


SNP discovery in common bean by restriction-associated DNA (RAD) sequencing for genetic diversity and population structure analysis

Paula Arielle M. R. Valdisser¹ · Georgios J. Pappas Jr.⁵ · Ivandilson P. P. de Menezes² · Bárbara S. F. Müller³ · Wendell J. Pereira⁴ · Marcelo G. Narciso¹ · Claudio Brondani¹ · Thiago L. P. O. Souza¹ · Tereza C. O. Borba¹ · Rosana P. Vianello¹ 

Received: 24 August 2015 / Accepted: 12 February 2016 / Published online: 1 March 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Researchers have made great advances into the development and application of genomic approaches for common beans, creating opportunities to driving more real and applicable strategies for sustainable management of the genetic resource towards plant breeding. This work provides useful polymorphic single-nucleotide polymorphisms (SNPs) for high-throughput common bean genotyping developed by RAD (restriction site-associated DNA) sequencing. The RAD tags were generated from DNA pooled from 12 common bean genotypes, including breeding lines of different gene pools and market classes. The aligned sequences identified 23,748 putative RAD-SNPs, of which 3357 were adequate for genotyping; 1032 RAD-SNPs with the highest ADT (assay design tool) score are presented in this article. The RAD-SNPs were structurally

annotated in different coding (47.00 %) and non-coding (53.00 %) sequence components of genes. A subset of 384 RAD-SNPs with broad genome distribution was used to genotype a diverse panel of 95 common bean germplasm and revealed a successful amplification rate of 96.6 %, showing 73 % of polymorphic SNPs within the Andean group and 83 % in the Mesoamerican group. A slightly increased H_e (0.161, $n = 21$) value was estimated for the Andean gene pool, compared to the Mesoamerican group (0.156, $n = 74$). For the linkage disequilibrium (LD) analysis, from a group of 580 SNPs (289 RAD-SNPs and 291 BARC-SNPs) genotyped for the same set of genotypes, 70.2 % were in LD, decreasing to 0.10 % in the Andean group and 0.77 % in the Mesoamerican group. Haplotype patterns spanning 310 Mb of the genome (60 %) were characterized in samples from different origins. However, the haplotype frameworks were under-represented for the Andean (7.85 %) and Mesoamerican (5.55 %) gene pools separately. In conclusion, RAD sequencing allowed the discovery of hundreds of useful SNPs for broad genetic analysis of common bean germplasm. From now, this approach provides an excellent panel of molecular tools for whole genome analysis, allowing integrating and better exploring the common bean breeding practices.

Communicated by R. K. Varshney.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-016-1182-3) contains supplementary material, which is available to authorized users.

✉ Rosana P. Vianello
rosana.vianello@embrapa.br

¹ Laboratório de Biotecnologia, Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, Brazil

² Laboratório de Genética e Biologia Molecular, Departamento de Biologia, Instituto Federal Goiano (IF Goiano), Urutá, GO, Brazil

³ Programa de Pós-Graduação em Biologia Molecular, Universidade de Brasília (UnB), Brasília, DF, Brazil

⁴ Programa de Pós-Graduação em Ciências Biológicas, Universidade Federal de Goiás (UFG), Goiânia, GO, Brazil

⁵ Laboratório de Biologia Molecular, Departamento de Biologia Celular, Universidade de Brasília (UnB), Brasília, DF, Brazil

Keywords Leguminosae · *Phaseolus vulgaris* · SNP discovery · Linkage disequilibrium · Haplotype blocks

Introduction

Many are the practical applications of genomics, such as the use of molecular tools aiming at individual identification, quantification of genetic variability in diverse germplasm and assisted selection for common bean disease (Blair et al.

2007, 2013; Cardoso et al. 2013, 2014). These applications are being incorporated routinely in several breeding programs for common bean, as a new paradigm with direct and quantifiable impact on the optimization of processes to accelerate the cycles of selection, as well as, the management and enrichment of genetic diversity in breeding populations. Common bean genomic approaches have the opportunity to reach new levels since the availability of the whole genome sequence of two important genotypes representatives of the Andean and Mesoamerican gene pools (Schmutz et al. 2014; Vlasova et al. 2016), opened new perspectives for the development of molecular tools with a high aggregated genetic value for plant breeding. Taking advantage of these advances, the SNP markers are useful in turning the approaches of whole-genome genotyping more accessible and affordable. Genome-wide SNP discovery by re-sequencing efforts has been performed in important crops and applied for different purposes such as genome-wide diversity studies, association mapping (Zhou et al. 2015) and genotyping in a genomic selection context (Poland et al. 2012; Crossa et al. 2013; Jarquin et al. 2014). For common bean, the continuous efforts to develop molecular tool useful in capturing the broad genetic diversity are moving towards the effective genomic assisted breeding for agronomic and quality traits (Gepts 2014; Rodriguez et al. 2015).

An increasing number of single-nucleotide polymorphism (SNP) markers for the common bean (*Phaseolus vulgaris* L.) crop, most derived from the polymorphic inter-Andean and Mesoamerican gene pools, have been described in the last ten years. The first effort on SNP discovery was based on the analysis of expressed sequence tags (ESTs) (Ramírez et al. 2005). Later, Gaitán-Solís et al. (2008) used the strategies of target genome sequencing and complexity reduction to identify a SNP set. Subsequently Galeano et al. (2009a) developed SNPs from SSCP (single strand conformation polymorphisms) markers and mismatch cleavage nuclease (Galeano et al. 2009b). McConnell et al. (2010) generated over a thousand gene-based SNP markers for the common bean. The first study that discovered a large number of SNPs (3487 SNPs) in the common bean was conducted by Hyten et al. (2010) by high-throughput DNA sequencing and was based on a reduced representation library. Souza et al. (2012) compared sequence-tagged sites (STSs) from a group of six diverse genotypes to identify a new set of 677 SNPs for genetic analysis. Based on conserved inter-gene pool gene sequences, Blair et al. (2013) evaluated hundreds of SNPs and determined their usefulness for genetic analysis and mapping saturation. More recently, Müller et al. (2015) filtered the SNP bank made available by Hyten et al. (2010) to genotype and perform a comparative analysis with SSR (simple sequence repeats) to estimate genetic diversity and population structure.

Based on SNP genotyping data, several studies have concluded that most SNP markers derived from the inter-gene pool accessions are useful for distinguishing Andean and Mesoamerican genotypes but are less useful for distinguishing within each gene pool. The tendency for the sampled accessions to cluster and follow the same pattern of differentiation of the target genotypes used to derive the SNPs is a problem associated with SNP development when only a few individuals are used for this purpose, as reported for barley (Moragues et al. 2010), rice (Thomson et al. 2012) and the common bean (Blair et al. 2013; Müller et al. 2015). This problem is a consequence of ascertainment biases existing in the panel of SNPs developed when exploring polymorphisms between genotypes used for their development. For the common bean, although SNP assay success has been high, the proportion of polymorphic SNPs tends to be high when contrasting Andean and Mesoamerican gene pools are considered, and decreases as genetic diversity is accessed within the gene pool (Cortés et al. 2011; Blair et al. 2013). In this context, depending on the purpose of the study and the origin of germplasm evaluated, different sets of SNPs should be developed, as previously proposed by Nagasaki et al. (2010).

The rapid and increasing advance in SNP marker discovery and genotyping has largely been facilitated by progress in new DNA sequencing technologies, the next-generation sequencing (NGS) (Margulies et al. 2005; Mardis 2008). The available NGS platforms reduced the DNA sequencing costs and time, requiring lower DNA quantities and increasing the efficiency and accuracy of high-throughput DNA sequencing (Ganal et al. 2009). However, the cost fully sequence individual genomes is still prohibitive, mainly for plants/crops with lower economic value and, consequently, with scarce financial support for genomic research. Accessible technologies that allow broad sequence polymorphism genotyping for a larger set of samples that is robust and cost-effective are desirable. The application of methods that reduce genome complexity followed by deep sequencing has been successfully used for SNP discovery in different organisms/plant species (Baird et al. 2008; Etter et al. 2011). Restriction site-associated DNA sequencing (RAD-seq) is a reduced-representation sequencing method that uses NGS to generate hundreds of thousands of short sequence tags to simultaneously discover and score SNP markers in hundreds of individuals for a reasonable cost (Willing et al. 2011). RAD-seq has been used to identify SNPs that are useful for population genetics and QTL mapping studies in eggplant (Barchi et al. 2011), barley (Chutimanitsakun et al. 2011), sunflower (Talukder et al. 2014), and other crops. In this work, we used RAD-seq to discover polymorphic SNPs across a set of 12 diverse common bean genotypes, followed by the evaluation of common bean genetic diversity and structure

using the Golden Gate genotyping system (Illumina). Additionally, we assessed the potential of RAD for direct genotyping-by-sequencing in the common bean. In conclusion, 1,032 RAD-SNPs were identified in common bean, aligned in coding and noncoding sequences and positioned along the bean genome. For a subset of 384 RAD-SNPs, $\geq 96\%$ presented a successful amplification rate and were broadened evaluated for a diverse set of genetic parameters, linkage disequilibrium, giving important insights to be used for the breeding programs. After integration with the previously developed SNPs (totaling 673 SNPs), under the perspective of the common bean breeding programs, the operational SNP panels made available provide the means to characterize the diverse common bean germplasm at the DNA level addressing different research questions. Please check and confirm Ganal (2009) has been changed to Ganal et al. (2009) so that this citation matches the list. Thanks. The correct is Ganal et al. (2009)

Materials and methods

Plant material and DNA extraction

A set of 12 common bean genotypes from the EMBRAPA Rice and Beans gene bank were used for RAD-seq development and included nine commercial cultivars and three breeding lines introduced by research institutions from Brazil and abroad (Supplementary Material 1). Total genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Hilden, NRW, Germany) according to the manufacturer's protocol. To perform SNP genotyping derived from RAD-seq, we sampled 95 common bean genotypes, including 45 commercial cultivars and 14 breeding lines developed by research institutions from Brazil (47) and abroad (12), 34 Brazilian landraces and two wild accessions, all belonging to Mesoamerican (74 genotypes) and Andean (21 genotypes) gene pools (Supplementary Material 1). Genomic DNA was isolated as described by Grattapaglia and Sederoff (1994). For both sets of samples, quantifications were performed in a Qubit[®] spectrophotometer (Invitrogen, Carlsbad, CA, USA).

RAD library development and sequencing

For SNP discovery, high quality DNA samples from the 12 selected genotypes above were divided in two equimolar pools with six individuals. The RAD-seq experiment was carried out by FLORAGENEX, Inc (Oregon, USA). In brief, the RAD-seq reduced representation libraries were prepared following digestion using the *Pst*I enzyme, followed by adaptor/barcode ligation, DNA purification and selective DNA amplification and size selection. The two resulting RAD libraries from each pool were sequenced

on an Illumina HiSeq platform (Illumina, San Diego, CA, USA) using 75 bp paired-end reads.

Identification of SNPs

RAD-seq derived sequencing reads were subjected to a series of steps in order to obtain a high quality SNP set, geared toward elimination of false positives and assay success for the Golden Gate Genotyping Technology (GGGT, Illumina) (Grattapaglia et al. 2011). The reads were submitted to systematic in silico SNP filtering leveraging the availability of a complete genome sequence of the Mesoamerican variety BAT93 (Vlasova et al. 2016). The following filtering scheme was used to retain polymorphic sites adhering to the following constraints: (1) presence in non-repeat region of the genome (avoiding transposable elements); (2) reads uniquely mapping in the reference genome (to avoid non-specific amplification); (3) only bi-allelic SNPs were maintained; (4) only SNPs with an allele frequency ranging from 40 to 60 %; and (5) adjacent SNPs separated by at least 50 kb along the genome. In the end, the selected set of SNPs was submitted to the proprietary Assay Design Tool (ADT, <http://icom.illumina.com/Custom/Index>) to select the subset of SNPs with the highest recommended design scores. The list of the sequences from selected primers and their positions in the BAT93 genome assembly is presented in Supplementary Material 2.

SNP genotyping

Based on the high quality SNPs above, a subset of 384 loci were used to synthesize an Oligo Pool Assay (OPA VC0013574) specific for use with the Illumina BeadXpress genotyping platform (Kim and Misra 2007), according to the manufacturer's instructions. Automatic allele calling was performed using Genome Studio version 1.8.4 (Illumina). The individual SNP genotyping performance was assessed by the call rate using a GeneCall score cutoff ≥ 0.25 . The clustering quality was performed by the GenTrain algorithm, provided as a module in the proprietary software Genome Studio (Illumina).

Genetic diversity and cluster analysis

The SNPs were characterized for genetic diversity by genotyping 95 common bean genotypes (Supplementary Material 1) and estimating several parameters, such as number of alleles (A), number of alleles per locus (\bar{A}), number of private alleles (A_p), expected heterozygosity or gene diversity (H_E), observed heterozygosity (H_O), fixation index (F_{IS}), genetic probability of identity (PI) and genetic probability of exclusion (PE) for each locus and multiloci, using the program GenAlexv6.5 (Peakall and Smouse 2012). A

cluster analysis using simple matching dissimilarity based on neighbor-joining with 1000 bootstrap replicates was obtained by DARwin v5.0.158 (Perrier and Jacquemoud-Collet 2006). All statistical analyses (overall significances) were performed using the non-parametric Wilcoxon test implemented by the Real Statistics Resource Pack software (Release 4.3) (Zaiontz 2013).

Population structure analysis

The Principal Coordinates Analysis (PCoA) and the genetic differentiation index (F_{ST}) were estimated with a confidence interval of 95 % for 9999 permutations, calculated using GenAlex v6.5 (Peakall and Smouse 2012). Analysis of PCoA was calculated using a simple matching dissimilarity matrix. The program STRUCTURE v2.3.4 (Pritchard et al. 2000) was used to infer the genotypes structure. A population number (K) ranging from 1 to 10 was assumed, with 10 interactions each, to identify the K with the highest log likelihood. In each run, the admixture model was applied under the condition of 1,000,000 Markov chain Monte Carlo (MCMC) replications followed by 500,000 burn-in periods. The criteria proposed by Evanno et al. (2005) was used to determine the most probable K through the program STRUCTURE HARVESTER v0.6.93 (Earl and vonHoldt 2012), followed by analysis using the software CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007) to determine the consensus among the 10 most probable K interactions. The program DISTRUCT v1.1 (Rosenberg 2004) was used for the graphical display of the population structure. An analysis of molecular variance (AMOVA) was performed to test the structure of the genetic diversity of the genotypes through the program Arlequin v3.5 (Excoffier and Lischer 2010), with 10,000 permutations.

Linkage disequilibrium

Linkage disequilibrium (LD) was estimated using Tassel v5.0 (Bradbury et al. 2007; Glaubitz et al. 2014). The standardized disequilibrium coefficient r^2 (parameter $r^2 > 0.10$) and corresponding p -values (two-sided Fisher's exact test) were estimated. For the calculation of LD, SNP loci with MAF < 0.05 (Minor Allele Frequency) were discarded from the analysis. The heterozygous genotypes were considered missing data, and the loci were considered to have a significant LD if $r^2 > 0.10$ and p value < 0.001. Haploview 4.2 (Barrett et al. 2005) was used to assess haplotypes based on 363 SNPs with MAF > 0.05 and missing data ≤ 25 %. The heterozygous alleles were considered missing data. Only SNPs incorporated up to a distance of 20 Mb were considered for the determination of haplotype blocks and D' and r^2 with confidence intervals according to Gabriel et al. (2002). The LD parameters and haplotype

blocks were calculated for the entire sample (95 genotypes) and, separately, for the Andean (21 genotypes) and Mesoamerican (74 genotypes) gene pools. In addition, the haplotype blocks were estimated using the RAD-SNP loci made available by this work and the BARC-SNP loci reported by Müller et al. (2015), totaling 708 SNPs (363 RAD-SNPs and 345 BARC-SNPs) analyzed using a common set of 88 genotypes.

Alignment of the SNPs in the *P. vulgaris* genome

Using BLASTN (Altschul et al. 1997) with E-value $\leq 1.0E-10$, the flanking sequences of the SNPs were aligned against the genome of *P. vulgaris*, Andean variety (G19833) (Schmutz et al. 2014). Annotation routines were run with the BLAST2GO tool (Conesa et al. 2005), and the ESTs were aligned to the non-redundant (Nr) *Viridiplantae* protein database at NCBI (<http://www.ncbi.nlm.nih.gov/>) and categorized by Gene Ontology (GO) terms.

Results

RAD-SNP performance

A total of 23,748 putative SNPs were identified between the two groups derived from the RAD-seq libraries. From these identified SNPs, approximately 58 % proved to be transitions, and a transition/transversion ratio of 1.40 was estimated. By applying successive filters, a set of 3357 (14 %) SNPs were adequate for Illumina Golden Gate genotyping, and 1032 SNPs with the highest score attributed by ADT (assay design tool) are available in this article (Supplementary Material 2). A sample of 384 SNPs, selected based on genome coverage, was assembled into a GoldenGate assay and used to genotype 95 common bean germplasms.

Of the 384 SNPs evaluated in this study, 371 (96.61 %) presented successful amplification with a call rate ranging from 0.6198 to 0.9917 and GenTrain ≥ 0.2620 , as provided by Genome Studio (Ritchie et al. 2011); while 363 SNPs (94.53 %) were polymorphic and eight (2.08 %) were monomorphic considering the 95 common bean genotypes. Of the 13 SNPs excluded from analyses, three had amplification failure and 10 did not have well-defined clusters. An analysis of the genotypes grouped by gene pool showed 264 (68.75 %) polymorphic SNPs within the Andean group and 302 (78.64 %) in the Mesoamerican group. With the greatest allele frequency estimated at 0.728, the mean H_E calculated was 0.384 ± 0.004 , ranging from 0.093 (EMBRA_1.0_Pv01_49615062_A_G) to 0.506 (EMBRA_1.0_Pv01_6538061_A_G and EMBRA_1.0_Pv05_5947407_T_C). The average estimates of H_E for the Brazilian germplasm ($n = 81$) was 0.372 and significantly

Table 1 Descriptive statistics of the 363 SNPs characterized in 95 genotypes of the common bean

Group	<i>S</i>	<i>P</i> (%)	NAP	H_O (SE)	H_E (SE)	F_{IS} (SE)	F_{ST} (SE)	PI	PE
Mesoamerican	74	83.3	99	0.007 ± 0.001	0.156 ± 0.008	0.923 ± 0.011	0.656 ± 0.011	4.1×10^{-52}	1
Andean	21	72.7	57	0.016 ± 0.002	0.161 ± 0.008	0.862 ± 0.015		3.9×10^{-52}	1
Total	95	100	156	0.009 ± 0.001	0.384 ± 0.004	0.976 (0.003)		1.6×10^{-123}	1

The sample size (*S*), percentage of polymorphic loci (*P*), number of private alleles (NAP), observed heterozygosity (H_O), gene diversity (H_E), fixation index (F_{IS}), genetic differentiation (F_{ST}), probability of identity (PI) and probability of exclusion (PE) are presented

Table 2 Comparison between the descriptive statistics for the 363 RAD-SNPs and 345 BARC-SNPs characterized for 88 genotypes of the common bean

Gene pool	Group	<i>N</i>	<i>A</i>	\bar{A} (SE)	Ap	H_E (SE)	H_O (SE)	F_{IS} (SE)	PI
363 SNPs-RAD									
Andean	Cultivar/Line	10	534	1.47 (0.03)	16	0.127 (0.008)	0.018 (0.003)	0.755 (0.023)	1.4×10^{-39}
	Landrace	11	611	1.70 (0.02)	93	0.195 (0.009)	0.015 (0.002)	0.895 (0.014)	1.5×10^{-60}
	Total	21	627	1.72 (0.02)	61	0.163 (0.008)	0.016 (0.002)	0.862 (0.015)	3.9×10^{-52}
Mesoamerican	Cultivar/Line	43	617	1.70 (0.02)	29	0.152 (0.008)	0.007 (0.001)	0.906 (0.013)	7.1×10^{-49}
	Landrace	22	563	1.56 (0.02)	18	0.131 (0.009)	0.008 (0.001)	0.869 (0.016)	2.1×10^{-41}
	Wild	2	451	1.20 (0.02)	21	0.170 (0.015)	0.003 (0.002)	0.971 (0.010)	–
	Total	67	665	1.83 (0.02)	99	0.162 (0.008)	0.007 (0.001)	0.917 (0.011)	2.1×10^{-52}
Grand total	All	88	726	2.00 (0.00)	–	0.396 (0.004)	0.010 (0.001)	0.975 (0.003)	6.1×10^{-127}
345 SNPs-BARC ^a									
Andean	Cultivar/Line	10	442	1.28 (0.02)	0	0.083 (0.008)	0.027 (0.008)	0.712 (0.032)	1.20E–24
	Landrace	11	403	1.17 (0.02)	0	0.056 (0.007)	0.023 (0.008)	0.704 (0.036)	7.60E–17
	Total	21	469	1.36 (0.03)	13	0.072 (0.007)	0.025 (0.008)	0.777 (0.028)	3.20E–22
Mesoamerican	Cultivar/Line	43	676	1.96 (0.01)	1	0.148 (0.007)	0.022 (0.007)	0.915 (0.016)	2.60E–45
	Landrace	22	453	1.31 (0.02)	0	0.088 (0.008)	0.022 (0.007)	0.791 (0.028)	7.20E–27
	Wild	2	428	1.24 (0.02)	1	0.168 (0.015)	0.022 (0.007)	0.798 (0.028)	–
	Total	67	677	1.96 (0.01)	221	0.142 (0.008)	0.022 (0.007)	0.914 (0.016)	1.10E–43
Grand total	All	88	690	2.00 (0.00)	–	0.390 (0.004)	0.022 (0.006)	0.937 (0.014)	4.70E–119

The sample size (*N*), number of alleles (*A*), number of alleles per locus (\bar{A}), number of private alleles (Ap), gene diversity (H_E), observed heterozygosity (H_O), fixation index (F_{IS}) and probability of identity (PI) are presented

^a Müller et al. (2015)

different ($p \leq 0.05$) from the overall value ($n = 95$; $H_E = 0.384$). For the Andean gene pool ($n = 21$), the H_E value was estimated at 0.161, while for the Mesoamerican gene pool ($n = 74$), it was 0.156. Considering only the cultivated/breeding lines ($n = 47$), the total H_E was 0.316 for the Brazilian (Andean $n = 7$, $H_E = 0.139$ and Meso $n = 40$, $H_E = 0.135$) and significantly lower ($p \leq 0.05$) than the abroad accessions ($n = 12$; $H_E = 0.438$). Considering only the Brazilian landraces ($n = 34$), the total H_E was 0.428 (Andean $n = 11$, $H_E = 0.195$ and Meso $n = 23$, $H_E = 0.130$). With regard to the H_O and F_{IS} indices, the values were estimated at 0.009 ± 0.001 and 0.976 ± 0.003 , respectively, for the entire set of samples (Table 1).

Considering the germplasm grouped by origin, of the 363 polymorphic SNPs, a set of 156 differentiated the Andean (57 private alleles) from the Mesoamerican group

(99 private alleles) based on the allelic pattern. The power of exclusion estimated was high, reaching 100 % (a set of 131 SNPs differentiate all genotypes). For the 363 SNPs, the combined PI was estimated at 1.6×10^{-123} (Table 1), with individual values ranging from 0.375 to 0.829. The genetic differentiation (F_{ST}) between the Andean and Mesoamerican population was estimated as 0.656 ($p < 0.01$). A comparative analysis of descriptive statistics performed between the SNPs previously published by Müller et al. (2015) and the data presented in this work for the same set of 88 accessions of common bean (Table 2).

For the Andean group, which is composed of cultivars, breeding lines and landraces, a high number of polymorphic SNPs was identified for the landraces (68.3 % polymorphic SNPs, $n = 11$), while for the Mesoamerican group, composed of the same subgroups, the higher number

Table 3 Evaluation of SNP polymorphism among the biparental crosses between the Andean (A) and Mesoamerican (M) inter- (A × M) and intra- (M × M) gene pools

Parental combination		Type of cross	Total SNP (363)	
Parental 1	Parental 2		N. Poly	% Poly
AND 277	Rudá	A × M	223	61.43
Ouro Branco	CNFP10132	A × M	187	51.51
AND 277	SEA 5	A × M	213	58.68
BAT 477	Pérola	M × M	38	10.47
Pérola	Red Mexican	M × M	65	17.91
CNFP10132	BRS Cometa	M × M	44	12.12
BRS Aurora	PI181996	M × M	84	23.14
PI181996	US Pinto 111	M × M	64	17.63
Common polymorphism				
Inter-gene pool		3(A × M)	147	40.50
Intra-gene pool		5(M × M)	2	0.55
Inter-gene pool	Intra-gene pool	3(A × M) + 5(M × M)	0	0.00

(70.0 %) was identified with the line/cultivar germplasm ($n = 49$). For the two wild accessions evaluated, 25.6 % of SNPs were polymorphic. A slightly increased of H_e was observed for the Andean landraces ($n = 11$, 0.195), compared to the Mesoamerican line/cultivar ($n = 49$, 0.152) and wild accessions ($n = 2$, 0.170). The genetic differentiation (F_{ST}), which is directly related to the variance in allele frequency among populations, indicated an overall estimate of ≥ 0.525 ($p < 0.01$) between the Andean and Mesoamerican gene pool accessions. Among the subgroups of accessions grouped by gene pool (line, cultivar and landrace), higher differentiation was observed between the Mesoamerican landrace and Andean line/cultivar subgroups (0.704, $p < 0.01$), and lower differentiation was observed between the Mesoamerican line/cultivar and Mesoamerican landrace subgroups (0.009, $p > 0.05$) (Table 2).

Allelic polymorphism of SNP markers was tested in eight biparental crosses, five involving parental lines of the Mesoamerican gene pool and three inter-gene pool crosses (Table 3). On average, 59 (16.25 %) and 208 (57.30 %) of the 363 SNPs were polymorphic among the intra- and inter-gene pool crosses, respectively. Also considering these 363 SNPs, the proportion of common polymorphic SNP markers was 40.50 % (147) in the inter-gene pool accessions and only 0.55 % (two) in the intra-gene pool accessions.

Cluster analysis

Genetic relationships calculated between common bean accessions on the panel of 384 SNPs showed a large genetic dissimilarity between gene pools and relatively small genetic dissimilarity within each gene pool, as verified in the neighbor-joining tree build (Fig. 1). This tree was generated for the group of 95 genotypes (Fig. 1) and

clustered the genotypes into two major groups corresponding to the Andean and Mesoamerican gene pools. We also analyzed the formation of an intermediate subgroup composed of accessions classified as Mesoamerican (composed of cultivar/lines and landraces), indicating that during the development of such materials, it were done introgressions from Andean gene pool genotypes by artificial or natural crosses. The Mesoamerican wild accessions G23554 and PHAS8328 grouped together and were distinguished from the domesticated accession of the Mesoamerican gene pool. The genetic dissimilarity coefficient varied between the genotypes, with an overall observed value of 0.375, ranged from a maximum of 0.267 (Mesoamerican wild) to a minimum of 0.135 (Andean line/cultivar).

Population structure and AMOVA

As well as by cluster analysis, PCoA showed a clear structuring of the genotypes based on Mesoamerican and Andean origins, corroborating the high F_{ST} values calculated. The first axis, which explained the largest molecular variation of the data, consistently separated the accessions by gene pool (Fig. 2) with a value of 91.4 %. In the structure analysis, $K = 2$ was selected as the best partition with a subdivision based on Andean (green) and Mesoamerican (red) origin (Fig. 3). The Andean group ($n = 21$) presented a strong genetic structuration among 18 accessions and the remaining three were hybrids resulting from hybridization between Mesoamerican and Andean. At $K = 3$ the group of Mesoamerican was subdivided, represented by Mesoamerican genotypes with a prevalence of black commercial grain type (63.6 %) and Mesoamerican with a prevalence of *Carioca* commercial grain type (61.1 %). The group with predominance of black grain type was subdivided in two ($K = 4$), with no evident

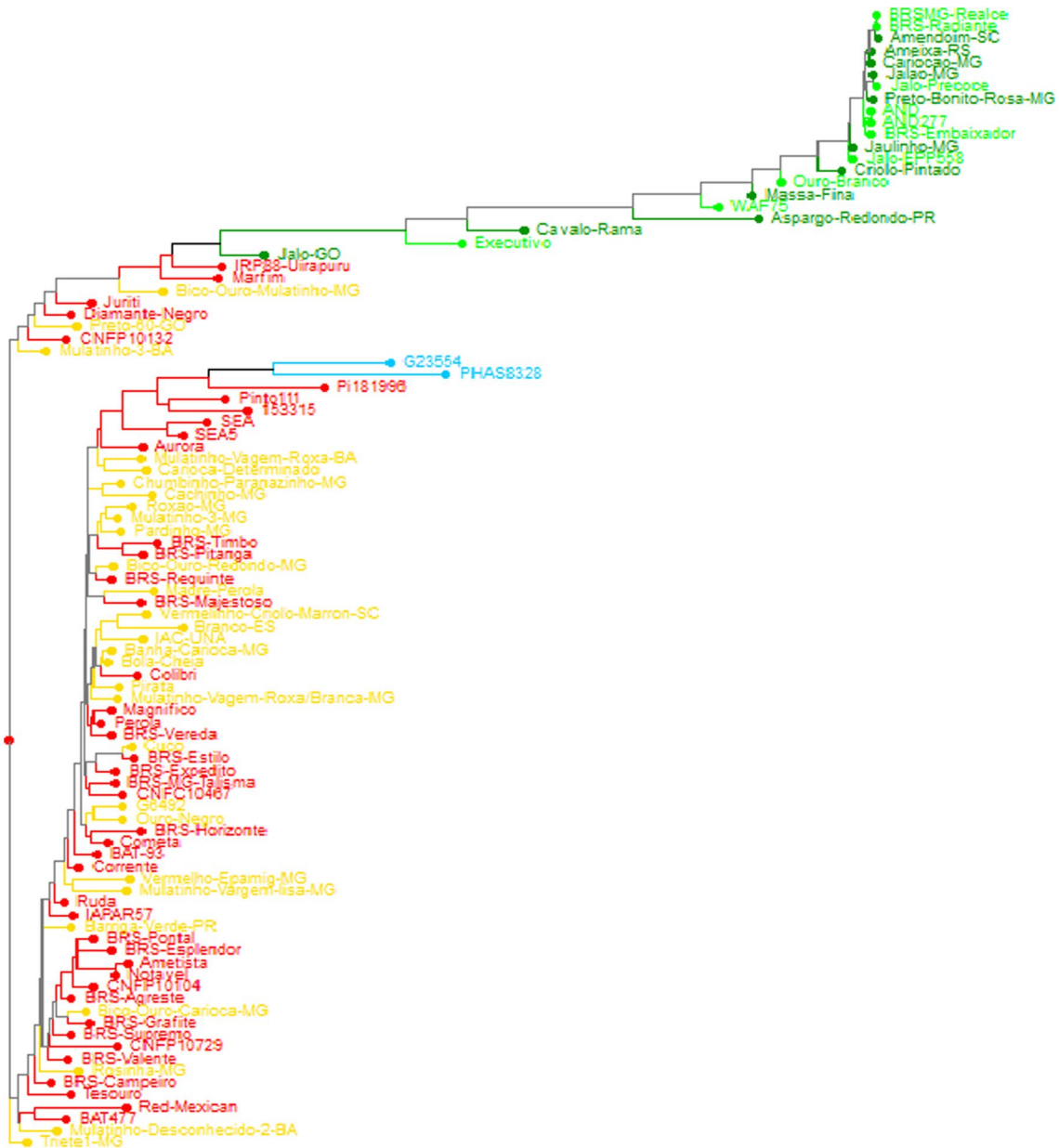


Fig. 1 Neighbor-joining trees representing common bean relationships based on 363 SNPs. The Andean genotypes are represented in *dark green* (landrace) and *light green* (cultivar/line); the Mesoameri-

can genotypes are shown in *red* (cultivar/line) and *yellow* (landrace); and the wild genotypes are indicated in *light blue*

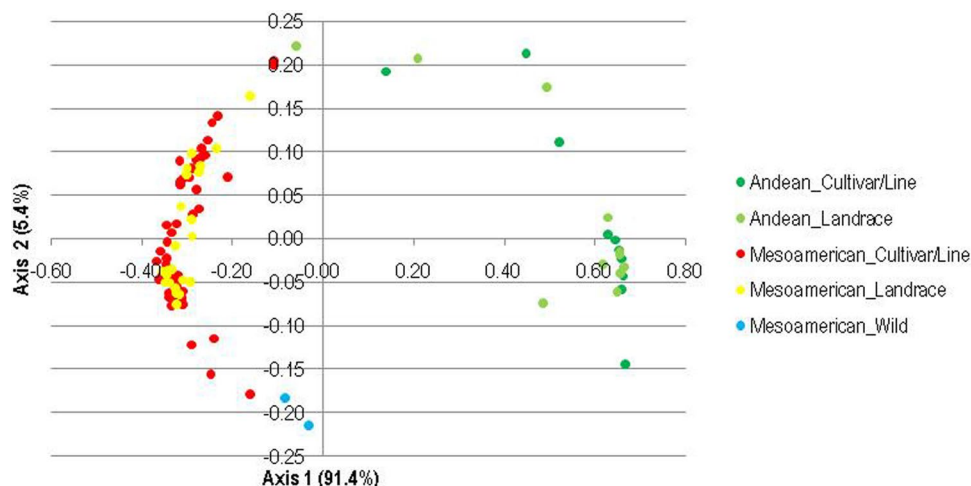
can explanation for this split, probably due to directed crosses of breeding programs. At $K = 5$, an additional fragmentation of the Mesoamerican group was observed and seven accessions were assigned as distinct from the remaining. From this seven, five accessions were classified as hybrids, resulting from hybridization among different germplasm not characterized at present study, and the remaining two accessions presented distinct genetic composition (US Pinto 111 and PHAS 8328). Through AMOVA, the RAD-SNPs explained, approximately, 80 % of the differentiation of the accessions

based on Andean and Mesoamerican origin, while 20 % was explained by populations. However, when the analysis was based on breeding (cultivar/line) and traditional (landraces) germplasm (2.84 %), greater differentiation (97.15 %) was observed between individuals (within populations).

Linkage disequilibrium and haplotype blocks

Of the 363 polymorphic SNP loci, four did not align to the reference genome of *P. vulgaris* (Andean v.1.0) through

Fig. 2 Principal coordinates analysis (PCoA) of the 95 common bean genotypes with 363 SNPs



Blast analysis. From the remaining 359 SNPs placed in the genome, only one, with $MAF < 0.05$, was removed. For the entire set of samples, of the 358 SNPs (total of 63,903 comparisons), 58.77 % were in LD (37,556 comparisons with $r^2 > 0.1$ and $p < 0.001$). This LD estimate decreased to 0.16 % (33 comparisons with $r^2 > 0.1$ and $p < 0.001$) in the Andean (21 genotypes) and 0.93 % (174 comparisons with $r^2 > 0.1$ and $p < 0.001$) in the Mesoamerican (74 genotypes) gene pools. The total number of excluded SNPs was 152 and 164 for the Andean and Mesoamerican genotype sets, respectively. Another comparative analysis between the RAD-SNP loci presented in this work and the SNPs previously published by Müller et al. (2015) is presented in Table 4.

Of the 359 RAD-SNP markers that were genotyped and aligned against the *P. vulgaris* genome, 290 with a call rate ≥ 0.75 and $MAF \geq 0.05$ were used for the haplotype analysis. A total of 21 haplotype blocks was identified in eight chromosomes, ranging from one (Chromosomes 7, 8 and 10) to six (Chromosome 1) and comprised 140 RAD-SNPs (40 %, average of 6.67 SNP/block). The largest block was identified on chromosome 2, with 18,735 kb and coverage of 14 SNPs, and the smallest block was on chromosome 1, with 13 kb and containing two SNPs. The total blocks represented 20 % of the entire genome. The frequency of haplotypes ranged from 0.011 to 0.810, with the most frequent haplotype located on chromosome 1. For the Andean accessions ($n = 21$), three blocks were formed: one located on chromosome 1 (with 12 different haplotypes, one of which was common to 70 % of the genotypes) and two blocks located on chromosome 9 (with four distinct haplotypes each, of which one haplotype was frequently in 80 % of the genotypes). For the Mesoamerican pool, no blocks were identified.

Considering all 704 SNPs (359 RAD-SNPs and 345 BARC-SNPs-PV loci) able of aligning the genome,

a selected set of 580 SNPs with a call rate ≥ 0.75 and $MAF \geq 0.05$ were selected and analyzed, presenting a LD of 70.21 % (117,883 comparisons with $r^2 > 0.1$ and $p < 0.001$). A total of 42 blocks were identified to span 11 chromosomes (Table 5), and the largest block was on chromosome 2 (9929 kb and 30 SNPs) and the smallest block was on chromosome 6 (5 kb and two SNPs). The haplotype frequencies ranged from 0.795 to 0.011, and the highest frequency belonged to block 1 of chromosome 1. *P. vulgaris* had haplotypes with an average block length of 8828 kb/block that covered 60 % of the sequence length. For the Andean group (Table 6), from a set of 325 SNPs (call rate ≥ 0.75 and $MAF \geq 0.05$), four blocks were identified with the largest on chromosome 1 (13,964 kb and 12 SNPs). For the Mesoamerican gene pool, from a set of 348 SNPs (call rate ≥ 0.75 and $MAF \geq 0.05$), six blocks were identified with the largest on chromosome 4 (11,011 kb and 11 SNPs) and the smallest on chromosome 6 (2180 kb and 4 SNPs).

SNP marker distribution in the common bean genome and annotation

Alignments of the 1377 SNPs against the genomes of *P. vulgaris*, including the 1032 SNPs made available by this work and the 345 SNPs previously reported by Müller et al. (2015), are represented in Fig. 4. A total of 1368 SNPs (1023 RAD-SNPs and 345 BARC-SNPs) mapped across the 11 common bean chromosomes and had high and low representation on chromosomes 1 (176 SNPs) and 4 (74 SNPs). From 1032 SNPs, 485 (47.00 %) were located in genes (Supplementary Material 3), and nine genes were represented by two SNPs. The GO enrichment analysis reveals that for the biological processes term, a large number of genes (70.31 %) encoded proteins associated with cellular and metabolic process, such as the specific terms “growth” (15 genes), “reproduction” (12 genes),

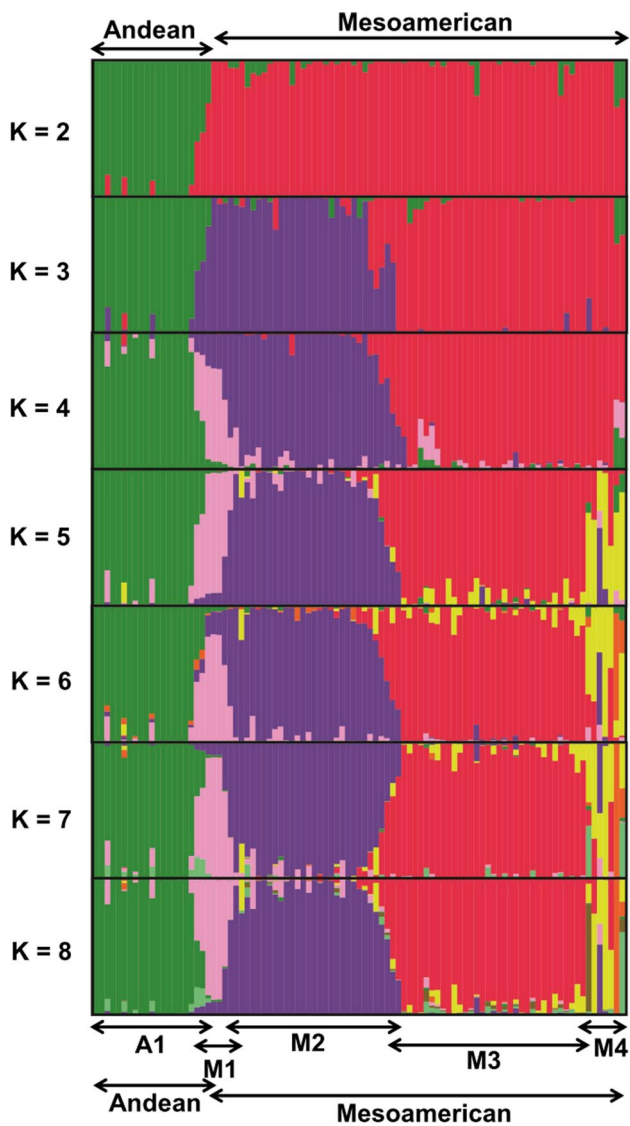


Fig. 3 Population structure inferred by the Bayesian approach based on RAD-SNP for $K = 2$ to 8. Each individual is represented by a vertical line that is divided into colored segments based on the proportion of the division identified for 2–8 subpopulations. The groups include: A1 (green): Andean; M1 (pink): cultivars *BRS Marfim* and *IPR88 Uirapuru* and landrace *Bico Ouro Mulatinho-MG* and one Landrace Andean (*Jalo-GO*); M2 (purple): Mesoamerican with a prevalence of 63.6 % of the cultivars with *Black* commercial grain type; M3 (red): Mesoamerican with a prevalence of 61.1 % of the cultivars with *Carioca* commercial grain type; M4 (different colors per genotype): line from Guatemala *PI 181996* (brown), Canadian line *153315* (yellow), North American lines *Red Mexican* (purple) and *US Pinto 111* (yellow), Mexican line *SEA* (yellow/red) and wild genotypes of Mexican origin *PHAS 8328* (orange) and *G23554* (light green)

“signaling” (28 genes) and “response to stimulus” (55 genes). The main distributions in the molecular function and cellular components were catalytic activity (38.35 %) and cell (28.66 %), respectively (Fig. 5).

Discussion

For common bean, RAD sequencing was very useful to discover genome-wide polymorphic SNP markers. The identification of a large number of RAD-SNPs (3357) and the amplification success rate of 96.61 % for the 384 tested SNPs demonstrate the potential use of this set of markers made available to the scientific community. The amplification success rate was comparable to the rate reported for common beans by Müller et al. (2015) (96.37 %) and higher than that described for soybean by Wu et al. (2010). The filter settings applied ensured a stringent selection of 3357 SNPs with high-probability amplification (≥ 3242 SNPs). However, further analyses are needed to validate these SNPs over the level of polymorphism. This study indicated that SNPs identified by the RAD strategy could offer an opportunity to genotype a population with different genetic structures. In addition, the use of variable restriction enzymes to cleave the DNA would improve polymorphism detection between genetically related genotypes.

The proportion of polymorphic RAD-SNP loci increased between gene pools (~ 7.0 %) and, overall, 363 RAD-SNP loci ($H_E = 0.396$) presented similar levels of genetic diversity for the same set of 88 genotypes of common bean when compared with the previously developed 345 SNPs ($H_E = 0.390$) described by Müller et al. (2015). Based on RAD-SNP markers, the number of polymorphic SNPs within the Andean group increased from 35.94 to 72.7 % (264 SNPs), and a slight decrease occurred in the Mesoamerican group (from 96.23 % to the actual proportion of 83.3 %). Consequently, the H_e value for the Andean group ($n = 21$; 0.163) was 2.2-fold higher compared to estimates previously obtained with BARC-SNP loci ($H_e = 0.072$) (Müller et al. 2015). Regarding the domesticated germplasm, the genetic diversity of the Andean accessions from Brazil were high ($n = 19$, $H_E = 0.161$) compared with those observed ($n = 60$, $H_E = 0.089$) by Rodriguez et al. (2015) including domesticated accessions from 12 countries. The differences most likely were consequence of the different sets of SNP markers used. The increased estimate of genetic diversity lies in the ease with which different samples belonging to the Andean group were pooled and sequenced ($n = 4$), allowing the detection of sequence variation among these genotypes and, consequently, captured a higher allele frequency distribution than previously obtained. We also detected higher genetic diversity ($p \leq 0.05$) for the cultivated germplasm from abroad ($n = 12$; $H_E = 0.438$) compared with the Brazilian ones ($n = 47$; $H_E = 0.316$), which is probably due to diverse nature of the international germplasm, representatives of six countries. Despite the higher recombination within Mesoamerican genepool, and the expectation of being detected

Table 4 Estimates of linkage disequilibrium calculated for pairs of loci ($r^2 > 0.1$ and $p < 0.001$) based on 345 BARC-SNPs, 363 RAD-SNPs and 580 BARC-SNPs and RAD-SNPs

	Pairs of linked loci			Total pairs of loci	Total loci (MAF > 0.05)
	Total (88) ^b	Andean (21) ^b	Mesoamerican (67) ^b		
SNPs-BARC ^a	50393 (84.92 %)	151 (0.25 %)	2751 (4.64 %)	59340	345
SNPs-RAD	36408 (56.66 %)	33 (0.16 %)	135 (0.74 %)	64261	358
SNPs-BARC ^a and SNPs-RAD	117883 (70.21 %)	34 (0.10 %)	207 (0.77 %)	167910	580 ^c

^a Müller et al. (2015)^b Sample size^c Loci with MAF > 0.05 and Call Rate > 0.75**Table 5** Overall common bean haplotype blocks based on 580 SNPs

Chr.	Total of blocks	Total SNPs/haplotype blocks	Average SNP/block	SNPs/haplotype blocks (%)	Blocks size (Kb)	Average block size (Kb)	Physical ^a length/ chrom (Kb)	Block genome coverage (%)
1	8	67	8.38	66.34	32895	4111.88	52183.5	63.03
2	3	44	14.67	75.86	38747	12915.67	49033.7	79.02
3	2	40	20.00	62.50	30481	15240.50	52218.6	58.37
4	2	24	12.00	77.42	22864	11432.00	45793.2	49.92
5	2	17	8.50	56.67	24281	12140.50	40237.5	60.34
6	5	26	5.20	56.52	12278	2455.60	31973.2	38.40
7	5	32	6.40	65.31	31427	6285.40	51698.4	60.78
8	3	24	8.00	52.17	28101	9367.00	59634.6	47.12
9	4	35	8.75	71.43	27557	6889.25	37399.6	73.68
10	3	36	12.00	70.59	29275	9758.33	43213.2	67.74
11	5	45	9.00	81.82	32595	6519.00	50203.6	64.92
Total	42	390	10.26	66.97	310501	8828.65	513589.1	60.30 ^b

^a Schmutz et al. (2014)^b Average genome block coverage**Table 6** Andean (AND) and Mesoamerican (MESO) haplotype block based on 580 SNPs covering the common bean genome

Chr.	Total of blocks		Total of SNPs/haplotype blocks		Average SNP/block		Captured SNP/block		Blocks size (Kb)		Average block size (Kb)	
	AND	MESO	AND	MESO	AND	MESO	AND	MESO	AND	MESO	AND	MESO
1	1	–	12	–	12	–	24.00	–	13964	–	13964	–
3	1	1	4	2	4	2	12.90	11.76	7246	1	7246	1
4	–	2	–	18	–	9	–	81.82	–	14907	–	7453.5
6	–	1	–	4	–	4	–	16.67	–	2180	–	2180
8	–	1	–	4	–	4	–	15.38	–	5822	–	5822
9	1	–	8	–	8	–	27.59	–	6376	–	6376	–
10	–	1	–	4	–	4	–	20.00	–	5617	–	5617
11	1	–	6	–	6	–	26.09	–	12774	–	12774	–
Total	4	6	30	32	7.5	4.60	22.64	29.12	40360	28527	10090	4214.7

high LDs due to the inbreeding nature of the *P. vulgaris*, the high level of LD in the whole sample was predominantly due to the gene pool structure. The slightly increased LD in

Mesoamerican compared to Andean (Table 4), and the high number of haplotype blocks with reduced size (Table 6), were probably favored by the process of selection imposed

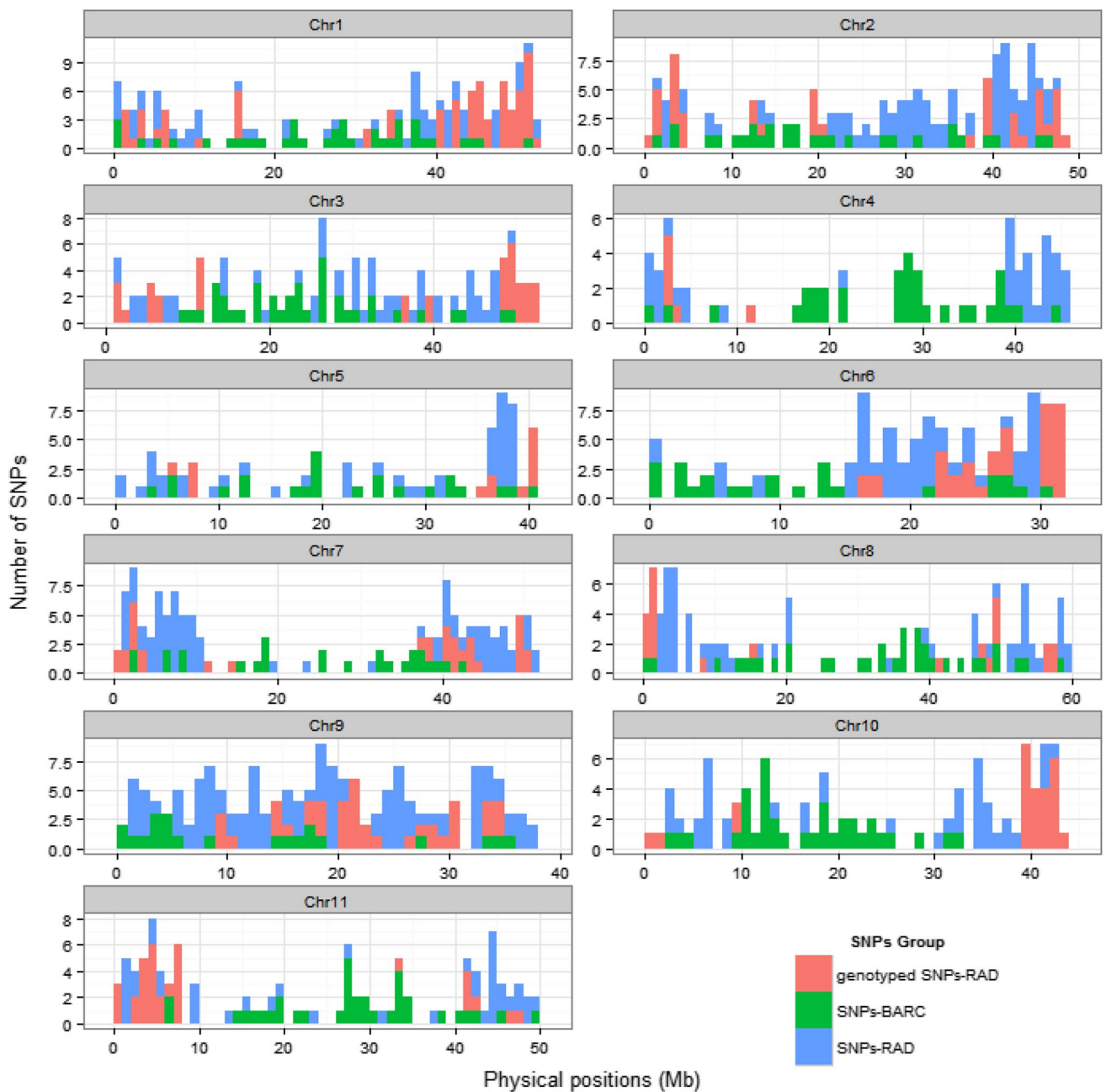


Fig. 4 SNP density physically mapped across eleven chromosomes of *P. vulgaris* L. and represented by the number of SNPs within 1 megabase (Mb) intervals. A total of 1377 SNPs were represented

throughout the genome. The number of genotyped RAD-SNPs, BARC-SNPs and combined SNPs were 359, 345 and 673, respectively

over the breeding programs. The use of crosses among elite parents, the most common type of cross made in a typical bean breeding program, could reduce the genetic gains with the selection, due to the high genetic similarity among genitors from the same gene pool. In Brazil, all the common bean grain consumed in the country is from the Mesoamerican gene pool. As expected, the genetic bottleneck following the successive breeding efforts in this gene pool have allowed the occurrence of increased linkage disequilibrium,

when compared to accessions from the Andean gene pool (Hudson 2001). It is expected that the number of haplotypes drastically increases with an increase in the number of genotyped SNPs.

Considering crosses involving common bean parents from distinct origins (Andean \times Mesoamerican), RAD-SNP loci were considerably less polymorphic (~57 %) when compared to the previously identified BARC-SNP loci (~78.0 %, Müller et al. 2015), while for the inter-gene

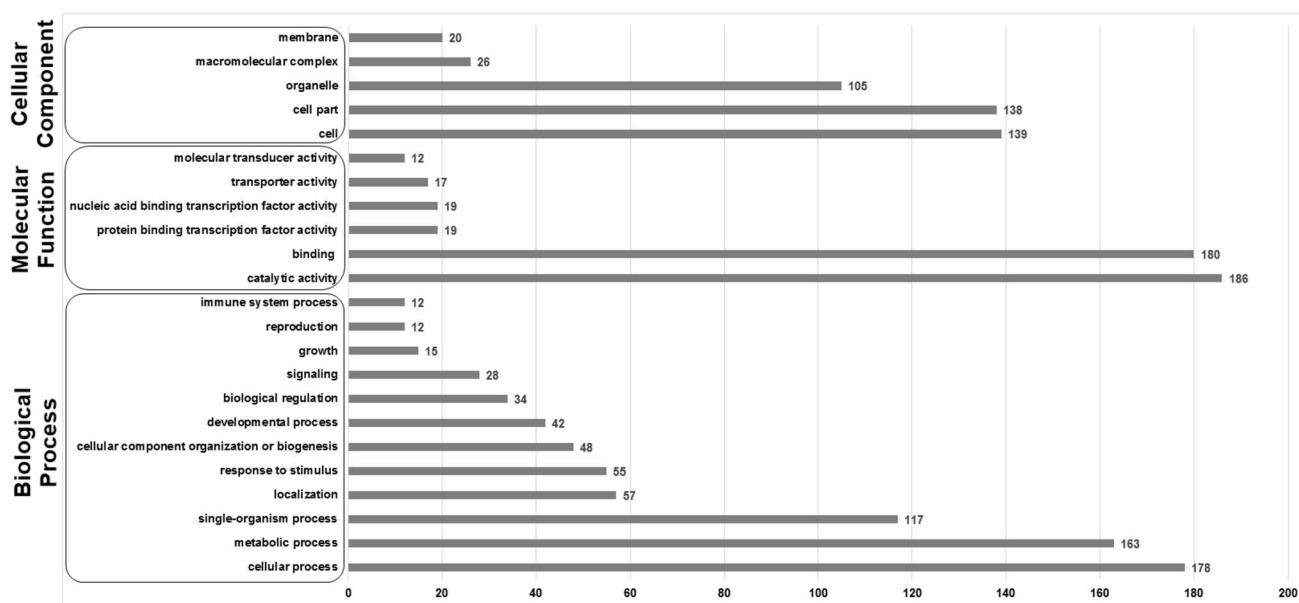


Fig. 5 Functional classification of 1032 RAD-SNP sequences derived from *P. vulgaris* annotated with 2nd level GO terms, including attributes based on cellular component, molecular function and biological process

pool, the rate of polymorphism was similar between these markers (~16.0 and ~17.0 %, respectively). Thereby, these set of markers complement each other for linkage analysis. An important common bean mapping population for drought tolerance (F_9 RIL population generated by SSD), derived from the cross between the Durango (SEA5; Singh et al. 2001) and Nueva Granada races (AND277; Blair et al. 2009) developed at the International Center for Tropical Agriculture (CIAT, Cali, Colombia), identified 495 high quality polymorphic SNPs to be used for linkage analysis. Focusing on bacterial wilt resistance genes, the biparental population Ouro Branco \times CNFP10132 ($F_{5.7}$ generation; Müller et al. 2015), of which the parental lines in this study were genotyped, also has great potential to be analyzed with a set of 442 polymorphic SNPs made available in the present study. Obviously, integrating genotyping with sequencing technology for QTL mapping of large recombinant inbred line (RIL) populations will provide quicker advances (Spindel et al. 2013). However, the immediate availability of molecular tool scan minimize efforts and accelerate the studies of QTL mapping.

The RAD-SNPs developed through two pools of DNA was very informative to capture the genetic diversity between and within gene pools, compared to the SNPs obtained from the sequence alignment of individual genotypes. Within the Andean group, the number of SNP loci with $MAF > 0.10$ was considerably higher for the RAD-SNP (total of 205 RAD-SNP loci vs. 55 BARC-SNPs), while for the Mesoamerican group, the total number of SNPs was similar (193 RAD-SNP loci vs. 141

BARC-SNPs). As these estimates were obtained for the same sample set, the efficient use of RAD sequencing to enable SNP development with adequate $MAF (\geq 1 \%)$ was demonstrated in this study. Gautier et al. (2013) showed that NGS of a pool of diploid individuals provides more accurate SNP allele frequency estimates than individual-based analysis and, in addition, provides a very cost-effective approach to estimate patterns of genetic diversity for a large set of populations. While 363 RAD-SNP markers tended to amplify the differences between the Andean and Mesoamerican gene pool ($F_{ST} = 0.656$), the dendrogram analysis demonstrated that landraces and line/cultivars within each gene pool share a high proportion of SNP loci. The differentiation was slightly reduced than previously estimated for the same set of samples ($F_{ST} = 0.759$, Müller et al. 2015). The strategy of pooled samples for SNP design showed to be more efficient to capture genetic variation in the Andean germplasm; however, the structuring by origin was still the predominant ($K = 2$). Even based on the screening of 12 individuals for the development of SNPs in the present work, a bias towards common polymorphisms and an overestimation of genetic differentiation was observed. Thus, according to the proposition of Goretti et al. (2014), more representative sample set should be used to develop the SNP markers, allowing more realistic estimates of genetic diversity and genetic differentiation, which will reduce the ascertainment bias.

As the set of genotypes used in this study is representative of the Brazilian germplasm, it could be supposed that due to the high extension of the LD (58.77 %) and small

number of haplotypes, a reduced number of SNP markers will be required to detect linkage between the marker and target genes related to the phenotypic variation. Certainly, due to the autogamous nature of the common bean and domestication process to which this crop was submitted (Bitocchi et al. 2013), this statement is true. Despite the significant LD (70 %) and haplotype block genome coverage (60 %) identified through analysis of 580 SNPs for the combined germplasm, the results suggest that dense SNP coverage will be needed to identify haplotypes that capture most of the genetic variation across inter-gene pool populations. SNP filtering by MAF (≥ 0.05) reduced the number of useful SNPs (~330) for LD and haplotype analysis by half and, consequently, limited block identification (four and six blocks for Andean and Mesoamerican, respectively). The identification of representative SNP tags spanning the entire genome in common beans has been a challenge and is dependent on the characterization of a larger number of SNPs; however, few SNPs will likely be required for whole genome sampling. This fact has implications for the design of genetic association studies. As the idea behind GWAS is to genotype a large number of markers distributed across the genome, the high LD observed within a gene pool could be efficiently explored in the breeding programs to increase the efficiency of selection. As more plants are genotyped and phenotyped, functional nucleotide polymorphisms, changes in specific genes causing desired phenotypes, will be identified for the development of functional markers. However, an additional effort to detect which gene included in the haplotype block is related to the evaluated trait could be performed by sequence similarity search against model plant databases or the development of tillering mutants (Porch et al. 2009).

In this study, we were interested in assessing the potential of RAD for direct genotyping-by-sequencing in the common bean. The developed markers added to the previously published SNPs (Müller et al. 2015) resulted in 708 high quality and genetically informative SNPs allowing a uniform physical distribution across the *P. vulgaris* genome. Both SNP panels have broad application in breeding and are recommended for use in genetic analyses of common bean germplasm. In addition, with the availability of several genotyping methods made available in recent years, the demand for SNP genotyping can be increasingly met. This work has allowed an accurate assessment of structuration and quantification of genetic diversity, and may be useful in whole-genome association studies. It is noteworthy that most of the germplasm used in the present study is from Brazilian origin (85 %) and, for this reason, the genetic diversity captured was suitable for answering questions related to such access. However, the public availability of these markers will enable a wide

variety of additional analyses to be carried out by scientists investigating and comparing the genetic variation across germplasm from other countries, as well as to explore the genetic basis of complex traits.

Acknowledgments We would like to thank the National Council for Scientific and Technological Development (CNPq) for the grants to BSFM, WJP, CB and RPV. The Brazilian Agricultural Research Corporation (EMBRAPA) for providing financial support for this research (Project Number 02.12.12.005.00.00).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389–3402
- Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376
- Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, Toppino L, Rotino G (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genom* 12:304
- Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265
- Bitocchi E, Bellucci E, Giardini A, Rau D, Rodriguez M, Biagetti E, Santilocchi R, Spagnoletti Zeuli P, Gioia T, Logozzo G, Attene G, Nanni L, Papa R (2013) Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. *New Phytol* 197(1):300–313
- Blair MW, Fregene MA, Beebe SE, Ceballos H (2007) Marker-assisted selection in common beans and cassava. In: Guimarães EP, Ruane J, Scherf BD, Sonnino A, Dargie JD (eds) Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish. FAO, Rome, pp 81–117
- Blair MW, Chavarro CM, Muñoz-Torres MC, Giraldo MC, Pedraza F (2009) Development and diversity of Andean-derived, gene-based microsatellites for common bean (*Phaseolus vulgaris* L.). *BMC Plant Biol* 9:100
- Blair MW, Cortés AJ, Penmetsa RV, Farmer A, Carrasquilla-Garcia N, Cook DR (2013) A high-throughput SNP marker system for parental polymorphism screening, and diversity analysis in common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 126:535–548
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–2635
- Cardoso PC, Veiga MM, de Menezes IP, Valdisser PA, Borba TC, Melo LC, Del Peloso MJ, Brondani C, Vianello RP (2013) Molecular characterization of high performance inbred lines of Brazilian common beans. *Genet Mol Res* 12(4):5467–5484
- Cardoso PC, Brondani C, Menezes IP, Valdisser PA, Borba TC, Del Peloso MJ, Vianello RP (2014) Discrimination of common bean

- cultivars using multiplexed microsatellite markers. *Genet Mol Res* 13(1):1964–1978
- Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistué L, Corey A, Filichkina T, Johnson EA, Hayes PM (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genom* 12:4
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676
- Cortés AJ, Chavarro MC, Blair MW (2011) SNP marker diversity in common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 123:827–845
- Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueno J, Windhausen VS, Buckler E et al (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3: Genes Genom Genet* 3:1903–1926
- Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* 4:359–361
- Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE* 6:e18561
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE : a simulation study. *Mol Ecol* 14:2611–2620
- Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10:564–567
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The Structure of Haplotype Blocks in the Human Genome. *Science* 296:2225–2229
- Gaitán-Solís E, Choi IY, Quigley C, Cregan P, Tohme J (2008) Single nucleotide polymorphisms in common bean: their discovery and genotyping using a multiplex detection system. *The Plant Genome* 1:125–134
- Galeano CH, Fernandez AC, Gomez 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 Genom* 10:629
- Galeano CH, Gomez M, Rodriguez LM, Blair MW (2009b) CEL I nuclease digestion for SNP discovery and marker development in common bean (*Phaseolus vulgaris* L.). *Crop Sci* 49:381–394
- Ganal MW, Altmann T, Röder MS (2009) SNP identification in crop plants. *Curr Opin Plant Biol* 12:211–217
- Gautier M, Foucaud J, Gharbi K, Cezard T, Galan M, Loiseau A, Thomson M, Pudlo P, Kerdelhue C, Estoup A (2013) Estimation of population allele frequencies at molecular markers produced by next generation sequencing methods: pooled versus individual DNA samples. *Mol Ecol* 22:3766–3779
- Gepts P (2014) The contribution of genetic and genomic approaches to plant domestication studies. *Curr Opin Plant Biol* 18:51–59
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9:e90346
- Goretti D, Bitocchi E, Bellucci E, Rodriguez M, Rau D, Gioia T, Attene G, McClean P, Nanni L, Papa R (2014) Development of single nucleotide polymorphisms in *Phaseolus vulgaris* and related *Phaseolus* spp. *Mol Breeding* 33(3):531–544
- Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross strategy and RAPD markers. *Genetics* 137:1121–1137
- Grattapaglia D, Silva-Junior OB, Kirst M, de Lima BM, Faria DA, Pappas GJ Jr (2011) High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across species. *BMC Plant Biol* 11:65
- Hudson RR (2001) Linkage disequilibrium and recombination. In: Bishop M, Cannings C, Balding DJ (eds) *Handbook of statistical genetics*. chichester. Wiley, USA, pp 309–324
- Hyten DL, Cannon SB, Song QJ, Weeks N, Fickus EW, Shoemaker RC, Specht JE, Farmer AD, May GD, Cregan PB (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genom* 11:38
- Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–1806
- Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genom* 15:740
- Kim S, Misra A (2007) SNP genotyping: technologies and biomedical applications. *Annu Rev Biomed Eng* 9:289–320
- Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24:133–141
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
- 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
- Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theor Appl Genet* 120:1525–1534
- Müller BSF, Pappas GJ, Valdisser PAMR, Coelho GRC, Menezes IPP, Abreu AG, Borba TCO, Sakamoto T, Brondani C, Barros EG, Vianello RP (2015) An operational SNP panel integrated to SSR marker for the assessment of genetic diversity and population structure of the common bean. *Plant Mol Bio Rep*. doi:10.1007/s11105-015-0866-x
- Nagasaki H, Ebana K, Shibaya T, Yonemaru J, Yano M (2010) Core single-nucleotide polymorphisms - a tool for genetic analysis of the Japanese rice population. *Breeding Sci* 60:648–655
- Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28:2537–2539
- Perrier X, Jacquemoud-Collet JP (2006) DARwin software. <http://darwin.cirad.fr/>. Accessed 21 January 2015
- Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sanchez-Villeda H, Sorrells M, Jannink J-L (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Gen* 5:103–113
- Porch TG, Blair MB, Lariguet P, Galeano C, Pankhurst CE, Broughton WJ (2009) Generation of a mutant population for tilling common bean genotype BAT 93. *J Am Soc Hort Sci* 134(3):348–355
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Ramírez M, Graham MA, Blanco-Lopez L, Silvente S, Medrano-Soto A, Blair MW, Hernandez G, Vance CP, Lara M (2005)

- Sequencing and analysis of common bean ESTs. Building a foundation for functional genomics. *Plant Physiol* 137:1211–1227
- Ritchie ME, Liu R, Carvalho BS; Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), Irizarry RA (2011) Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. *BMC Bioinformatics* 12:68
- Rodriguez M, Rau D, Bitocchi E, Bellucci E, Biagetti E, Carboni A, Gepts P, Nanni L, Papa R, Attene G (2015) Landscape genetics, adaptive diversity and population structure in *Phaseolus vulgaris*. *New Phytol*. doi:10.1111/nph.13713
- Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. *Mol Ecol Notes* 4:137–138
- Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Gefroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MM, Miklas PN, Osorno JM, Rodrigues J, Thareau V, Urrea CA, Wang M, Yu Y, Zhang M, Wing RA, Cregan PB, Rokhsar DS, Jackson SA (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet* 46:707–713
- Singh SP, Terán H, Gutierrez JA (2001) Registration of SEA5 and SEA13 drought tolerant dry bean germplasm. *Crop Sci* 41:276–277
- Souza TLPO, Barros EG, Bellato CM, Hwang EY, Cregan PB, Pastor-Corrales MA (2012) Single nucleotide polymorphism discovery in common bean. *Mol Breeding* 30:419–428
- Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S, Lorieux M, Ahmadi N, McCouch S (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor Appl Genet* 126:2699–2716
- Talukder ZI, Gong L, Hulke BS, Pegadaraju V, Song Q, Schultz Q, Qi L (2014) A High-Density SNP Map of Sunflower Derived from RAD-Sequencing Facilitating Fine-Mapping of the Rust Resistance Gene R12. *PLoS One* 9:e98628
- Thomson MJ, Zhao KY, Wright M, McNally KL, Rey J, Tung CW, Reynolds A, Scheffler B, Eizenga G, McClung A, Kim H, Ismail AM, de Ocampo M, Mojica C, Reveche MY, Dilla-Ermita CJ, Mauleon R, Leung H, Bustamante C, McCouch SR (2012) High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. *Mol Breeding* 29:875–886
- Vlasova A, Capella-Gutiérrez S, Rendón-Anaya M, Hernández-Oñate M, Minoche A, Erb I, Câmara F, Prieto-Barja P, Corvelo A, Sanseverino W, Westergaard G, Dohm JC, Pappas Jr GJ, Saburido-Alvarez S, Kedra D, Gonzalez I, Cozzuto L, Gómez-Garrido J, Aguilar-Morón MA, Andreu N, Aguilar OM, Garcia-Mas J, Zehnsdorf M, Vázquez MP, Delgado-Salinas A, Delaye L, Lowy E, Mentaberry A, Vianello-Brondani RP, García JL, Alioto T, Sánchez F, Himmelbauer H, Santalla M, Notredame C, Gabaldón T, Herrera-Estrella A, Guigó R (2016) The genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. *Genome Biol*. doi:10.1186/s13059-016-0883-6
- Willing EM, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics* 27:2187–2193
- Wu X, Ren C, Joshi T, Vuong T, Xu D, Nguyen HT (2010) SNP discovery by high-throughput sequencing in soybean. *BMC Genom* 11:469
- Zaiontz C (2013) Real Statistics using Excel. <http://www.real-statistics.com/reliability/cronbachs-alpha>. Accessed 12 November 2015
- Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C, Shen Y, Liu T, Li C, Li Q, Wu M, Wang M, Wu Y, Dong Y, Wan W, Wang X, Ding Z, Gao Y, Xiang H, Zhu B, Lee SH, Wang W, Tian Z (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* 33(4):408–414