Plant Mol Biol (2014) 86:455–470
DOI 10.1007/s11103-014-0240-7

Analysis of BAC-end sequences in common bean (*Phaseolus vulgaris* L.) towards the development and characterization of long motifs SSRs

Bárbara Salomão de Faria Müller · Tetsu Sakamoto · Ivandilson Pessoa Pinto de Menezes · Guilherme Souza Prado · Wellington Santos Martins · Claudio Brondani · Everaldo Gonçalves de Barros · Rosana Pereira Vianello

Received: 20 November 2013 / Accepted: 14 August 2014 / Published online: 28 August 2014

Abstract The increasing volume of genomic data on the *Phaseolus vulgaris* species have contributed to its importance as a model genetic species and positively affected the investigation of other legumes of scientific and economic value. To expand and gain a more in-depth knowledge of the common bean genome, the ends of a number of bacterial artificial chromosome (BAC) were sequenced, annotated and the presence of repetitive sequences was determined. In total, 52,270 BESs (BAC-end sequences), equivalent to 32 Mbp (~6 %) of the genome, were processed. In total, 3,789 BES-SSRs were identified, with a distribution of one SSR (simple sequence repeat) per 8.36 kbp and 2,000 were suitable for the development of SSRs, of which 194 were evaluated in low-resolution screening. From 40 BES-SSRs based on long motifs SSRs (≥trinucleotides) analyzed in high-resolution genotyping, 34 showed an equally good amplification for the Andean and for the Mesoamerican genepools, exhibiting an average gene diversity (H_E) of 0.490 and 5.59 alleles/locus, of which six classified as Class I showed a $H_E \geq 0.7$. The PCoA and structure analysis allowed to discriminate the gene pools ($K = 2$, $F_{ST} = 0.733$). From the 52,270 BESs, 2 % corresponded to transcription factors and 3 % to transposable elements. Putative functions for 24,321 BESs were identified and for 19,363 were assigned functional categories (gene ontology). This study identified highly polymorphic BES-SSRs containing tri- to hexanucleotides motifs and bringing together relevant genetic characteristics useful for breeding programs. Additionally, the BESs were incorporated into the international genome-sequencing project for the common bean.

Electronic supplementary material The online version of this article (doi:10.1007/s11103-014-0240-7) contains supplementary material, which is available to authorized users.

B. S. F. Müller · E. G. Barros
Laboratório de Genética Molecular de Plantas, Instituto de Biotecnologia Aplicada à Agropecuária (BIOAGRO), Universidade Federal de Viçosa (UFV), Viçosa, MG, Brazil

B. S. F. Müller · G. S. Prado · C. Brondani · R. P. Vianello (✉)
Laboratório de Biotecnologia, Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, Brazil
e-mail: rosana.vianello@embrapa.br

T. Sakamoto
Laboratório de Biodados, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brazil

I. P. P. Menezes
Laboratório de Genética e Biologia Molecular, Departamento de Biologia, Instituto Federal Goiano (IF Goiano), Urutaí, GO, Brazil

W. S. Martins
Instituto de Informática, Universidade Federal de Goiás (UFG), Goiânia, GO, Brazil

E. G. Barros
Departamento de Biologia Geral, Universidade Federal de Viçosa (UFV), Viçosa, MG, Brazil

E. G. Barros
Programa de Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília (UCB), Brasília, DF, Brazil

Keywords Leguminosae · Genomic analysis · BAC library · Transposable elements · Transcription factors · Microsatellites

Introduction

The Leguminosae family is one of the three major families of the Plantae Kingdom (Doyle and Luckow 2003), comprising approximately 18,000 species divided between the Mimosoideae, Caesalpinoideae, and Papilionoideae subfamilies. The Papilionoideae subfamily is the most numerous and includes species of high economic value, such as peanut (*Arachis hypogaea*), soybean (*Glycine max*), common bean (*Phaseolus vulgaris*), and cowpea (*Vigna unguiculata*) (Choi et al. 2004). *P. vulgaris* is the most important of the grain legumes consumed by humans in almost all tropical and subtropical climate countries. This species is of great importance in human nutrition because of its balanced chemical composition. The bean is used as a source of protein, which unlike most cereal proteins has an adequate amount of the essential amino acid lysine. The legume is also considered an important source of dietary fiber, exhibiting hypocholesterolemic and hypoglycemic properties, and it has a high content of complex carbohydrates, polyunsaturated fatty acids, minerals (Ca, Fe, Cu, Zn, P, K, and Mg), and B complex vitamins (Martin-Cabrejas et al. 1997). Health benefits associated with the consumption of beans have been demonstrated, indicating that their inclusion in the diet is associated with lower rates of chronic diseases, such as cardiovascular disease (mainly because of the effect on cholesterol), diabetes, and cancer (Mitchell et al. 2009; Zhu et al. 2012). The per capita consumption of this legume (kg/year) varies considerably between countries, with estimates ranging from 2 to 3.5 kg in the USA to, approximately, 16 kg in Brazil (FAOSTAT 2009). Genetic improvement programs for bean, which are essentially public and multidisciplinary undertakings, can now make use of biotechnological tools to make the process of developing new cultivars more efficient and competitive in global agribusiness.

There are several attributes that make the common bean a suitable experimental organism, such as the small genome size, which is comparable to that of the model legume species *Medicago truncatula* and *Lotus japonicus* (Cannon et al. 2009). Additionally, the bean exhibits a low index of genome duplication; most loci are single copy (Broughton et al. 2003). From a phylogenetic viewpoint, *P. vulgaris* occupies a privileged position within the group of tropical climate legumes (subtribe Phaseoleae), in addition to the demonstrated high degree of homology with the transcribed genomic regions of the soybean genome (Schmutz et al. 2010). Because of the growing importance of the common bean on the world sustainable agriculture scenarios, the genomes of two representative genotypes (from the Mesoamerican and Andean origin) were recently sequenced. The sequencing of the Mesoamerican variety (BAT93) was an integrated initiative undertaken by four Ibero-American countries, Brazil, Mexico, Spain, and Argentina (http://www.cyted.org/), whereas the Andean variety (G19833, Schmutz et al. 2014) was sequenced in the US (http://www.phytozome.net/commonbean.php). Free access to the sequences and data will contribute to increase even more the use of the common bean as a genetic model species and will have a positive impact on the investigation of soybean and other legumes. Currently, the analysis of the sequences is in progress; however, the large collection of genome information has already generated many research opportunities, such as establishing mining approaches for repetitive sequences (e.g., simple sequence repeats—SSRs) (Blair et al. 2012a) and single nucleotide polymorphisms (SNPs) (Cortés et al. 2011; Galeano et al. 2012; Blair et al. 2013).

Phaseolus vulgaris has been investigated from a genomic perspective over the past 10 years. The emphasis of the studies was on the development of genetic maps with expanded genome coverage (Galeano et al. 2011), the location of quantitative trait loci (QTL) (Asfaw et al. 2012), the generation of expressed sequence tags databases (ESTs) with an emphasis on biotic and abiotic stresses (Ramírez et al. 2005; Kalavacharla et al. 2011), and the construction of BAC (bacterial artificial chromosome) clone libraries to aid the manipulation and assembly of the structural genome (Kami et al. 2006; Schlueter et al. 2008). Libraries containing large genomic DNA inserts are considered extremely useful tools for establishing cytogenetic and physical mapping and are of great importance for genomic sequencing, positional cloning, characterization of the genome for gene structure, functional analysis and recently, transgenic approaches (Schlueter et al. 2008; Ragupathy et al. 2011; Kang and Hérbert 2012). BACs have been widely used as vectors for cloning of large genomic fragments (100–150 kb), which is demonstrated for several economically important species in a review by Yu (2012). Mining approaches for simple sequence repeats (SSRs) or microsatellites (Tautz 1989) using databases derived from sequencing the ends of BACs (BES-SSRs) have been successfully applied to various plant species, including legumes such as soybean (Shultz et al. 2007), common bean (Córdoba et al. 2010), pigeon pea (Bohra et al. 2011), and cultivated peanut (Wang et al. 2012).

A growing number of SSR markers (currently ≥2,000) derived from genomic sequences and genes are available for the common bean (Yu et al. 2000; Blair et al. 2003; Hanai et al. 2007; Garcia et al. 2011) making it possible to perform studies using different scientific approaches. The

main applications include studies to estimate genetic diversity and population structure in the cultivated (Cardoso et al. 2014), landraces (Burle et al. 2010) and wild germplasm (Kwak and Gepts 2009; Blair et al. 2012b) as well as genetic mapping experiments (Galeano et al. 2011) with an emphasis on the identification of QTLs (Hanai et al. 2010; Blair et al. 2011a). Almost all of the SSRs described in the literature for the common bean are derived from dinucleotide repeats, followed to a lesser extent by trinucleotide derivatives of gene sequences. With recent efforts in the genomic sequencing of several species of plants SSRs based on tetra-, penta- and hexanucleotides have been reported (Sonah et al. 2011) and showing to be less abundant in the genome. In *Medicago sativa*, the estimated joint proportion of tetra-, penta-, and hexanucleotides in a 12,371 ESTs was approximately 25.1 %, followed by 48.8 % trinucleotides, and 26.1 % dinucleotides (Wang et al. 2013). In *P. vulgaris*, the few SSRs that have been evaluated based on repetitions ≥tetra-nucleotides (only 12 markers) were confirmed as monomorphic (Blair et al. 2011b). However, the large volume of genome information currently available created an opportunity for the development of a practically unlimited number of SSRs, expanding the possibilities for the larger-scale characterization of new classes of repeats with analytical robustness and a potential for the detection of polymorphisms.

In this study, the ends from a genomic BAC library derived from the *P. vulgaris* genotype BAT93 (Kami et al. 2006) were sequenced, and the sequences informations were characterized with respect to nucleotide composition, presence of transcription factors and transposable elements, gene annotation, genomic similarities with different legume species, and presence of sequences with long repetitive motifs (≥tri- and ≤hexanucleotides, as well as compound SSRs) aiming to develop useful SSRs for the genetic improvement of the species.

Materials and methods

Plant material

The genotype BAT93, cloned in a BAC library, used for sequencing and SSR discovery is a Mesoamerican breeding line of *P. vulgaris* developed at the International Center for Tropical Agriculture (CIAT, Cali, Colombia). A total of 88 common bean (*P. vulgaris*) genotypes used for SSR characterization comprise 67 Mesoamerican and 21 Andean genotypes were characterized (Supplementary Table 1). From the Mesoamerican, 43 were cultivars/lines from Brazil, Colombia, Guatemala, Germany and USA; 22 were all landraces from Brazil and the two wild were represented accessions from Mexico. Among the Andean, 10 were advanced cultivars/lines from Brazil and Colombia and the 11 remaining were landraces from Brazil. Germplasm accessions were representatives of diverse classes of grains, including Black, *Carioca*, Cranberry, Dark Red Kidney, *Jalo*, *Mulatinho*, Pinto, Purple, *Rajado*, Red, *Rosinha*, White, among others. The genomic DNA was extracted from leaf tissue samples in accordance with the extraction procedure based on the CTAB method described by Grattapaglia and Sederoff (1994).

Sequencing and processing of BAC-end sequences (BESs)

The complete genome of *P. vulgaris* genotype BAT93, cloned and maintained as long fragments in a genomic library developed by Kami et al. (2006), was used in this study. The average size of the inserts in the PVMBBa library (also known as BAT 93 *Hin*dIII library) was 125 kbp, with a 20X coverage of the genome. In total, 30,000 BAC clones were randomly selected for individual end-sequencing in the forward and reverse directions using the 3730xl ABI (Applied Biosystems, CA, USA) automated sequencing platforms at the Arizona Genome Institute (AGI, AZ, USA) in accordance with the method described by Kim et al. (2007). The BESs were processed using the Phred program (Ewing et al. 1998), and nucleotides corresponding to vector sequences were excluded using the LUCY program (Chou and Holmes 2001). The CAP3 program (Huang and Madan 1999) was used to determine consensus (contigs) sequences.

Identification of BES-SSRs and design of primers

The BESs were processed according to the methodology described by Martins et al. (2006) using the TROLL module (Castelo et al. 2002) with the Staden package (Staden et al. 1999). Specific primers flanking the identified SSRs were designed using the Primer3 software (Rozen and Skaletsky 2000). The TROLL module parameters were set to identify sequences containing SSRs of perfect and compound types based on di-, tri-, tetra-, penta- and hexanucleotides, with at least four repeats. The significance of differences among the motif sizes were evaluated using the Fisher's exact test computed in the software R (available at CRAN, http://cran.r-project.org/) at P value ≤0.05. The parameters for the design of the primers were similar to those described by Garcia et al. (2011). The nomenclature "BES" was added to the nomenclature "SSRs" to indicate its origin (BES-SSRs).

Genetic analysis of BES-SSRs

To develop functional SSR markers, 194 SSR markers were evaluated with priority given to tetra-, penta- and hexa-repeats

and sequences with the highest number of repeats. These markers were evaluated for amplification pattern and possible polymorphism against a panel of four common bean genotypes, one representative of the Andean gene pool (BRS Executivo) and three of the Mesoamerican gene pool (Rudá, BRS Esplendor, and BRS Agreste). The markers were amplified, subjected to electrophoresis on a low-resolution 2 % agarose gel stained with SYBR® (Applied Biosystems, USA).

Based on the ease of interpretation of the electrophoresis profile, robustness of the amplified product, and the potential for polymorphism detection, 40 BES-SSRs were selected for the synthesis of fluorescently-labeled primers (Supplementary Table 2) to be genotyped on a high-resolution capillary electrophoresis. Accounting for the expected size of the amplified fragment (in bp) and the complementarity between primers as determined by the AutoDimer program (Vallone and Butler 2004), the SSRs were grouped in multilocus genotyping systems (termed multiplex systems) using four distinct fluorescent dyes, 6-FAM™, HEX™, NED™, and PET™ (Applied Biosystems, USA).

For the panel of 40 selected BES-SSRs, the evaluation was performed on a set of 88 common bean genotypes (Supplementary Table 1). The genomic DNA was extracted from leaf tissue samples in accordance with the extraction procedure based on the CTAB method described by Grattapaglia and Sederoff (1994). The amplification reactions of the microsatellite DNA, set in tetraplex systems, were performed using the commercial 2× QIAGEN Multiplex PCR kit (Qiagen, NRW, Germany) in a GeneAmp 9700 Thermal Cycler (Applied Biosystems, USA) with an initial denaturation step at 95 °C for 15 min, followed by 40 cycles consisting of a denaturation step (94 °C for 30 s), an annealing step (ranging from 56 to 60 °C for 90 s), and an extension step (72 °C for 1.5 min), followed by a final extension step at 72 °C for 10 min. The electrophoresis was performed on the ABI3500 platform (Applied Biosystems, USA). The analysis of the fragments was performed using GeneScan Analysis 2.1 (Applied Biosystems, USA), and the size determination of alleles was performed using the GeneMapper 4.1 program (Applied Biosystems, USA).

Genetic diversity and structure analysis

The genetic parameters were analyzed using the *GenAlEx* v6.5 program (Peakall and Smouse 2012) by estimating the allele frequency, number of alleles per locus ($\bar{A}$), expected heterozygosity (H_E), observed heterozygosity (H_O), probability of identity (PI), private alleles (*Ap*) and Wright's fixation index (F_{IS}) based on a 95 % statistical confidence interval using 10,000 bootstrap replicates. Standard errors (SE) over loci were calculated between genepools and among the groups of accessions. All statistical analyses (overall significances) were performed using the non-parametric Wilcoxon test implemented by the Real Statistics program (Zaiontz 2013). The Principal Coordinates Analysis (PCoA) based on the genetic distances estimated by the Rogers' coefficient modified by Wright and the genetic differentiation (F_{ST}) were calculated (95 % statistical confidence interval using 10,000 bootstrap replicates) using the *GenAlEx* v6.5 program (Peakall and Smouse 2012). Population structure based on the Bayesian clustering method was determined by STRUCTURE v2.2.4 program (Pritchard et al. 2000) to infer the number of populations (k) without prior information on the number of groups of accessions. The membership of each genotype was run for the range of genetic cluster (*K*) from 1 to 20, with 30 interactions each. The runs were implemented with a burn-in period of 500,000 steps followed by 1,000,000 Monte Carlo Markov Chain (MCMC) replicate. The most probable *K* was determined as proposed by Evanno et al. (2005) using the Structure Harvester v0.6.93 (Earl and vonHoldt 2012). The software CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007) was used to find consensus among the 10 most probable *K* interactions, and the results were displayed by DISTRUCT v1.1 (Rosenberg 2004).

Functional annotation of BESs

The BLASTN (Altschul et al. 1990) algorithm was used to annotate the BESs against the GenBank GSS (genomic sequences) database with a minimum reliability value (E value) ≤1E−06, whereas the BLASTX (Altschul et al. 1997) algorithm was used against the non-redundant protein database at the default settings. BLAST2GO was used for the functional analysis of the sequences using the terms of the gene ontology (GO) from the BLASTX hits (Conesa et al. 2005) at default settings (E value of 1E−06 and annotation cutoff of 55). The BESs were aligned against the transcriptome database of common bean (BLASTN; E value ≤1E−06), Andean variety (G19833) available at (http://www.phytozome.com/commonbean.php). Using BLASTN, the BESs were also compared against the Andean *P. vulgaris* (G19833) genome database with an E value ≤1E−05. The IDs generated were subjected to automatic comparison with the transcription factors (TFs) database of plants (PlnTFDB, Plant Transcription Factor (http://plntfdb.bio.uni-potsdam.de/v3.0/), Pérez-Rodríguez et al. 2010), selecting only the TFs described for the common bean, and with the soybean transposable elements (TEs) database (SoyTEdb, SoyBase and the Soybean Breeder's Toolbox (http://soybase.org/soytedb/), Du et al. 2010) at a cutoff E value ≤1E−05.

Comparative genomic analysis

The total BESs, as well as the BES-SSRs TFs, and TEs were aligned separately with the genome of *M. truncatula*

(Young et al. 2011), *G. max* (Schmutz et al. 2010) and *P. vulgaris* Andean variety (G19833, Schmutz et al. 2014) (http://www.phytozome.com/commonbean.php). The alignments were performed using the BLASTN algorithm at an E value ≤1E−05 and a minimum alignment length of 50 bp. The best hits determined from each BES were used to develop maps for synteny visualization using the Circos program (Krzywinski et al. 2009).

Results

Sequencing of BES

The sequencing of 30,000 BAC clones was sufficient to obtain an overview of the common bean genome and to derive useful information that can potentially increase the knowledge and improve the genetics of this species. In total, 60,000 genomic sequences from BAC clones were sequenced, comprising approximately 36 Mbp and corresponding to an estimated 6.5 % of the *P. vulgaris* genome. The sequencing of the BAC ends generated 52,270 BES reads of high quality (87.12 %) with an average size of 690 bp. The total length analyzed was 31,702,945 bp with an estimated GC content of 38.83 %.

Identification and characterization of BES-SSRs

A preliminary search for SSRs in the 52,270 BESs resulted in the identification of 3,789 SSRs, with a frequency of one SSR per 8.36 kb. For a more stringent analysis to identify complete SSR regions useful for the development of molecular markers, short and redundant sequences were removed from the BESs, resulting in 47,698 useful sequences, which were pooled into 1,028 contigs and 42,082 singlets (total 43,110). From the contigs and singlets, 2,472 SSRs were identified, with some of BES containing more than one SSR. Considering only 1 SSR per BES, 2,000 SSRs were identified and made available to the scientific community by NCBI (BES-SSRs) (Supplementary Table 3).

Among the common bean BES-SSRs, the perfect type trinucleotide (45.75 %) and dinucleotide (27.95 %) repeats were the most frequent, followed by the compound repeats (19.4 %), which are formed by the contiguous or adjacent repetition of different motifs. The SSRs with tetra- (3.85 %), penta- (0.95 %) and hexanucleotide repeats (2.1 %) were less abundant (Table 1). Trinucleotides with AAG motifs occurred more frequently (17.12 %), followed by AAT (9.83 %) and ATG (6.04 %). Regarding the 559 dinucleotides, the AT (16.57 %), AG (7.09 %) and AC (4.29 %) motifs were the most frequent, and the GC motif was represented by only 1 sequence. The most common motifs for the tetra-, penta-, and hexanucleotides were AAAT (2.4 %), AAAAT (0.8 %), and GGGCTT (0.65 %), respectively (Table 2).

Table 1 Characterization of identified BES-SSRs including the total number of markers classified by the types of repeat motifs, frequency estimation of the motifs in relation to the total number of identified SSRs, and classification based on the lengths (in base pairs) of the repeats in Class I (SSRs ≥ 20 bp) and Class II (SSRs ≤ 19 bp)

Repeat motif	Number of SSRs	Frequency (%)	Class I	Class II
Dinucleotide	559	27.95	167	392*
Trinucleotide	915	45.75	37	878*
Tetranucleotide	77	3.85	17	60*
Pentanucleotide	19	0.95	19*	0
Hexanucleotide	42	2.10	42*	0
Compound	388	19.40	362*	26
Total SSRs	2000	100	644	1,356*

* Significant difference (95 % IC, $P < 0.05$)

Based on the length of the DNA sequences containing repetitions, the SSRs were categorized into Class I (SSRs ≥ 20 bp) and Class II (SSRs ≤ 19 bp), as proposed by Temnykh et al. (2001). Of the 2,000 BES-SSRs, 644 (32.2 %) were representative of Class I and were predominantly compound repeats (56 %), whereas 1,356 (67.8 %) were Class II with a predominance of trinucleotides (65 %) followed by dinucleotides (29 %). Penta- and hexanucleotide repeats were identified only as representatives of Class I (Table 1).

Genetic diversity and structure analysis of BES-SSRs

Longer SSRs tend to be more variable, and together with the tetra-, penta-, and hexanucleotide repeats they are prioritized for use as molecular markers. Primers were designed for the 194 SSRs derived from the 2,000 BES-SSRs. From them, 164 SSRs have perfect repeats (18 di-, 8 tri-, 77 tetra-, 19 penta-, and 42 hexanucleotides) and 30 were compound SSRs. Of these, 31 % were representatives of Class II (60 tetra-nucleotide repeats) and 69 % were Class I, which included all the remaining SSR motifs identified in this study. In the first stage of the screening, 64 % of loci products were successfully amplified in at least one sample tested, of which the highest rates of successful amplification were for pentanucleotide (79 %) and tetranucleotide (78 %) repeats, whereas the rate for hexanucleotides was only 48 %. Among the SSRs SSRs that were amplified, 24 % were polymorphic. Of the total markers evaluated, 40 BES-SSRs were selected for fluorescence synthesis.

From 40 BES-SSRs, 33 showed perfect repeats (4 tri-, 10 tetra-, 10 penta-, and 9 hexanucleotides) and seven compound repeats. In order to evaluate the BES-SSRs amplification in multiplexes, 10 tetraplex systems were tested, and in eight multiplexes, all SSR markers amplified. The

Table 2 Frequency of each BES-SSR motif type most representative in the common bean genome categorized according to the number of repeats that make up the motifs

Repeat motif	Number of SSRs	Frequency (%)
Dinucleotide		
AT|TA	332	16.57
AG|GA|CT|TC	141	7.09
AC|CA|TG|GT	85	4.29
GC|CG	1	0.05
Trinucleotide		
AAG|AGA|GAA|CTT|TTC|TCT	343	17.12
AAT|ATA|TAA|ATT|TTA|TAT	197	9.83
ATG|TGA|GAT|CAT|ATC|TCA	121	6.04
AAC|ACA|CAA|GTT|TTG|TGT	95	4.74
AGG|GGA|GAG|CCT|CTC|TCC	58	2.89
AGC|GCA|CAG|GCT|CTG|TGC	55	2.74
ACC|CCA|CAC|GGT|GTG|TGG	35	1.75
GGC|GCG|CGG|GCC|CCG|CGC	5	0.25
Tetranucleotide		
AAAT|AATA|ATAA|TAAA|TTTA|TTAT|TATT|ATTT	48	2.40
Pentanucleotide		
AAAAT|AAATA|AATAA|ATAAA|TAAAA|TTTTA| TTTAT|T TATT|TATTT|ATTTT	16	0.80
Hexanucleotide		
GGCTTG|GCTTGG|CTTGGG|TTGGGC|TGGGCT| GGG CTT|CCGAAC|CGAACC|GAACCC|AACCCG| ACCCGA|CCCGAA	13	0.65

five markers that did not amplify in multiplex were individually analyzed (PvPenta 5, PvComp8, PvPenta8 and PvPenta10). From the analysis of 88 genotypes, 34 (85 %) polymorphic and six monomorphic (PvTetra7, PvPenta18, PvTri4, PvPenta9, PvHexa26, and PvHexa28) BES-SSRs were identified showing an equally good amplification for the Andean and for the Mesoamerican genepools. Considering the polymorphism of 34 BES-SSRs, the average number of alleles per locus was 5.59, ranging from two (PvHexa12, PvTetra32, PvPenta8, PvPenta14, and PvTetra76) to 19 (PvComp10). The average H_E was 0.490, ranging from the minimum of 0.135 in the PvPenta5 locus to the maximum of 0.930 in the PvComp10 locus. The *Ho* was 0.004, corroborating with the high level of inbreeding (F_{IS} = 99.3) estimated for the common bean genotypes, so the correction for null alleles was not considered. The number of alleles per locus ($\bar{A}$) was 4.76 for Mesoamerican genotypes and 2.73 for Andean, at $P = 0.007$, while the H_E was 0.396 and 0.286, at $P = 0.027$ (as obtained from a nonparametric test considering 67 and 21 accessions, respectively). The average number of alleles per marker locus ($\bar{A}$) was significantly different among cultivars/lines (4.56), landraces (3.44) and wild (1.55), at $P < 0.05$ (nonparametric test with 53, 33 and 2 accessions, respectively), whereas the average estimates *of* H_E were not significantly different ($P > 0.05$) with values of 0.480, 0.455 and 0.338, respectively. In relation to the private alleles (*Ap*) (Table 3) in the BES-SSRs, 55 alleles where identified in cultivars/line, 20 in landraces and 12 in wild accessions. In relation to the classes of motifs, trinucleotides (represented by three markers) showed the highest average estimate of H_E (0.733), followed by the compound SSRs (0.659) (represented by seven markers). The H_E for the tetranucleotides was 0.512, whereas the values for the penta- and hexanucleotides were lower but similar, with indices of 0.351 and 0.332, respectively (Table 4). Of the 34 markers, six (with a repeat length > 40 bp) showed a $H_E \geq 0.7$; in descending order, the H_E estimates were 0.918 (PvComp10), 0.821 (PvTri8), 0.816 (PvComp4), 0.743 (PvTri6), 0.741 (PvTetra25), and 0.703 (PvComp9). For $H_E \geq 0.5$, the number increased to 16 loci, which corresponded to 47 % of the SSRs characterized. A high standard of amplification profiles was observed with rare occurrences of nonspecific products and stuttering (Supplementary Figure 1).

PCoA revealed two distinct cluster from the entire population which is related to their germplasm Mesoamerican and Andean origin (Fig. 1). The first axis explained 29 % of the molecular variation and separated the accessions by origin, while the second explained 10 %. The two wild accessions were placed in an intermediate position between the Andean and Mesoamerica germplasm. A high estimate of F_{ST} (0,733, $P < 0.05$) between the Andean and

Table 3 Descriptive statistics of the 34 BES-SSRs characterized in 88 genotypes of common bean

Sample size (*N*), number of alleles (*A*), number of alleles per locus ($\bar{A}$), number of private alleles (*Ap*), genetic diversity (H_E), observed heterozygosity (H_O), endogamy index (F_{IS}) and standard error (*SE*)

	N	*A*	$\bar{A}$ (*SE*)	*Ap*	H_E (*SE*)	H_O (*SE*)	F_{IS} (*SE*)
Gene pool							
Andean	21	93	2.73 (0.32)	28	0.286 (0.042)	0.001 (0.001)	0.995 (0.004)
Mesoamerican	67	162	4.76 (0.54)	97	0.396 (0.036)	0.005 (0.002)	0.990 (0.003)
Group							
Cultivar/line	53	155	4.56 (0.53)	55	0.480 (0.036)	0.002 (0.001)	0.966 (0.002)
Landrace	33	117	3.44 (0.47)	20	0.455 (0.043)	0.007 (0.003)	0.989 (0.004)
Wild	2	51	1.55 (0.11)	12	0.338 (0.059)	0.015 (0.015)	0.953 (0.033)
Total	88	190	5.59 (0.72)	–	0.490 (0.036)	0.004 (0.001)	0.993 (0.002)

Mesoamerican groups was observed for the 34 BES-SSRs analyzed indicating a genetic differentiation by origin. The structure analysis, through the results of ΔK, indicates that K = 2 was the most probable cluster for BES-SSRs, corresponding to the Mesoamerican and Andean genepools (Supplementary Figure 2). A total of 66 genotypes, constitute of 42 cultivars/lines and 22 landraces, plus the two wild genotypes were grouped as the Mesoamerican origin, while the Andean cluster constitute of 11 cultivars/lines and 11 landraces. The accession BRSMG Talismã was previously identified as of Mesoamerican origin and, a posteriori, placed in the Andean cluster.

Annotation of BES

The annotation of 52,270 sequences (BLASTN against the GenBank GSS database) produced 47,194 alignments (90.3 %) of which 31,564 (60.4 %) aligned with a high similarity (E value = 0) to the genomic sequences of *P. vulgaris* available in NCBI. Of the 2,000 BES-SSRs, 1,611 (80.5 %) demonstrated high similarity with sequences from GSS database. A total of 14.165 BESs (27.1 %) showed similarity with the common beans transcriptome sequence data, of which 11.200 (79 %) were annotated representing 2.032 different transcripts. Based on BLASTX, putative functions were identified for 24,321 (46.53 %) BESs, of which 609 (30.45 %) were BES-SSRs, and 13 (32.5 %) belonged to the 40 BES-SSRs genotyped in 88 accessions (Supplementary Table 1). The reliability estimates (E values) of the sequences submitted for the mapping and annotation process using BLAST2GO ranged from 1E−06 to 1E−175. A higher similarity (BLAST "Top-Hits") to sequences from *P. vulgaris* (17 %), followed by *G. max* (13.5 %), *Vitis vinifera* (6 %), *M. truncatula* (4 %), and *Oryza sativa* (1 %), was observed. Considering only the BES-SSRs, a high proportion of alignment with the *G. max* (12 %) genome was observed, whereas the alignment estimate was only 8.65 % for *P. vulgaris*.

Of the 24,321 BESs associated with genes, 19,363 (80 %) were assigned to functional categories in accordance with the GO database (Fig. 2). Within the biological processes category (10,130 gene products or sequences in total), which describes the processes in which genes are placed, the most representative were metabolic (9,521), cellular (9,204), and DNA metabolic (6,837) process. In the molecular functions category (13,632 gene products or sequences), the most represented gene activities were binding (12,555), organic cyclic compound binding (8,692) and nucleic acid binding (7,478). In the cellular components category (4,225 gene products or sequences), which describes the cellular structure in which the gene product is located, the most abundant were cell part (3,936), intracellular part (3,500), and organelle (3,285). Among the 609 BES-SSRs corresponding to genes, 416 (68 %) were assigned in accordance with the GO terms, with representation from the three functional categories, including genes predominating in metabolic process (196) in the biological processes category, binding (256) in the molecular function category, and the cell part (119) in the cellular component category.

The analysis for the presence of TFs demonstrated that 2 % of the BESs (1,039 sequences) were represented by 65 families previously described in the literature by Pérez-Rodríguez et al. (2010). Among the 65 families identified, four predominated, comprising 57 % of the total (Fig. 3a); the largest representation was the transcription factor mTERF (37 %), followed by DDT (9 %), Tify (6 %), and ARR-B-MYB (5 %). Of the 2,000 BES-SSRs, 58 (2.9 %) were associated with TFs, representing 30 different families of which the bHLH and NAC families were the most abundant.

Of the BESs, 3 % (1,643) were associated with transposable elements (TEs), assigned to two classes according to their mechanism of transposition, which can be described as either *copy and paste* (class I TEs) or *cut and paste* (class II TEs) (Wicker et al. 2007). In the present study, Class I retroelements were predominant (98 %), and only 2 % were Class II, consisting of DNA transposons. Among the retroelements, the LTR retrotransposons (corresponding to long terminal repeats) were represented predominantly by the *Gypsy* (81 %) and *Copia* (17 %) families. For the Class II TEs, although much less abundant, the families

Table 4 Genetic descriptors related to 34 BES-SSRs of the analyzed common bean, including the number of alleles per locus ($\bar{A}$), observed heterozygosity (*Ho*), genetic diversity (H_E), and probability of identity (*PI*)

Marker	$\bar{A}$	H_o	H_E	*PI*
Compound				
PvComp10	19	0.013	0.918	1E−02
PvComp4	17	0.000	0.816	5E−02
PvComp9	9	0.023	0.703	1E−01
PvComp2	5	0.024	0.674	2E−01
PvComp8	9	0.012	0.605	2E−01
PvComp21	5	0.000	0.508	3E−01
PvComp27	3	0.000	0.388	4E−01
Mean	9.57	0.010	0.659	3.2E−07[a]
Standard error	2.34	0.004	0.068	–
Hexanucleotide				
PvHexa20	3	0.000	0.503	3.6E−01
PvHexa10	3	0.012	0.453	3.9E−01
PvHexa19	3	0.000	0.416	4.2E−01
PvHexa12	2	0.000	0.334	5.0E−01
PvHexa36	4	0.000	0.229	6.1E−01
PvHexa39	3	0.000	0.208	6.4E−01
PvHexa15	9	0.000	0.184	6.7E−01
Mean	3.86	0.002	0.332	7.6E−03[a]
Standard error	0.88	0.002	0.049	–
Pentanucleotide				
PvPenta13	3	0.000	0.611	2.3E−01
PvPenta4	3	0.000	0.555	2.8E−01
PvPenta10	6	0.012	0.435	3.4E−01
PvPenta14	2	0.000	0.334	5.0E−01
PvPenta19	4	0.000	0.312	4.9E−01
PvPenta8	2	0.000	0.263	5.8E−01
PvPenta16	3	0.000	0.167	7.0E−01
PvPenta5	3	0.000	0.134	7.5E−01
Mean	3.25	0.001	0.351	1.6E−03[a]
Standard error	0.450	0.001	0.061	–
Tetranucleotide				
PvTetra25	12	0.012	0.741	9.8E−02
PvTetra65	5	0.012	0.653	1.8E−01
PvTetra73	4	0.000	0.618	2.0E−01
PvTetra57	6	0.000	0.588	2.1E−01
PvTetra49	3	0.014	0.475	3.5E−01
PvTetra32	2	0.012	0.456	4.0E−01
PvTetra50	4	0.000	0.450	3.3E−01
PvTetra47	4	0.000	0.428	3.9E−01
PvTetra76	2	0.000	0.198	6.6E−01
Mean	4.67	0.006	0.512	9.2E−06[a]
Standard error	1.01	0.002	0.054	–
Trinucleotide				
PvTri8	11	0.000	0.821	5.2E−02
PvTri6	7	0.000	0.743	1.1E−01
PvTri5	10	0.000	0.635	1.8E−01
Mean	9.33	0.000	0.733	1.0E−03[a]
Standard error	1.20	0.000	0.054	–

[a] *PI* for all locus combinations

identified were *Mutator* (1.7 %), followed by *PIF/Harbinger* (0.25 %) and *Helitron* (0.05 %). Among the 2,000 BES-SSRs, only 38 (2 %) were identified as associated with TEs of which 33 were from the *Gypsy* family (87 %), four from *Copia* (10 %), and one from *Mutator* (3 %) (Fig. 3b).

Comparison of genomes of legumes

The alignment of BESs against the genomes of *P. vulgaris* Andean landrace G19833 (Pv*And*), *M. truncatula* (Mt), and *G. max* (Gm) is shown in Fig. 4a. Of the 52,270 BESs analyzed, 99.4, 14.4, and 2.7 % exhibited similarities with the genomic sequence of Pv*And*, Gm, and Mt, respectively. Of the 2,000 BES-SSRs, 99.4 % aligned to the genome of Pv*And*, 16.7 % to Gm and only 2.7 % to Mt. Only 53 BES-SSRs were similar between the Mt and Pv*And* genomes, whereas 331 were similar between Gm and Pv*And*. For the Pv*And* and Gm genomes with 10 or more BES-SSRs in common alignment was exhibited. The sequences that mapped to chromosome one of the common bean aligned primarily with sequences in soybean chromosome Gm19, whereas the sequences in chromosomes Pv3, Pv5, Pv7, and Pv11 aligned with chromosomes Gm17, Gm12, Gm10/Gm20, and Gm6/Gm12, respectively (Fig. 4a).

For the TFs identified in *P. vulgaris* (1,039 BESs), 100 % of them showed alignment over the Pv*And* genome, whereas 41 and 8.5 % of TFs aligned over the *Gm* and *Mt* genomes, respectively. Considering the chromosomes containing 10 or more TFs found in both the common bean and soybean genomes, the distribution of sequences over the genome was quite variable. The TFs located in chromosome one of *P. vulgaris* (Pv1) were identified on chromosome 19 of soybean (Gm19). Similarly, the sequences in Pv2, Pv3, Pv4, Pv9, and Pv11 corresponded to those on Gm1, Gm7, Gm16, Gm4, and Gm12, respectively. A broader distribution and dispersion was observed for the TFs on Pv5, which were identified on Gm1, Gm12, and Gm13, and the TFs on Pv7 were found on Gm10 and Gm20 (Fig. 4b).

For the BESs identified as TEs (1,643 BESs), 99.9 % aligned to the Pv*And* genome, 72.5 % to the *Gm* genome, and 7.4 % to the *Mt* genome. Regarding the distribution of the TE sequences throughout the genomes, an increased distribution was observed in the *Gm* compared with the Pv*And* genome. This was true for the TEs located in Pv5, Pv8, and Pv11, which were scattered over five (Gm1, Gm6, Gm15, Gm18, and Gm19), eight (Gm1, Gm2, Gm4, Gm5, Gm6, Gm15, Gm18, and Gm19) and six (Gm1, Gm5, Gm10, Gm15, Gm18, and Gm19) soybean chromosomes, respectively. For the TEs common to the *P.* vulgaris and *Mt* genomes, 35 TEs were located on Pv11 and their corresponding location was on Mt3 (Fig. 4c).

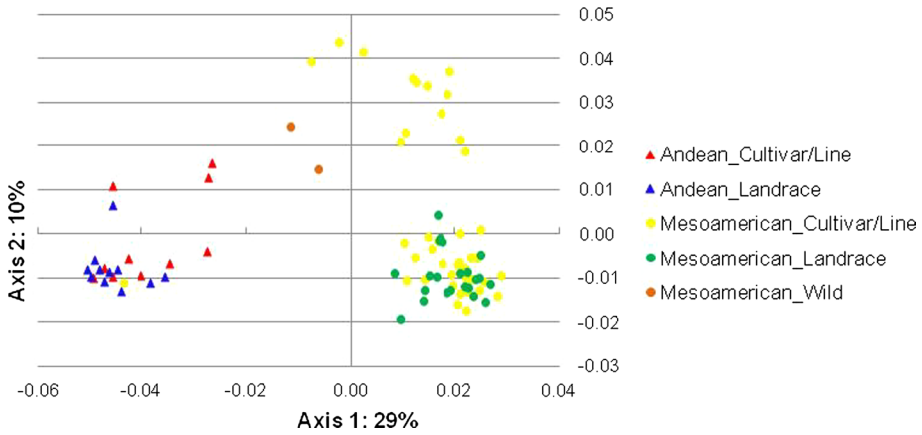

Fig. 1 PCoA analysis for 88 common bean genotypes using 34 BES-SSRs based on the average genetic distance of Rogers' coefficient modified by Wright

Discussion

The analysis of BESs has been a valuable resource for the study of plant genomes. In this study, the information derived from the BESs was incorporated immediately into the international project for genome sequencing of the Mesoamerican common bean variety, assisting in the investigation and assembly of the whole-genome reference sequence. Correspondingly, the availability of these sequences has allowed the analysis of the composition of the common bean genome and the identification SSRs based on long repeats, which have been little explored to date in common bean, providing genome information that can be used in genetic studies, including germplasm characterization, QTL mapping and associative genetics.

Identification and distribution of SSR

It has been shown that data mining of genomic sequences derived from BESs is useful for the generation of SSR markers in various legumes, such as the 3,290 SSRs identified in soybean (Shoemaker et al. 2008), 18,149 in pigeon pea (Bohra et al. 2011), 1,424 in peanut (Wang et al. 2012), 6,845 in chickpea (Thudi et al. 2011), and 875 in the common bean (Córdoba et al. 2010). The number of SSR regions identified in *P. vulgaris* in this study demonstrated that this type of sequence is fairly abundant; the number of SSRs (3,789) reported was four times more than that reported by Córdoba et al. (2010). The frequency of SSRs in the genome of the common bean (8.36 Kb/SSR) was close to the estimates reported for other legumes, such as soybean (6.82 Kb/SSR), pigeon pea (5.65 Kb/SSR) and chickpea (4.85 Kb/SSR) (Saini et al. 2008; Bohra et al. 2011; Thudi et al. 2011), whereas for cultivated peanut, the frequency was estimated at 27.32 Kb/SSR (Wang et al. 2012).

Indeed, the abundance and density of SSRs along the genome varies because of a number of factors and can be determined directly by the characteristics of the genome of each species (Biswas et al. 2012; Tóth et al. 2000; Morgante et al. 2002), or indirectly, depending on the strategy used to identify the SSRs. Features of the different restriction enzymes used for the construction of the BAC libraries increases the likelihood of incorporating all regions present in the genome and consequently, increases the chances of identifying more SSRs (Wang et al. 2012). In addition, the parameter settings of the softwares used to identify and quantify the SSRs from the genome sequences contribute largely to the observed discrepancies (Leclercq et al. 2007). A recent study conducted by Blair and Hurtado (2013) showed that three different bioinformatics procedures had influenced the estimates of the number and types of SSRs over the same set of 21,000 ESTs. In the present study, due to the large number of SSRs found in the *P. vulgaris* genome, it was possible to identify and select markers taking into consideration the types and abundance of motifs.

The other important finding of this study was that the estimates of SSR distribution observed in the non-coding regions (BES-SSRs: 0.026 %) were close to the SSR frequency found in the coding regions (EST-SSRs: 0.021 %) by Garcia et al. (2011). According to Victoria et al. (2011), SSRs in plant genomes follow a random distribution pattern, however, studies in *Arabidopsis* (Morgante et al. 2002) and *Citrus* (Biswas et al. 2012) demonstrated a higher frequency of SSRs in coding compared to non-coding regions. From a practical viewpoint, a more homogeneous distribution of SSRs throughout the genome is interesting because it allows the direct mapping of the SSRs in *P. vulgaris* sequences isolated either from sources of expressed sequences or from random genomic sequences, ensuring a broad distribution and sampling of this genome.

BAC end sequences: SSRs

As for the practical goal of developing molecular tools for the genotyping of the common bean, sequences containing

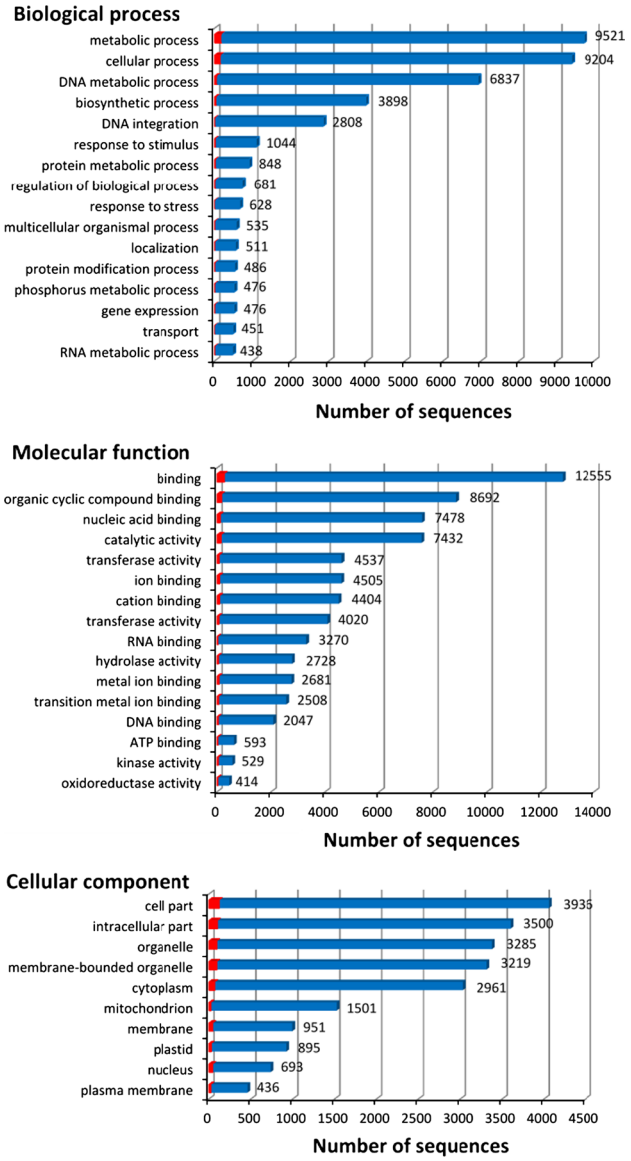

Fig. 2 Functional classification of 52,270 BESs (*blue bars*) and of 2,000 BESs (*red bars*) with microsatellites (BES-SSRs) derived from *P. vulgaris* annotated with GO terms, including attributes based on biological process, molecular function and cellular component. Numbers along the *bars* represent the total number of BES contained in each GO terms. Only major categories are presented

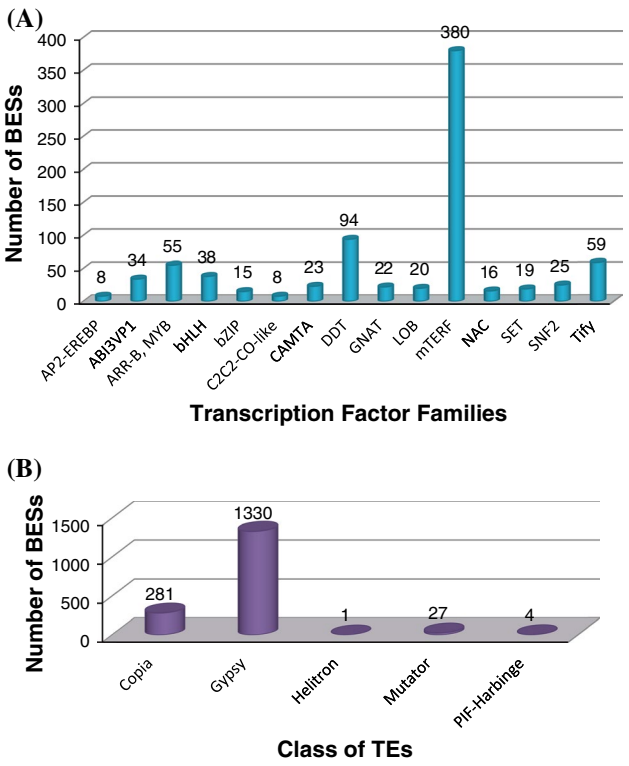

Fig. 3 Association of (**a**) transcription factor (TFs) families of (**b**) transposable elements (TEs) with BESs. Only major families are represented for TFs

tandem repeats of a single nucleotide were excluded from this study. The most abundant SSRs observed in this study were the trinucleotide motifs (46 %) and not the dinucleotide motifs (26 %) as described in other legumes (Wang et al. 2012; Thudi et al. 2011; Bohra et al. 2011; Córdoba et al. 2010). However, the predominance of trinucleotide motifs was also found in soybean (Shoemaker et al. 2008) and *Arabidopsis* (Lawson and Zhang 2006). Although it is expected that trinucleotide motifs will be found more frequently in coding regions because of potential negative selection against mutations that alter the reading frame of codons (Morgante et al. 2002), more recent studies have demonstrated a distinct pattern of motif distribution between different gene regions (5′UTR, 3′UTR, introns, and exons) (Lawson and Zhang 2006). Therefore, polymorphisms along gene region extensions can be exploited as efficiently as in the structural genome, ensuring that estimates of genetic diversity are efficiently sampled throughout the genome.

The identification, characterization and quantification of a larger number of SSR motifs, as well as the evaluation of their genome distribution are of fundamental importance to the development of new markers for the common bean. Among the dinucleotide repeats, the AT motifs (16.57 %) were more abundant, which is consistent with previous studies in beans (Córdoba et al. 2010) and other legumes, such as peanut, pigeon pea, and soybean (Wang et al. 2012; Bohra et al. 2011; Shultz et al. 2007). Despite the higher abundance of AT/TA repeats, sampling this motif through the construction of enriched libraries has not been explored yet, since the AT repeat is self-complementary and is difficult to screen for by colony hybridization. In this study, a single GC motif was identified among the 559 dinucleotide motifs characterized, emphasizing that the (GC)n motif is extremely rare in eukaryotic genomes as previously

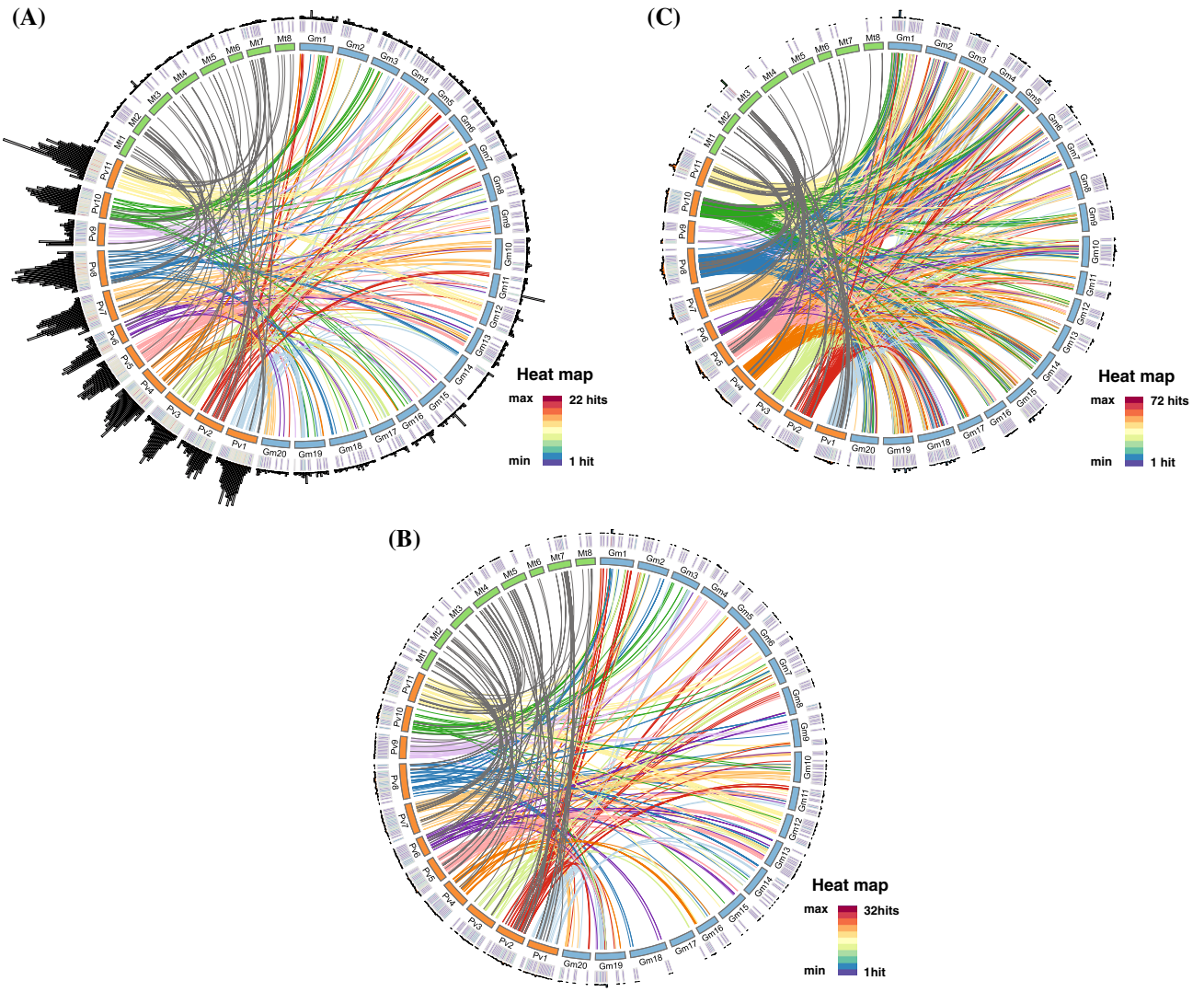

Fig. 4 Comparative analysis of high-resolution BES from *P. vulgaris* (Mesoamerican variety) relative to the genome of *P. vulgaris* Andean variety (*orange bars*), *M. truncatula* (*green bars*), and *G. max* (*blue bars*). **a** Alignment of BES-SSRs in relation to the other 3 genomes is represented by *3 concentric circles*; the most *external circle* represents the alignment of 52,270 BESs (*gray*) against the 3 legume genomes, the *intermediate circle* represents the heat map in which the *colored bars* indicate the density of the BLAST hits by 2-Mb sequences (2 Mb), and the *inner circle* represents the haploid chromosomes of the *P. vulgaris* (Pv1-11), *M. truncatula* (Mt1-8), and *G. max* (Gm1-20) genomes. **b**, **c** Represent the alignments of TFs and TEs, respectively. The *lines* in the three panels connect the chromosome regions of *P. vulgaris* Andean variety with of *G. max* (*colored lines*) or *M. truncatula* (*gray lines*) which showed high similarity with the same BES

reported by Katti et al. (2001). The low frequency of poly(C) and (G) n repeats was specifically observed in common beans (Córdoba et al. 2010; Chen et al. 2014) and other legume (Mun et al. 2006; Bohra et al. 2011). According to Eustice et al. (2008), the (CG)*n* repeats tend to form secondary structures (hairpins) which are unstable in the genome thus; negative selection may be involved in the scarcity of this repeats over DNA sequences. For the trinucleotide motifs identified in this study, a predominance of AAG (17.12 %) was observed, which has also been reported as more abundant in peanuts (Wang et al. 2012). Building libraries for tetra-, penta-, and hexanucleotides is difficult because these motifs are less abundant in the genome, which restricts their availability for developing primer pairs and for use in the molecular characterization of the common bean. Therefore, the isolation of these repetitive regions from random genomic fragments was an attractive strategy, generating significant and important results for the development of functional SSR markers in the common bean.

BAC libraries of *P. vulgaris* are a good source of SSRs, and although only a fraction of these identified markers

were investigated in this study, there did not appear to be a limit to their level of polymorphism or to the use of this method for large-scale genotyping. In this study, a fraction of the tri-, tetra-, and hexanucleotides and compound SSRs were examined, as well the pentanucleotides identified from the BESs. The initial screen allowed the selection of markers with consistent amplification profiles and a potential for polymorphism detection from the initially evaluated genotype group exhibiting an amplification success rate of 79 %. Similar results for the amplification efficiency were reported for the common bean (76.5 %) and chickpea (79 %) in which the SSRs were derived from BESs (Córdoba et al. 2010; Thudi et al. 2011). Regarding the proportion of SSRs classified according to motif size using established criteria, there was a predominance of compound SSRs (56 %) categorized as Class I ($P \leq 0.05$), comprising repeats with longer DNA sequences. Previous studies have reported ratios above 30 % for this class of repeats in the common bean (Hanai et al. 2007; Garcia et al. 2011), which together with the data from this study, will enable the investigation of these repeats for potential use in the genetic analysis of this species. The availability of large genomic DNA sequences will continuously promote the identification of class I SSRs, overcoming the limitation of *short-sequence DNA* repeat (*SSR*) loci identified from cDNA and enriched genomic libraries. Longer repeats in compound motifs than perfect motifs has been reported and may result from more complex evolutionary patterns by assuming that a compound motif may have more chances to gain a new repeat unit since two or more individual microsatellites are found directly adjacent to each other (Li et al. 2003). This explains a higher mean number of repeats found in compound SSRs, predominantly class I, than perfect ones as reported in rice (Temnykh et al. 2001).

Molecular characterization of BES-SSRs

Among the SSRs evaluated in this study, all six with a $H_E \geq 0.7$ were Class I, comprising three compound SSRs and three perfect SSRs (two tri-and one tetranucleotide), following the pattern of higher polymorphism for having more than one type of repeat and/or larger SSR region, as previously demonstrated (Byrne et al. 1996; Cho et al. 2000). Regarding the nature of the SSR, the average number of alleles per locus ranged from 9.57 for this compound SSR to 3.25 for pentanucleotides, and according to Cardoso et al. (2013), the average number of alleles observed was 7.79 in dinucleotides, indicating that compound SSRs are extremely interesting and should be prioritized for molecular characterization. In the tetra-, penta- and hexanucleotide type loci, the average number of alleles was 4.67, 3.25, and 3.86, respectively, significantly lower than the values reported for dinucleotides, which because of the higher mutation rates in dinucleotide type repeats, is consistent with expected values (Renwick et al. 2001). Meanwhile, seven SSRs (one hexa-, two penta-, and four tetranucleotides) were identified with H_E average value of 0.610 and 5.14 alleles/locus number, predominantly on PvTetra25, which has 12 alleles and a H_E of 0.741. These results suggested that it is possible to identify a higher number of SSRs based on repeats ≥trinucleotides with high levels of genetic information content, which is comparable or superior to currently available dinucleotides. For the common bean, there are no known reports on the development of SSRs based on large motifs, and the tetra-, penta- and hexanucleotides identified by Blair et al. (2011b) were not polymorphic, indicating that the SSRs developed in this study are the first of this type for *P. vulgaris*. Furthermore, the tri-, tetra- and penta- and hexanucleotide microsatellites exhibited less "stutter" and therefore, allowed for a clearer identification of alleles, reducing genotyping errors.

Taking advantage of the annotated common bean genome, 21.43 % of the BES in this study presented similarity with the Andean genome. With minimal effort, the functional annotation of the BES-SSRs revealed 13 loci that mapped within genic regions, including genes related to plant defense (*calcium-transporting ATPase*) (Frei dit Frey et al. 2012) and the response to biotic and abiotic stresses (*constans-like zinc finger protein*) (Noguero et al. 2013), of which are important mechanisms that can assist in developing of higher common bean plants. Among the 13 SSR markers associated with gene products, 11 were polymorphic including one tri- (PvTri6), two tetra- (PvTetra32 and PvTetra57), one penta- (PvPenta4), five hexa- (PvHexa12, PvHexa15, PvHexa19, PvHexa20 and PvHexa39) and two compound (PvComp2 and PvComp21). The identification of marker loci linked to genomic regions of interest allows for new study approaches to assess the allelic variability at loci of interest, identification of potential markers for use in assisted selection, and development of new and different common bean cultivars.

Genomic composition of BES

TFs are proteins that bind to DNA in eukaryotes to enable gene transcription and play a fundamental role in gene regulation (Pérez-Rodríguez et al. 2010). Therefore, the identification of TFs (2 %) in the common bean represents a relevant contribution towards the understanding of the regulatory networks of genes that may be associated with traits of interest for plant breeding. Members of the mTERF family, represented in 37 % of the TFs identified in this study, are regulators of mitochondrial transcription and control of mtDNA replication (Roberti et al. 2009). Although *Arabidopsis* has a more compact genome compared to *P. vulgaris*, a wide variety of mTERF family genes derived from

genomic duplication processes have been reported in that species, and appears to play an important role in the expression of genes in organelles, photosynthesis processes, and cellular respiration (Kleine 2012).

TEs are the most abundant repetitive components in higher eukaryote genomes and play important roles in regulating gene expression and affecting the composition and structure of genomes. Therefore, the identification and characterization of TEs in the common bean genome is important for the study of phylogeny, application of repetitive genetic elements as molecular markers, understanding the genome organization and gene annotations, and the immediate application of assisting in the assembly of the *P. vulgaris* genome (Du et al. 2010). The diversity and abundance of TEs in plant genomes are highly variable among species (Kriedt et al. 2014) being influenced by several evolutionary factors, such as life history, population size and matting system (de la Chaux et al. 2012), making it difficult to perform a direct inter-species comparisons. In legumes, a wide variation is evident, and estimates close to those reported in this study (3 %) have been described for peanut (5.86 %), whereas those for soybean (17 %) and chickpea (49 %) are higher (Wang et al. 2012; Schmutz et al. 2010; Varshney et al. 2013). The predominance of class I (*retroelements*) relative to class II (DNA *transposons*) TEs demonstrated that the common bean resembles several plants with extended variable genomes (Bohra et al. 2011; Wang et al. 2012; Biswas et al. 2012; Wu et al. 2012; Kim et al. 2013; Dereeper et al. 2013). LTRs were the most represented among the retroelements, demonstrating a wide disparity in numbers between the *Gypsy* (81 %) and *Copia* (17 %) families. This same pattern was observed in soybean (Schmutz et al. 2010), which is most likely related to the differential inclusion of these elements during the evolution of these species.

In this study, the high similarity at the genome level between the common bean and soybean reflects the phylogenetic relationship of these legumes that belong to the same Phaseoloid clade, whereas limited synteny was presented with *Medicago* species that are included to the Hologalegina clade (Gepts et al. 2005). The BESs and TFs located on chromosome 7 in the common bean aligned with chromosomes 10 and 20 in soybean, which is consistent with a report by Galeano et al. (2011) in which this relationship was first demonstrated. The complete genome sequencing of soybean demonstrated that Gm20 is highly homologous to the long arm of Gm10, and a number of parts of the chromosome 20 are distributed over several other places in the genome (Schmutz et al. 2010). However, the wide distribution of most common bean sequences in the genome of soybean does not exhibit a 1:2 ratio, which is borne out by chromosomal breakage and union, possibly associated with a tetraploidization event in the soybean genome (Galeano et al. 2011); this is consistent with evidence that the soybean genome was split and reassembled at some point after the duplication event (McClean et al. 2010).

Common beans genomic resources

The analysis of the *P. vulgaris* genome by BES sequencing proved informative, enabling access to some features of the genome not described to date. This technique was shown to be an excellent strategy to obtain molecular tools for analysis of the common bean, a crop of great economic and social interest. The sequenced crop legume genomes, including common bean, allowed to evaluate the SSRs as their broad distribution over the entire genome, and also corroborated the BES sequencing strategy as adequate to be used for other species without genome information available. As a benefit of the common bean genome sequencing (Schmutz et al. 2014), where a large amounts of quality DNA sequence data were made available in a public database, an almost unlimited number of SSR markers are likely to be developed. As effect, breeders will benefit with a genome-wide SSRs markers to be applied to many fields of the molecular breeding. In addition, the deeper understanding of the expressed common bean genes from the reference genome will expand knowledge of the genetic architecture of complex traits, enabling to determine new breeding approaches. Major benefits of the presented study included the generation of a panel of SSRs based on long motifs ($\geq$tetranucleotides) of a satisfactory analytical quality, the ease of interpretation, and the clear identification of alleles. Once used by other research groups, the incorporation of the data generated herein will create a real possibility for sharing and comparing molecular profiles from the international germplasm database of the common bean, in which the international genotypes BAT93, JALO EEP558 and G19833 may be used for data integration. Therefore, from a practical viewpoint, at least four operating tetraplex systems, including SSRs with high informative power (average H_E of 0.668) and discrimination (PI combined of 1.26E−14), were made available as an operational tools for international use. Additionally, new SSR markers sequences are available to be evaluated in future studies, through alignment with the common bean reference genome, with the possibility of alternative designs for new primers aimed at increasing the compatibility of co-amplifications, reducing costs and generating flexibility in the genotyping process. In practice, both the generated sequences as well as the data extracted from the analyses can be used in other large-scale projects.

Acknowledgments We thank the National Council for Scientific and Technological Development (CNPq) for the grants to GSP, CB, EGB and RPV; the Coordination for the Improvement of Higher Education Personnel/Ministry of Education (CAPES/MEC) for the grants

to BSFM, the Minas Gerais State Research Foundation (FAPEMIG) for the grant to TS; and the Brazilian Agricultural Research Corporation (EMBRAPA) for financial support for this research. We also thank the anonymous referees for providing important comments that improved the first version of this manuscript.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

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 Res 25(17):3389–3402

Asfaw A, Blair MW, Struik PC (2012) Multienvironment quantitative trait loci analysis for photosynthate acquisition, accumulation, and remobilization traits in common bean under drought stress. G3 (Bethesda) 2(5):579–595. doi:10.1534/g3.112.002303

Biswas MK, Chai L, Mayer C, Xu Q, Guo W, Deng X (2012) Exploiting BAC-end sequences for the mining, characterization and utility of new short sequences repeat (SSR) markers in *Citrus*. Mol Biol Rep 39(5):5373–5386. doi:10.1007/s11033-011-1338-5

Blair MW, Hurtado N (2013) EST-SSR markers from five sequenced cDNA libraries of common bean (*Phaseolus vulgaris* L.) comparing three bioinformatic algorithms. Mol Ecol Resour 13(4):688–695. doi:10.1111/1755-0998.12099

Blair MW, Pedraza F, Buendia HF, Gaitán-Solís E, Beebe SE, Gepts P, Tohme J (2003) Development of a genome-wide anchored microsatellite map for common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 107:1362–1374. doi:10.1007/s00122-003-1398-6

Blair MW, Astudillo C, Rengifo J, Beebe SE, Graham R (2011a) QTL analyses for seed iron and zinc concentrations in an intragenepool population of Andean common beans (*Phaseolus vulgaris* L.). Theor Appl Genet 122(3):511–521. doi:10.1007/s00122-010-1465-8

Blair MW, Hurtado N, Chavarro CM, Muñoz-Torres MC, Giraldo MC, Pedraza F, Tomkins J, Wing R (2011b) Gene-based SSR markers for common bean (*Phaseolus vulgaris* L.) derived from root and leaf tissue ESTs: an integration of the BMc series. BMC Plant Biol 11:50. doi:10.1186/1471-2229-11-50

Blair MW, Hurtado N, Sharma P (2012a) New gene-derived simple sequence repeat markers for common bean (*Phaseolus vulgaris* L.). Mol Ecol Resour 12:661–668. doi:10.1111/j.1755-0998.2012.03136.x

Blair MW, Soler A, Cortés AJ (2012b) Diversification and population structure in common beans (*Phaseolus vulgaris* L.). PLoS ONE 7(11):e49488. doi:10.1371/journal.pone.0049488

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. doi:10.1007/s00122-012-1999-z

Bohra A, Dubey A, Saxena RK et al (2011) Analysis of BAC-end sequences (BESs) and development of BES-SSR markers for genetic mapping and hybrid purity assessment in pigeonpea (*Cajanus* spp.). BMC Plant Biol 11:56. doi:10.1186/1471-2229-11-56

Broughton WJ, Hernández G, Blair M, Beebe S, Gepts P, Vanderleyden J (2003) Beans (*Phaseolus* spp.)—model food legumes. Plant Soil 252:55–128. doi:10.1023/A:1024146710611

Burle ML, Fonseca JR, Kami JA, Gepts P (2010) Microsatellite diversity and genetic structure among common bean (*Phaseolus vulgaris* L.) landraces in Brazil, a secondary center of diversity. Theor Appl Genet 121:801–813. doi:10.1007/s00122-010-1350-5

Byrne N, Marquez-Garcia MI, Smith DS, Moran GF (1996) Conservation and genetic diversity of microsatellite loci in the genus *Eucalyptus*. Aust J Bot 44(3):331–341. doi:10.1071/BT9960331

Cannon SB, May GD, Jackson SA (2009) Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. Plant Physiol 151(3):970–977

Cardoso PCB, Veiga MM, Menezes IPP, Valdisser PAMR, Borba TCO, 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. doi:10.4238/2013

Cardoso PCB, Brondani C, Menezes IPP, Valdisser PAMR, Borba TCO, Del Peloso MJ, Vianello RP (2014) Discrimination of common bean cultivars using multiplexed microsatellite markers. Genet Mol Res 13(1):1964–1978. doi:10.4238/2014.March.24.1

Castelo AT, Martins W, Gao GR (2002) TROLL—tandem repeat occurrence locator. Bioinformatics 18(4):634–636

Chen M, Wu J, Wang L, Zhang X, Blair MW, Jia J, Wang S (2014) Development of mapped simple sequence repeat markers from common bean (*Phaseolus vulgaris* L.) based on genome sequences of a Chinese landrace and diversity evaluation. Mol Breeding 33(2):489–496. doi:10.1007/s11032-013-9949-2

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100(5):713–722

Choi H, Kim D, Uhm T, Limpens E, Lim H, Mun JH, Kalo P, Penmetsa RV, Seres A, Kulikova O, Roe BA, Bisseling T, Kiss GB, Cook DR (2004) A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*. Genetics 166(3):1463–1502

Chou H, Holmes MH (2001) DNA sequence quality trimming and vector removal. Bioinformatics 17(12):1093–1104

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(18):3674–3676. doi:10.1093/bioinformatics/bti610

Córdoba JM, Chavarro C, Schlueter JA, Jackson SA, Blair MW (2010) Integration of physical and genetic maps of common bean through BAC-derived microsatellite markers. BMC Genom 11:436. doi:10.1186/1471-2164-11-436

Cortés AJ, Chavarro MC, Blair MW (2011) SNP marker diversity in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 123(5):827–845. doi:10.1007/s00122-011-1630-8

de la Chaux N, Tsuchimatsu T, Shimizu KK, Wagner A (2012) The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. Mob DNA 3(1):2. doi:10.1186/1759-8753-3-2

Dereeper A, Guyot R, Tranchant-Dubreuil C et al (2013) BAC-end sequences analysis provides first insights into coffee (*Coffea canephora* P.) genome composition and evolution. Plant Mol Biol 83(3):177–189. doi:10.1007/s11103-013-0077-5

Doyle JJ, Luckow MA (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol 131(13):900–910. doi:10.1104/pp.102.018150.groups

Du J, Grant D, Tian Z, Nelson RT, Zhu L, Shoemaker RC, Ma J (2010) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genom 11:113. doi:10.1186/1471-2164-11-113

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conser Genet Resour 4(2):359–361. doi:10.1007/s12686-011-9548-7

Eustice M, Yu Q, Lai CW, Hou S, Thimmapuram J, Liu L, Alam M, Moore PH, Presting GG, Ming R (2008) Development and

application of microsatellite markers for genomic analysis of papaya. Tree Genet Genomes 4(2):333–341. doi:10.1007/s11295-007-0112-2

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. doi:10.1111/j.1365-294X.2005.02553.x

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8(3):175–185. doi:10.1101/gr.8.3.175

FAOSTAT (2009) Food and Agriculture Organization of the United Nations—Statistical Database. http://faostat.fao.org/. Accessed 10 Oct 2013

Frei dit Frey N, Mbengue M, Kwaaitaal M et al (2012) Plasma membrane calcium ATPases are important components of receptor-mediated signaling in plant immune responses and development. Plant Physiol 159(2):798–809. doi:10.1104/pp.111.192575

Galeano CH, Fernandez AC, Franco-Herrera N, Cichy KA, McClean PE, Vanderleyden J, Blair MW (2011) Saturation of an intra-gene pool linkage map: towards a unified consensus linkage map for fine mapping and synteny analysis in common bean. PLoS ONE 6(12):e28135. doi:10.1371/journal.pone.0028135

Galeano CH, Cortés AJ, Fernández AC, Soler Á, Franco-Herrera N, Makunde G, Vanderleyden J, Blair MW (2012) Gene-based single nucleotide polymorphism markers for genetic and association mapping in common bean. BMC Genet 13:48. doi:10.1186/1471-2156-13-48

Garcia RAV, Rangel PN, Brondani C, Martins WS, Melo LC, Carneiro MS, Borba TC, Brondani RP (2011) The characterization of a new set of EST-derived simple sequence repeat (SSR) markers as a resource for the genetic analysis of *Phaseolus vulgaris*. BMC Genet 12:41. doi:10.1186/1471-2156-12-41

Gepts P, Beavis WD, Brummer EC, Shoemaker RC, Stalker HT, Weeden NF, Young ND (2005) Legumes as a model plant family genomics for food and feed report of the cross-legume advances through genomics conference. Plant Physiol 137(4):1228–1235. doi:10.1104/pp.105.060871

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross strategy and RAPD markers. Genetics 137(4):1121–1137

Hanai LR, de Campos T, Camargo LE et al (2007) Development, characterization, and comparative analysis of polymorphism at common bean SSR loci isolated from genic and genomic sources. Genome 50(3):266–277. doi:10.1139/G07-007

Hanai LR, Santini L, Camargo LE, Fungaro MH, Gepts P, Tsai SM, Vieira ML (2010) Extension of the core map of common bean with EST-SSR, RGA, AFLP, and putative functional markers. Mol Breed 25(1):25–45. doi:10.1007/s11032-009-9306-7

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9(9):868–877. doi:10.1101/gr.9.9.868

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(14):1801–1806. doi:10.1093/bioinformatics/btm233

Kalavacharla V, Liu Z, Meyers BC, Thimmapuram J, Melmaiee K (2011) Identification and analysis of common bean (*Phaseolus vulgaris* L.) transcriptomes by massively parallel pyrosequencing. BMC Plant Biol 11:135. doi:10.1186/1471-2229-11-135

Kami J, Poncet V, Geffroy V, Gepts P (2006) Development of four phylogenetically-arrayed BAC libraries and sequence of the APA locus in *Phaseolus vulgaris*. Theor Appl Genet 112(6):987–998. doi:10.1007/s00122-005-0201-2

Kang W, Hérbert JM (2012) A Sox2 BAC transgenic approach for targeting adult neural stem cells. PLoS ONE 7(11):e49038. doi:10.1371/journal.pone.0049038

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18(7):1161–1167

Kim H, San Miguel P, Nelson W, Collura K, Wissotski M, Walling JG, Kim JP, Jackson SA, Soderlund C, Wing RA (2007) Comparative physical mapping between *Oryza sativa* (AA genome type) and *O. punctata* (BB genome type). Genetics 176(1):379–390. doi:10.1534/genetics.106.068783

Kim C, Lee TH, Compton RO, Robertson JS, Pierce GJ, Paterson AH (2013) A genome-wide BAC end-sequence survey of sugarcane elucidates genome composition, and identifies BACs covering much of the euchromatin. Plant Mol Biol 81(1–2):139–147. doi:10.1007/s11103-012-9987-x

Kleine T (2012) *Arabidopsis thaliana* mTERF proteins: evolution and functional classification. Front Plant Sci 3:233. doi:10.3389/fpls.2012.00233

Kriedt RA, Cruz GMQ, Bonatto SL, Freitas LB (2014) Novel transposable elements in solanaceae: evolutionary relationships among Tnt1-related sequences in wild petunia species. Plant Mol Biol Rep 32(1):142–152. doi:10.1007/s11105-013-0626-8

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639–1645. doi:10.1101/gr.092759.109

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(5):979–992. doi:10.1007/s00122-008-0955-4

Lawson MJ, Zhang L (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biol 7(2):R14. doi:10.1186/gb-2006-7-2-r14

Leclercq S, Rivals E, Jarne P (2007) Detecting microsatellites within genomes: significant variation among algorithms. BMC Bioinformatics 8:125. doi:10.1186/1471-2105-8-125

Li YC, Fahima T, Röder MS, Kirzhner VM, Beiles A, Korol AB, Nevo E (2003) Genetic effects on microsatellite diversity in wild emmer wheat (*Triticum dicoccoides*) at the Yehudiyya microsite, Israel. Heredity 90(2):150–156. doi:10.1038/sj.hdy.6800190

Martin-Cabrejas MA, Esteban RM, Perez P, Maina G, Waldron KW (1997) Changes in psyscochimical properties of dry beans (*Phaseolus vulgaris* L.) during long term storage. J Agric Food Chem 47:3223–3227

Martins W, de Sousa D, Proite K, Guimarães P, Moretzsohn M, Bertioli D (2006) New softwares for automated microsatellite marker development. Nucleic Acids Res 34(4):e31. doi:10.1093/nar/gnj030

McClean PE, Mamidi S, McConnell M, Chikara S, Lee R (2010) Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. BMC Genom 11:184. doi:10.1186/1471-2164-11-184

Mitchell DC, Lawrence FR, Hartman TJ, Curran JM (2009) Consumption of dry beans, peas, and lentils could improve diet quality in the US population. J Am Diet Assoc 109(5):909–913

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30(2):194–200. doi:10.1038/ng822

Mun JH, Kim DJ, Choi HK, Gish J, Debellé F, Mudge J, Denny R, Endré G, Saurat O, Dudez AM, Kiss GB, Roe B, Young ND, Cook DR (2006) Distribution of microsatellites in the genome of Medicago truncatula: a resource of genetic markers that integrate genetic and physical maps. Genetics 172(4):2541–2555. doi:10.1534/genetics.105.054791

Noguero M, Atif RM, Ochatt S, Thompson RD (2013) The role of the DNA-binding One Zinc Finger (DOF) transcription factor family in plants. Plant Sci 209:32–45. doi:10.1016/j.plantsci.2013.03.016

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28(19):2537–2539. doi:10.1093/bioinformatics/bts460

Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LG, Rensing SA, Kersten B, Mueller-Roeber B (2010) PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res 38(1):D822–D827. doi:10.1093/nar/gkp805

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

Ragupathy R, Rathinavelu R, Cloutier S (2011) Physical mapping and BAC-end sequence analysis provide initial insights into the flax (*Linum usitatissimum* L.) genome. BMC Genom 12:217. doi:10.1186/1471-2164-12-217

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(4):1211–1227

Renwick A, Davison L, Spratt H, King JP, Kimmel M (2001) DNA dinucleotide evolution in humans : fitting theory to facts. Genetics 159(2):737–747

Roberti M, Polosa PL, Bruni F, Manzari C, Deceglie S, Gadaleta MN, Cantatore P (2009) The MTERF family proteins: mitochondrial transcription regulators and beyond. Biochim Biophys Acta 1787(5):303–311. doi:10.1016/j.bbabio.2009.01.013

Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138. doi:10.1046/j.1471-8286.2003.00566.x

Rozen S, Skaletsky H (2000) Primer3 for general users and for biologist programmers. In: Misener S, Krawetz SA (eds) Bioinformatics methods and protocols. Methods molecular biology, vol 132. The Humana Press Inc., Totowa, New Jersey, pp 365–386

Saini N, Shultz J, Lightfoot DA (2008) Re-annotation of the physical map of *Glycine max* for polyploid-like regions by BAC end sequence driven whole genome shotgun read assembly. BMC Genom 9:323. doi:10.1186/1471-2164-9-323

Schlueter JA, Goicoechea JL, Collura K et al (2008) BAC-end sequence analysis and a draft physical map of the common bean (*Phaseolus vulgaris* L.) genome. Trop Plant Biol 1(1):40–48. doi:10.1007/s12042-007-9003-9

Schmutz J, Cannon SB, Schlueter J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463(7278):178–183. doi:10.1038/nature08670

Schmutz J, McClean PE, Mamidi S et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46(7):707–713. doi:10.1038/ng.3008

Shoemaker RC, Grant D, Olson T et al (2008) Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. Genome 51(4):294–302. doi:10.1139/G08-010

Shultz JL, Kazi S, Bashir R, Afzal JA, Lightfoot DA (2007) The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor Appl Genet 114(6):1081–1090. doi:10.1007/s00122-007-0501-9

Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, Rana JC, Singh NK, Sharma TR (2011) Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. PLoS ONE 6(6):e21298. doi:10.1371/journal.pone.0021298

Staden R, Beal KF, Bonfield JK (1999) The staden package, 1998. In: Misener S, Krawetz SA (ed) Bioinformatics methods and protocols. Methods molecular biology, vol 132. The Humana Press Inc., Totowa, New Jersey, pp 115–130

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17(16):6463–6471. doi:10.1093/nar/17.16.6463

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11(8):1441–1452. doi:10.1101/gr.184001

Thudi M, Bohra A, Nayak SN et al (2011) Novel SSR markers from BAC-end sequences, DArT arrays and a comprehensive genetic map with 1,291 marker loci for chickpea (*Cicer arietinum* L.). PLoS ONE 6(11):e27275. doi:10.1371/journal.pone.0027275

Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10(7):967–981. doi:10.1101/gr.10.7.967

Vallone PM, Butler JM (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotechniques 37(2):226–231

Varshney RK, Song C, Saxena RK et al (2013) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat Biotechnol 31(3):240–246. doi:10.1038/nbt.2491

Victoria FC, Maia LC, de Oliveira AC (2011) In silico comparative analysis of SSR markers in plants. BMC Plant Biol 11:15. doi:10.1186/1471-2229-11-15

Wang H, Penmetsa RV, Yuan M et al (2012) Development and characterization of BAC-end sequence derived SSRs, and their incorporation into a new higher density genetic map for cultivated peanut (*Arachis hypogaea* L.). BMC Plant Biol 12:10. doi:10.1186/1471-2229-12-10

Wang Z, Yan H, Fu X, Li X, Gao H (2013) Development of simple sequence repeat markers and diversity analysis in alfalfa (*Medicago sativa* L.). Mol Biol Rep 40(4):3291–3298. doi:10.1007/s11033-012-2404-3

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8(12):973–982. doi:10.1038/nrg2165

Wu J, Gu YQ, Hu Y, You FM, Dandekar AM, Leslie CA, Aradhya M, Dvorak J, Luo MC (2012) Characterizing the walnut genome through analyses of BAC end sequences. Plant Mol Biol 78(1–2):95–107. doi:10.1007/s11103-011-9849-y

Young ND, Debellé F, Oldroyd GED et al (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480(7378):520–524. doi:10.1038/nature10625

Yu K (2012) Bacterial artificial chromosome libraries of pulse crops: characteristics and applications. J Biomed Biotechnol 2012:493186. doi:10.1155/2012/493186

Yu K, Park SJ, Poysa V, Gepts P (2000) Integration of simple sequence repeat (SSR) markers into a molecular linkage map of common bean (*Phaseolus vulgaris* L.). J Hered 91(6):429–434. doi:10.1093/jhered/91.6.429

Zaiontz C (2013) Real statistical analysis using Excel. http://www.real-statistics.com. Accessed 20 May 2014

Zhu Z, Jiang W, Thompson HJ (2012) Edible dry bean consumption (*Phaseolus vulgaris* L.) modulates cardiovascular risk factors and diet-induced obesity in rats and mice. Br J Nutr 108(Suppl 1):66–73. doi:10.1017/S0007114512000839