1,373	11	0	1	1,447	723	99.18
G19833 (A)	130	1,312	16	0	2	1,442	719	98.63
G19839 (A)	124	1,317	17	0	1	1,441	720	98.77

[a] Genotype identification for Mesoamerican (Meso) and Andean alleles A and B, respectively, with situation of dominant markers (missing the A or B allele) indicated for the number of SNPs tested

called allele differences (2.1 %). Jalo EEP558 had similar numbers with 0.9, 0.8 and 1.7 % in these categories, respectively.

Apart from the two controls above, ten other genotypes were repeated twice in the SNP analysis across two assay plates to measure repeatability of the assay in genotypes different from those used to obtain the SNPs (Table 2). In this test, the GoldenGate assays were found to be very repeatable with no discrepancies in allele calls found for any of the genotypes in any of the duplicate pairs. The Mesoamerican control genotypes had a slightly higher number of missing data in the assay compared to the control Andean genotypes. The average rate of successful allele calls in the control Andean genotypes was 99.11 % while for control Mesoamerican genotypes it was 98.06 %, a small but consistent difference (unpaired t test, $P \leq 0.01$).

Relationships between accessions

The dendogram built with the results of the GoldenGate assay run on the minicore diversity panel showed a large genetic distance between genepools and relatively small genetic distances within genepools (Fig. 5). The wild genotypes G24404 and G24390 were located in between the two genepools, with the Colombian wild (G24404) slightly closer to the Andean genepool group and the Mexican wild (G24390) slightly closer to the Mesoamerican genepool group.

Despite the small intra-genepool genetic distances, the relationships of the genotypes within each genepool were accurate based on biological considerations, race morphology and previous classifications based on SSR diversity analysis (Blair et al. 2006a, 2009). Among the Andean genepool individuals, Jalo EEP558 anchored the group as the most distinct as would be expected since it was used to define which alleles were Andean. Nueva Granada race genotypes, AFR298, G122 and IJR shared the same branch with Jalo EEP558. Meanwhile, the advanced bush bean lines or varieties BRB191, Calima (G4494), Cerinza and SEQ1027 were slightly more distant from the control genotype. G21078 grouped close to these genotypes followed by another branch that diverged to the two Peru race genotypes, G19833 and G19839, which were closely related to each other. Finally, further down the dendogram were branches to G21657, Montcalm and G21242.

The Mesoamerican genotypes had a similar level of genetic diversity as the Andean genotypes, although some genotypes had long branch lengths from the main dendogram axis. After the wild Mexican accession, the first genotype to branch off was G685, a Guatemala race genotype, followed by the cluster of G855, G2333 and G11360, a group of climbing beans from Central America. After this G14519 was found, followed by the advanced lines SEA15 and SEA5, then a branch to BAT477 and MAR1 and another branch to BAT881 and SEA21. The inter-specific (*P. vulgaris* × *P. acutifolius*) line SEL1309 was found on a long branch from the last two genotypes. After this, one cluster was found with G3513, G11530 and G21212; and another cluster with G4825, VAX1, VAX3 and VAX6, these last three being CBB-resistant advanced lines. At the end of the dendogram were BAT93 and MAM38 on one branch, followed by the black beans DOR390, G5883 and Jamapa and then the small red beans DOR364, DOR476 and Tio Canela. The inset next to the dendogram shows the results from SSR marker analysis of the mini-core by Blair et al. (2006a), and it is interesting that both with SSRs and with the SNPs analyzed here, we found the separation of black seeded versus red-seeded subgroups for race Mesoamerica.

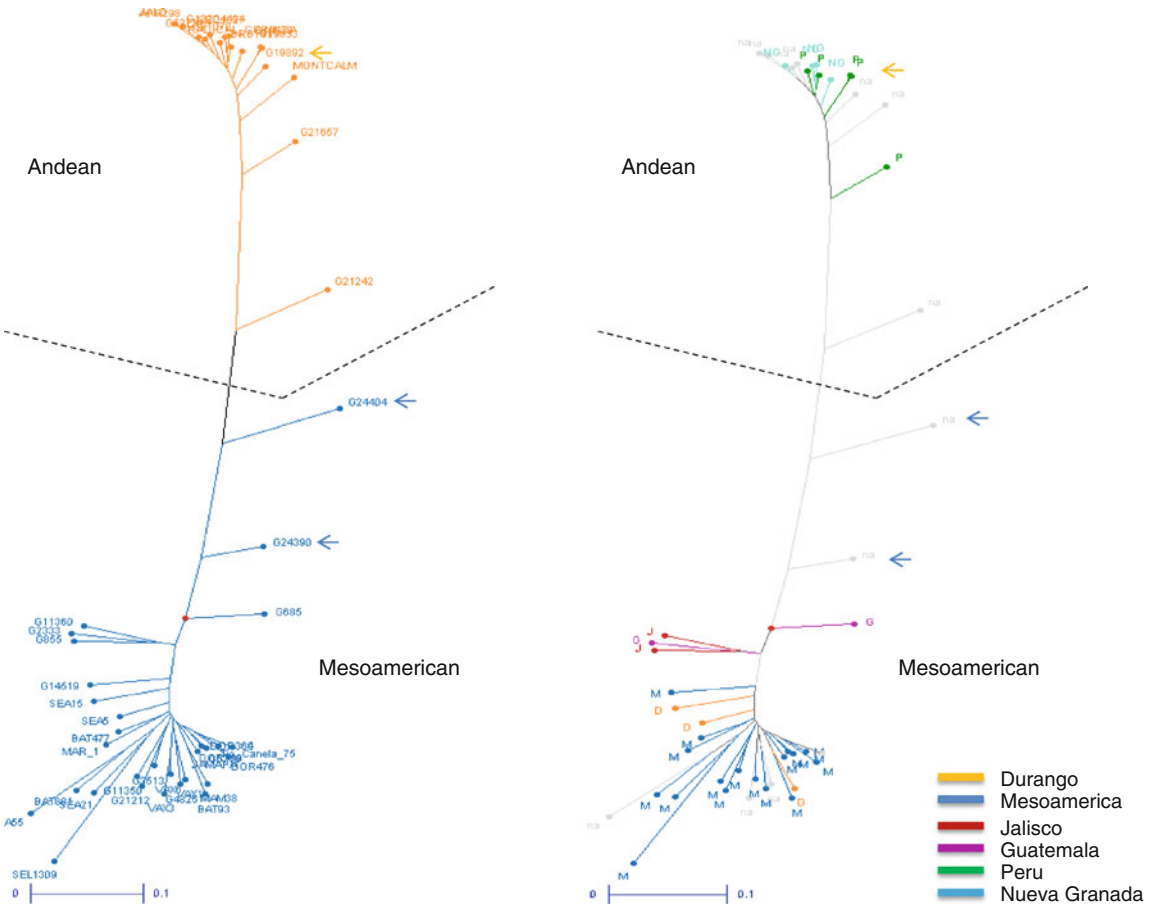

Fig. 5 Neighbour joining dendogram for 50 common bean genotypes with tree generated from genetic distance matrix for mini-core common bean genotypes used in this study and 768 single nucleotide polymorphism (SNP) markers from the GoldenGate assay. Germplasm identification and classification in **a** is according to Andean and Mesoamerican genepools in contrasting light (*orange*) and dark (*blue*) lines. Meanwhile **b** shows the identification by races as given in the *colored legend*. The race name abbreviations are: *D* Durango, *J* Jalisco, *M* Mesoamerica, *NG* Nueva Granada, *P* Peru. *Arrows* in both subfigures indicate wild accessions, with all others being cultivars. *NA* not appropriate to classify as with wild accessions (color figure online)

The genetic diversity, MAF and PIC values were much lower when considering the Andean and Mesoamerican genepools separately, which indicates the low polymorphism of the SNPs within genepools. The average genetic diversity and PIC values were 0.119 and 0.103, respectively, for the Andean genotypes alone and 0.194 and 0.162, respectively, for the Mesoamerican genepool genotypes alone. Meanwhile, the average MAF values were 0.078 and 0.135 for the Andean and Mesoamerican genepools, respectively.

Inter- and intra-genepool polymorphism levels

Polymorphism analysis was based on the studies of Blair et al. (2006a) and Cortés et al. (2011), where we were interested in specific parental combinations represented by the minicore genotypes. These included crosses within genepools (intra-genepool crosses) and between genepools (inter-genepool crosses) as described in Table 3. The latter crosses were represented by Andean × Mesoamerican, which had much higher polymorphism than the within Andean or within Mesoamerican crosses. Considering that JaloEEP558 and BAT93 were the basis for the SNP markers, we considered their alleles to represent Andean and Mesoamerican genotypic classes, respectively, for the identification of polymorphisms in the other inter-genepool combinations. Therefore, the BAT93 × Jalo EEP558 combination had the highest polymorphism levels observed with the SNPs. Apart from this cross, high polymorphism of 76.5–81 % was found for the inter-genepool crosses DOR364 × G19833, BRB191 × MAM38 and G5273 × MAM38. Analysis of the other crosses was instructive due to notable differences between inter- and intra-genepool crosses. For example, two additional inter-genepool crosses in the first part of Table 3, G855 × BRB191 and G2333 × G19839, had approximately 65 % polymorphism; while

Table 3 Level of polymorphism in parental combinations across or within Mesoamerican (M) and Andean (A) gene pools for the 786 SNP markers used in the study

	Parental combination		Type of cross	SNP markers (736)	
	Female parental	Male parental		No. of poly	% poly
Blair et al. (2006a, b)	BAT93	JaloEEP558	M(m) × A(ng)	707	96.1
	BAT881	G21212	M(m) × M(m)	122	16.8
	BRB191	MAM38	A(ng) × M(d)	568	77.2
	DOR390	G19892	M(m) × A(w)	502	68.2
	DOR364	G19833	M(m) × A(p)	563	76.5
	DOR364	BAT477	M(m) × M(m)	60	8.6
	DOR364	G3513	M(m) × M(m)	74	10.1
	DOR476	SEL1309	M(m) × M(m)	188	25.5
	Cerinza	G24404	A(p) × A(w)	37	5.0
	Cerinza	G24390	A(p) × M(w)	248	33.7
	G855	BRB191	M(j) × A(ng)	473	64.2
	G2333	G19839	M(g) × A(p)	484	65.8
	G5273	MAM38	A(ng) × M(d)	596	81.0
	G11360	G11350	M(j) × M(m)	159	21.6
	G21657	G21078	A(p) × A(p)	21	2.9
	G21078	G21242	A(p) × A(na)	242	32.9
	G14519	G4825	M(m) × M(m)	113	15.5
	VAX6	MAR1	M(m) × M(m)	98	13.3
Cortés et al. (2011)	A55	G122	M × A	478	64.9
	SEA5	CAL96	M × A	532	72.2
	SEA15	CAL96	M × A	525	71.3
	SEA5	CAL143	M × A	519	70.5
	SEA15	CAL143	M × A	507	68.9
	SEA5	BRB191	M × A	532	72.3
	SEA15	BRB191	M × A	521	70.8

Inter-genepool and inter-race combinations indicated by abbreviations, where *A* Andean, *M* Mesoamerican. Genepools followed by an additional letter in parenthesis where *d* Durango, *g* Guatemala, *j* Jalisco, *m* Mesoamerica, *ng* Nueva Granada, *p* Peru race, *w* wild accession

the intra-genepool combinations were variable but much lower (from 3 to 30 %).

A greater number of Mesoamerican × Mesoamerican (M × M) crosses were analyzed than Andean × Andean (A × A) crosses. Among the M × M crosses, DOR364 × BAT477 had the lowest polymorphism (5.0 %) and DOR476 × SEL1309 (25.5 %) had the highest. Among the A × A crosses, the inter-racial cross G21078 × G21242, where one parent was from race Nueva Granada and one from race Peru, had higher polymorphism (32.9 %) than the intra-racial cross G21078 × G21657, where both parents were from race Peru and polymorphism was very low (2.9 %).

Finally for the cultivated × wild crosses, the A × M combination of Cerinza × G24390 had intermediate to low polymorphism (33.7 %). However, the cultivated × wild, inter-genepool combination DOR390 × G19892 has higher polymorphism (68.2 %). Since this raised the question if some inter-genepool crosses might be lower in polymorphism, we analyzed an additional set of combinations evaluated by Cortés et al. (2011). The second part of Table 3 shows that the polymorphism was between 64.9 and 72.2 % for seven parental combinations of interest for drought and heat tolerance breeding, a narrow range that confirms high polymorphism for inter-genepool combinations.

Analysis of the validation germplasm set

Since original SSR analysis by Blair et al. (2006a) found the mini-core to detect more diversity in Andean genotypes than in Mesoamerican genotypes, but later SSR analysis by Blair et al. (2009) found even diversity in the two genepools, we were interested in extending the SNP analysis of the minicore into a larger validation set. With this in mind, two plates of additional genotypes (186 entries) were evaluated with the same GoldenGate assay used for the mini-core and produced a dataset of nearly 150,000 datapoints. Control genotypes were used twice per plate and consisted in the same Andean and Mesoamerican control genotypes. There were slightly more Mesoamerican genotypes than Andean genotypes in the analysis to reflect

the predominance of the former genepool among bush beans. Climbing beans, in which Andean genotypes predominate, were not analyzed as the purpose of this extra genotyping was to distinguish bush beans only. For that reason a set of advanced lines were also included.

Results of the validation germplasm set of 186 genotypes, showed that the SNP assay was effective at separating Andean and Mesoamerican genotypes as predicted in the mini-core set. Population structure analysis at $K = 2$ confirmed the clear separation of the genepools, which was the best K value based on an Evanno's test (Fig. 6). To a certain extent, the SNP assay was also useful at separating the Mesoamerican racial groups Durango–Jalisco and Mesoamerica, which were identified at $K = 3$ and in separation the Andean races Nueva Granada and Peru, which were identified at $K = 4$, although these K values were not ideal when defined by statistical test. SSR analysis from Blair et al. (2009) defined these groups with higher precision. In the SNP study, advanced lines did show mixture of the races within each genepool as would be expected since they are the products of plant breeding programs. The dendogram for these genotypes showed the tight clustering of the genotypes within each respective genepool, confirming the results of the population structure analysis.

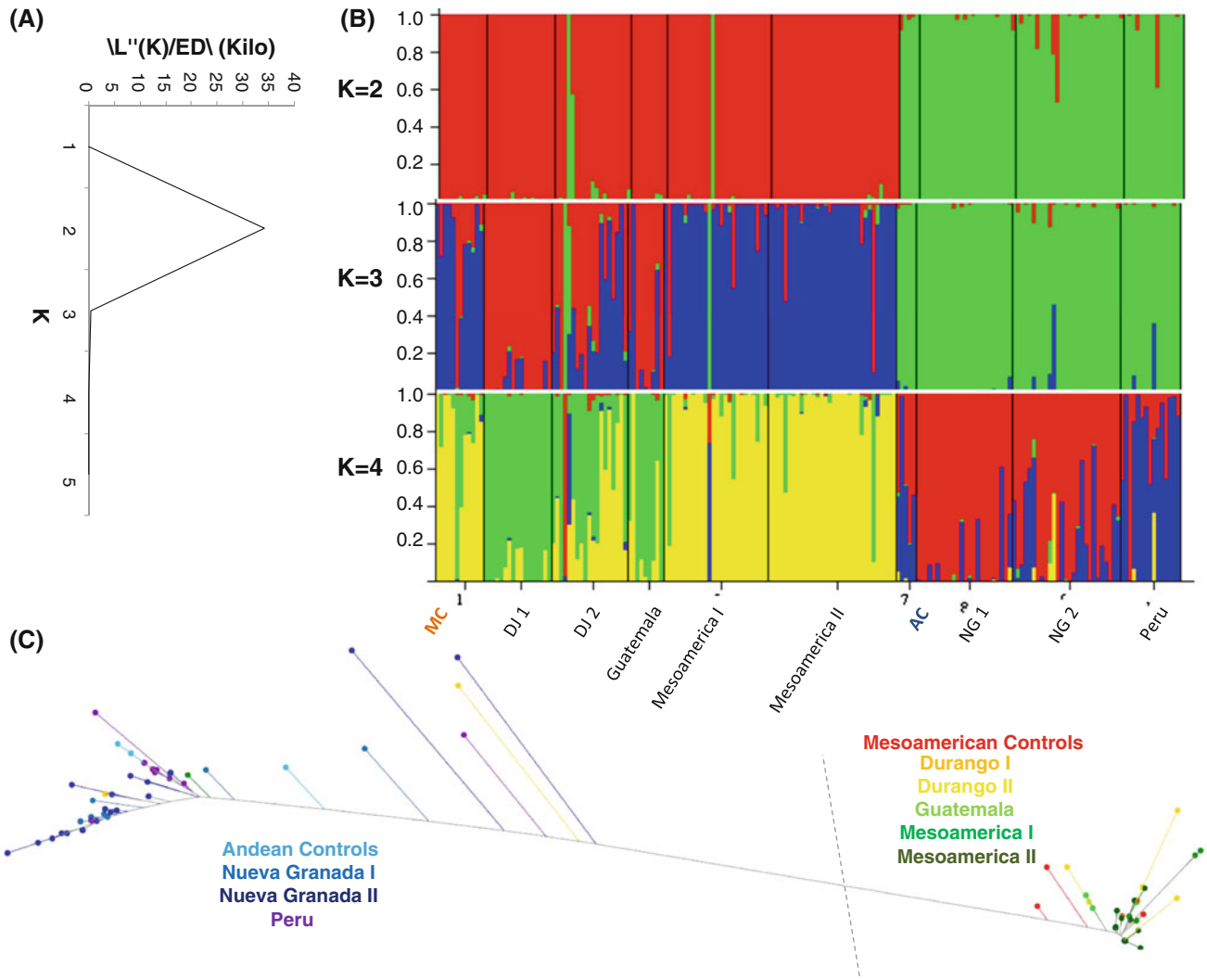

Fig. 6 Population structure and diversity analysis for 186 genotypes in the common bean validation genrmplasm set. *Subfigures* showing **a** results of Evanno's test; **b** results of 100,000 iterations of STRUCTURE software with K values from $K = 2$ to $K = 4$ with sub-populations as indicated in Blair et al. (2009) and shown with *lettering* below the graph; Abbreviations used include: *AC* Andean Controls, *DJ* Durango–Jalisco group, *MC* Mesoamerican controls, *NG* Nueva Granada race. All other races (race Mesoamerica and race Peru) are fully designated as are the subgroups of race Mesoamerica and race Nueva Granada in and **c** results of neighbor joining dendogram construction showing divisions of Andean and Mesoamerican genotypes and respective races as defined in previous subfigure and indicated in *different colors* (color figure online)

Discussion

In this study, we implemented a multi-locus platform for SNP marker analysis for diversity assessment in common bean based on the GoldenGate assay. We evaluated two diverse germplasm sets based on a mini-core of cultivars, advanced lines and wild accessions that have been used to create genetic or breeding populations as described in Blair et al. (2006a) and Cortés et al. (2011) as well as a validation set evaluated for SNPs in drought candidate genes by Cortés et al. (2012). As distinct from those studies we evaluated the distribution of genetic diversity and minimum allele frequency values for a large number of SNP loci at a time giving a new picture of gene-based diversity in the crop. Similar to the previous studies, we found that the Andean and Mesoamerican genepools are very distinct. Overall, we found the GoldenGate assay to be a useful genetic tool for rapid analysis of parental combinations, germplasm studies or for evaluation of association panels. However, SNP markers both in this study and those of Cortés et al. (2011, 2012) detected lower polymorphism compared to SSR markers analyzed by Blair et al. (2006a, 2009).

We evaluated PIC values in the mini-core set as a parameter to compare SNP markers to the previous studies using SNP and SSR markers. In this case, we found that the PIC values of the TOG markers used in this GoldenGate assay were lower than for other gene-based and genomic SNP markers (PIC values of 0.436 and 0.440, respectively) used in Kaspar SNP analysis by Cortés et al. (2011). The TOG markers were based on conserved legume sequences and therefore a slightly lower average PIC value of 0.328 was not out of the ordinary. Some studies using full core collections or large germplasm sets use low frequency alleles as a reason for eliminating SNPs from consideration, but we did not do this since we were interested in all the genes analyzed. In any case, average minor allele frequency was high at 0.336 (theoretical maximum 0.5).

In a SNP study using the Illumina GoldenGate assay in garden pea, *Pisum sativum*, a similar average minor allele frequency was found. (Deulvot et al. 2010), showing that in most germplasm sets if they are selected to be comprehensive do not have an excess of rare alleles. However, in an analysis of maize diversity, Yan et al. (2009) found that minor allele frequency was continuously distributed in two sets of inbreds from different regions, but that many SNPs were rare so they eliminated these SNPs for the purpose of estimating kinship. This appears not to be a great problem for SNPs in common bean as long as both genepools are evaluated. However, for association mapping within each genepool the extent of rare alleles should be carefully considered for common bean.

Ascertainment bias is often a problem with SNP markers where the sequence of the genotypes used to develop the assay influences the correct evaluation of alleles in other genotypes. We must take into account that the use of certain control genotypes for the development of our GoldenGate assay may have influenced the detection of SNPs and in some cases their allele calling. The fact that BAT93 is an advanced breeding line with some history of introgression breeding makes it likely that some of the SNPs discovered could be atypical of the Mesoamerican genepool that it represents. Similarly, Jalo EEP558 is a landrace from a secondary center of diversity (Brazil) rather than from the primary center of diversity for the Andean genepool, and therefore may also have atypical alleles for certain SNPs. Finally, since both control genotypes were from the cultivated genepool, in many cases SNPs from BAT93 and Jalo EEP558 were not functional in the wild genotypes used in the diversity panel, especially G24390 (Mexican Mesoamerican) and G24404 (Colombian wild) which are more divergent than G19892 (Argentinean Andean). G24404 was found to be a very distinct wild accession in SSR diversity studies (Blair et al. 2006a) and in advanced backcrossing (Blair et al. 2006b), while both G24390 and G24404 were found to be distinct in our previous SNP diversity study (Cortés et al. 2011).

Returning to the garden pea study of Deulvot et al. (2010), ascertainment bias was less of a problem in wild germplasm evaluation with SNPs developed from cultivar sequences than in seen here with common bean. Other studies in agricultural crops with GoldenGate SNPs generally have mainly considered cultivated germplasm rather than wild relatives (Yan et al. 2010; Zhao et al. 2010). The results for common bean indicate that re-sequencing will probably be needed for the development of SNP markers that are functional in wider germplasm sets, especially outside cultivated groups. In any case, re-sequencing especially with low-cost next-generation technology will be useful for SNP validation in each genepool of common bean, anyway.

Apart from the results for the wild accessions, the diversity assessment showed similar results to SNP analysis by Cortés et al. (2011). In both studies, the genetic distance between the cultivated Andean and Mesoamerican genepools was much greater than within either group of cultivars separately. However, some evidence for differences between common bean races as defined by Singh et al. (1991) was observed, especially for race Guatemala and race Jalisco versus the two Mesoamerica race subgroups in the Mesoamerican genepool, and race Nueva Granada versus race Peru in the Andean genepool. In contrast to the results of Cortés et al. (2011) and Blair et al. (2006a), the genetic diversity in the Mesoamerican genepool was slightly larger than within the Andean genepool. This is perhaps due to the difference in markers analyzed. Contrasting results have been obtained by various authors

for the comparative diversity levels found in each genepool (Benchimol et al. 2007; Blair et al. 2009; Cortés et al. 2011; Kwak and Gepts 2009). While SSR markers tend to find high diversity in each genepool and similar diversity levels, SNP markers or sequence-based analysis finds lower diversity and a tendency for higher diversity in the Mesoamerican genepool. The validation set in the present study was biased toward a greater number of Mesoamerican genotypes which had slightly higher diversity.

Race assignments in the validation set seemed to show the value of SNPs within genepool analysis, but the level of polymorphism among Andean races or among Mesomerican races was low and within genepool population structure could not be confirmed with an Evanno's test. Similarly, polymorphism levels among pairs of mini-core genotypes in this study were correlated with whether the comparison was across genepools or races. For example, SNP polymorphism was very low in parental comparisons from the same race within the Andean genepool, intermediate or low for parents from different races within the Andean or Mesoamerican genepool, but high for crosses between genepools. These results agree with those of Cortés et al. (2011), who found above 65 % average SNP polymorphism for inter-genepool comparisons, but 10 and 25 % for intra-genepool comparisons in Andean and Mesoamerican genepools, respectively. Within either genepool, the intra-race combinations appear to show lower average polymorphism than the inter-race combinations.

It was notable that the Mesoamerican alleles were over-represented with 18,063 data points, while the Andean alleles were under-represented with 13,785 data points. When we consider that in the mini-core diversity panel, 18 of the genotypes were Andean and 29 were Mesoamerican and we found that on a per genotype basis the Andean allele was slightly more frequent than the Mesoamerican allele. This may represent introgression of the Andean allele into the Mesoamerican genepool. All of these assumptions depend on the genepool purity of the control genotypes, BAT93 and Jalo EEP558 so this must be taken into account when evaluating the results.

Overall the utility of the new SNP markers was highest in inter-genepool comparisons especially for source genotypes used to develop the Illumina set (BAT93 and Jalo EEP558). Indeed the polymorphism for these two genotypes was above 90 % as would be expected since the SNP discovery process was based on these genotypes. Crosses based on genotypes that had similar origins as the control genotypes were the next most polymorphic. Surprisingly this included crosses with MAM38, which is a diverse Mesoamerican advanced line which includes various diverse parents in its pedigree. In the evaluation by Cortés et al. (2011), the most polymorphic cross evaluated with 94 SNP markers was DOR364 × G19833 (Mesoamerican × Peru races), which had an average level of polymorphism of 86.2 %, followed by other combinations between cultivated Mesoamerican and Andean beans. In that study, wild × cultivated crosses were of similar polymorphism as inter-genepool crosses. Differences between the two analyses could be due to the different sources of the SNPs, where in Cortés et al. (2011) these were from a wide range of sources.

The functional SNP markers in our GoldenGate assay were equivalent in number to the only other GoldenGate assay prepared for common bean by Hyten et al. (2010). There, a total of 827 genomic SNPs were functional and technologically successful out of 1,050 that were assayed (79 %). In comparison to our study, 736 SNPs were successful out of 768 that were assayed (96 %). These differences were perhaps due to the source of the SNPs used in our study based on conserved legume gene sequences versus the source used in Hyten et al. (2010), which was a multi-tier genomic DNA representation library based on a 454 next-generation sequencing run.

The high success rate observed with the TOG-based markers in the GoldenGate assay is comparable to evaluations in *Pisum sativum*, where re-sequencing was used to create a 384 SNP set evaluated with the Beadexpress platform from Illumina and success rate was over 92 % (Deulvot et al. 2010). Finally, reproducibility of the present GoldenGate assay was also very good due to the high quality of the SNPs. All the allele calls were in agreement between duplicates and the most widespread genotyping error was between a parental allele versus a heterozygous call. This is a low prevalence of errors, which is typical of other studies using GoldenGate assays as well (Yan et al. 2009). Comparisons of SNP technologies can now be made with this assay versus other technologies.

Other technological options for SNP analysis include Taqman-based applications such as Kaspar and other oligonucleotide applications such as single-base extension assays. In terms of failure rate, the GoldenGate assay had missing data at a similar rate (2.7 %) as the Kaspar analysis (2.5 %) of Cortés et al. (2011). Cost comparisons would be slightly higher for the Kbiosciences Kaspar assay compared to the Illumina GoldenGate assay per datapoint, but lower per genotype due to the flexibility in the number of SNPs evaluated per genotype. The requirement of GoldenGate is to evaluate a full array of SNPs at a time and therefore targeted SNP genotyping as for genetic map construction would be better done on a Kaspar platform after polymorphism evaluation on either system. The GoldenGate assay would be ideal for large diverse germplasm sets, for example in association mapping studies (Blair et al. 2009). For this kind of analysis, gene-based SNPs would be ideal.

In summary, we developed an Illumina GoldenGate assay of 768 SNP markers for common bean from

conserved legume genes and applied this to genomics tool to genotyping and diversity analysis of a mini-core germplasm set. Parental comparisons made in this study were representative of the types of parental combinations used in common bean genetics research, and show the value of the recently developed SNPs for efficient genetic analysis of common bean especially for inter-genepool crosses and between genepool comparisons of cultivars. Therefore, we have shown that conserved gene sequences are useful in uncovering polymorphisms in common bean, which is useful for defining polymorphisms found in introgressions from one genepool to another. The true advantage of these markers is that they form cross-comparable markers between the multiple grain legumes which is important for synteny mapping and for having fixed landmarks at even distances within the genomes of various legumes including less-well studied orphan crops.

A final conclusion is that although gene-based SNPs are not ideal for intra-genepool or intra-race crosses where SSRs are more informative, they are useful for any combination of inter-genepool parents. The low mutation rate of gene-based SNPs allows us to be assured that the same transition or transversion event does not occur at the same nucleotide position, allowing confident analysis of alleles and haplotypes in diversity studies. Furthermore, the probability that the same allele at the inter-genepool level comes from a common ancestor (identity by descent) is high. This makes the high-throughput GoldenGate assay a good platform for genetic analysis of breeding line pedigrees. In conclusion, the development of the GoldenGate assay for common bean will facilitate various genetic analyses to be carried out on the crop.

Acknowledgments The authors wish to thank Lucy M. Díaz and Carolina Chavarro for DNA extraction and technical work, as well as Agobardo Hoyos and Alcides Hincapie for seed multiplication and greenhouse management. This research was part of the Tropical Legume project (to MWB as PI for common beans and DRC as PI for comparative legume genomics). Additional funding was from the US National Science Foundation award DBI 0605251 (to DRC).

References

Afanador LK, Hadley SD, Kelly JD (1993) Adoption of a mini-prep DNA extraction method for RAPD marker analysis in common bean. Bean Improv Coop 35:10–11

Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181–186

Benchimol LL, de Campos T, Carbonell SAM, Colombo CA, Chioratto AF, Formighieri EF, Gouvêa LRL, de Souza AP (2007) Structure of genetic diversity among common bean (*Phaseolus vulgaris* L.) varieties of Mesoamerican and Andean origins using new develop microsatellite markers. Genet Resour Crop Evol 54:1747–1762

Blair MW, Giraldo MC, Buendia HF, Tovar E, Duque MC, Beebe SE (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 × wild common bean (*Phaseolus vulgaris* L.) cross. Theor Appl Genet 112:1149–1163

Blair M, Diaz LM, Buendía HF, Duque MC (2009) Genetic diversity, seed size associations and population structure of a core collection of common beans (*Phaseolus vulgaris* L.). Theor Appl Genet 119:955–972

Broughton WJ, Hernández G, Blair MW, Beebe S, Gepts P, Vanderleyden J (2003) Beans (*Phaseolus* spp.) model food legumes. Plant Soil 55:55–128

Chagné D, Batley J, Edwards D, Forster JW (2007) Single nucleotide polymorphism genotyping in plants. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, Silva HNd (eds) Association mapping in plants, pp 77–94

Cortés A, Chavarro C, Blair MW (2011) SNP marker diversity in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet. doi: 10.1007/s00122-011-1630-8

Cortés AJ, This D, Chavarro MC, Madriñan S, Blair MW (2012) Nucleotide diversity patterns at the drought related DREB encoding genes in wild and cultivated common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 125(5):1069–1085

Deulvot C, Charrel H, Marty A, Jacquin F, Donnadieu C, Lejeune-Hénaut I, Burstin J, Aubert G (2010) Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. BMC Genomics 11:468

Evano 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

Galeano CH, Fernández AC, Gómez M, Blair MW (2009a) Single strand conformation polymorphism based SNP and Indel markers for genetic mapping and synteny analysis of common bean (*Phaseolus vulgaris* L.). BMC Genomics 10:629

Galeano CH, Gomez M, Rodriguez LM, Blair MW (2009b) CEL I nuclease for SNP discovery and marker development in common bean (*Phaseolus vulgaris* L.). Crop Sci 49:381–394

Gepts P, Aragao F, Barros E, Blair MW, Brondani R, Broughton W, Hernández G, Kami J, Lariguet P, McClean P, Melotto M, Miklas P, Pedrosa-Harand A, Porch T, Sánchez F (2008) Genomics of Phaseolus beans, a major source of dietary protein and micronutrients in the tropics. In: Moore PH, Ming R (eds) Genomics of tropical crops, chap 5. Springer, Berlin, pp 113–143

Hyten DL, Song Q, Choi IY, Yoon MS, Specht JE, Matukumalli LK, Nelson RL, Shoemaker RC, Young ND, Cregan PB (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116(7):945–952

Hyten DL, Song Q, Fickus EW, Quigley CV, Lim J-S, Choi I-Y, Hwang E-Y, Pastor-Corrales M, Cregan PB (2010) High-throughput SNP discovery and assay development in common bean. BMC Genomics 11:475

Kwak M, Gepts P (2009) Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). Theor Appl Genet 118:979–992

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

McConnell M, Mamidi S, Lee R, Chikara S, Rossi M, Papa R, McClean P (2010) Syntenic relationships among legumes revealed using a gene-based genetic linkage map of common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 121:1103–1116

Muchero W, Diop NN, Bhat PR, Fenton RD, Wanamaker S, Pottor M, Hearne S, Cisse N, Fatokun C, Ehlers JD, Roberts PA, Close TJ (2009) A consensus genetic map of cowpea [*Vigna unguiculata*

(L) Walp.] and synteny based on EST-derived SNPs. Proc Natl Acad Sci USA 106:18159–18164

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small numbers of individuals. Genetics 89:583–590

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS (2002) BeadArray technology: enabling an accurate, cost-effective approach to high throughput genotyping. Biotechniques Suppl 5:6–58

Perrier X, Flori A, Bonnot F (2003) Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC (eds) Genetic diversity of cultivated tropical plants. Enfield Science Publishers, Montpellier, pp 43–76

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Singh SP, Gepts P, Debouck DG (1991) Races of common bean (*Phaseolus vulgaris*, Fabaceae). Econ Bot 45:379–396

Yan JB, Shah T, Warburton M, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS ONE 4(12):e8451

Yan JB, Yang XH, Shah T, Sanchez-Villeda H, Li JS, Warburton M, Zhou Y, Crouch JH, Xu YB (2010) High-throughput SNP genotyping with the GoldenGate assay in maize. Mol Breed 25:441–451

Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Liakat Ali ML, Tung C-W, Reynolds A, Bustamante CD, McCouch SR (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. PLoS ONE 5(5):e10780