UC Davis

UC Davis Previously Published Works

Title

Genome-wide identification of SNPs and copy number variation in common bean (*Phaseolus vulgaris* L.) using genotyping-by-sequencing (GBS)

Permalink

<https://escholarship.org/uc/item/6sj302tq>

Journal

Molecular breeding : new strategies in plant improvement, 36(7)

ISSN

1380-3743

Authors

Ariani, Andrea
Berny Mier y Teran, Jorge Carlos
Gepts, Paul

Publication Date

2016-07-01

DOI

10.1007/s11032-016-0512-9

Peer reviewed

Molecular Breeding

Genome-wide identification of SNPs and Copy Number Variation in common bean (*Phaseolus vulgaris* L.) using Genotyping-By-Sequencing (GBS) --Manuscript Draft--

Manuscript Number:	MOLB-D-15-00576	
Full Title:	Genome-wide identification of SNPs and Copy Number Variation in common bean (<i>Phaseolus vulgaris</i> L.) using Genotyping-By-Sequencing (GBS)	
Article Type:	Original Article	
Keywords:	Common Bean; Copy Number Variation (CNV); Genome-wide SNPs calling; Genotyping-by-Sequencing (GBS); Next-generation Sequencing	
Corresponding Author:	Andrea Ariani, Ph.D. University of California Davis Davis, UNITED STATES	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	University of California Davis	
Corresponding Author's Secondary Institution:		
First Author:	Andrea Ariani, Ph.D.	
First Author Secondary Information:		
Order of Authors:	Andrea Ariani, Ph.D.	
	Jorge Carlos Berny	
	Paul Gepts	
Order of Authors Secondary Information:		
Funding Information:	U.S. Department of Agriculture (2013-67013-21224)	Paul Gepts
Abstract:	<p>Next Generation Sequencing (NGS) technologies have increased markedly the throughput of genetic studies, allowing the identification of several thousands of SNPs within a single experiment. Even though sequencing cost is rapidly decreasing, the price for whole genome re-sequencing of a large number of individuals is still costly, especially in plants with a large and highly redundant genome. In recent years, several reduced representation library (RRL) approaches has been developed for reducing the sequencing cost per individual. Among them, Genotyping-By-Sequencing (GBS) represents a simple, cost-effective, and highly multiplexed alternative for species with or without an available reference genome. However, this technology requires specific optimization for each species, especially for the restriction enzyme (RE) used. Here we report on the application of GBS in a test experiment with 18 genotypes of wild and domesticated <i>Phaseolus vulgaris</i>. After an in silico digestion with different RE of the <i>P. vulgaris</i> genome reference sequence, we selected CviAII as the most suitable RE for GBS in common bean based on the high frequency and even distribution of restriction sites. A total of 44,875 SNPs, 1,940 deletions and 1,693 insertions were identified, with 50% of the variants located in genic sequences and tagging 11,027 genes. SNPs and InDels distribution was positively correlated with gene density across the genome. In addition, we were able to also identify putative copy number variations (CNVs) of genomic segments between different genotypes. In conclusion, GBS with the CviAII enzyme results in thousands of evenly spaced markers and provides a reliable, high-throughput and cost-effective approach for genotyping both wild and domesticated common beans.</p>	
Suggested Reviewers:	<p>Elena Bitocchi Universita Politecnica delle Marche e.bitocchi@univpm.it She is an expert and published several papers regarding common bean genotyping</p>	

	and evolution.
	Leah McHale Professor, Ohio State University mchale.21@osu.edu She is an expert in the filed of soybean breeding and genomics
	Rajeev Varshney Scientist, ICRISTAT r.k.varshney@cgiar.org He is and expert in genotyping, genomics and breeding

1 **Genome-wide identification of SNPs and Copy Number Variation in Common Bean**
2 **(*Phaseolus vulgaris* L.) using Genotyping-By-Sequencing (GBS)**

3 Andrea Ariani*, Jorge Carlos Berny Mier y Teran, Paul Gepts

4 Department of Plant Sciences/MS1, University of California, 1 Shields Avenue, Davis, CA 95616-
5 8780, USA

6

7

8

9

10

11 **Corresponding author:**

12 Andrea Ariani

13 email: aariani@ucdavis.edu

14 Tel: +1-530-220-3208

15 Fax: +1-530-752-4361

16

17

18 **Acknowledgement**

19 This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported
20 by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. This project was supported
21 by Agriculture and Food Research Initiative (AFRI) Competitive Grant No. 2013-67013-21224
22 from the USDA National Institute of Food and Agriculture.

23

24 Abstract

25 Next Generation Sequencing (NGS) technologies have increased markedly the throughput of
26 genetic studies, allowing the identification of several thousands of SNPs within a single experiment.
27 Even though sequencing cost is rapidly decreasing, the price for whole genome re-sequencing of a
28 large number of individuals is still costly, especially in plants with a large and highly redundant
29 genome. In recent years, several reduced representation library (RRL) approaches has been
30 developed for reducing the sequencing cost per individual. Among them, Genotyping-By-
31 Sequencing (GBS) represents a simple, cost-effective, and highly multiplexed alternative for
32 species with or without an available reference genome. However, this technology requires specific
33 optimization for each species, especially for the restriction enzyme (RE) used. Here we report on
34 the application of GBS in a test experiment with 18 genotypes of wild and domesticated *Phaseolus*
35 *vulgaris*. After an *in silico* digestion with different RE of the *P. vulgaris* genome reference
36 sequence, we selected *CviAII* as the most suitable RE for GBS in common bean based on the high
37 frequency and even distribution of restriction sites. A total of 44,875 SNPs, 1,940 deletions and
38 1,693 insertions were identified, with 50% of the variants located in genic sequences and tagging
39 11,027 genes. SNPs and InDels distribution was positively correlated with gene density across the
40 genome. In addition, we were able to also identify putative copy number variations (CNVs) of
41 genomic segments between different genotypes. In conclusion, GBS with the *CviAII* enzyme results
42 in thousands of evenly spaced markers and provides a reliable, high-throughput and cost-effective
43 approach for genotyping both wild and domesticated common beans.

44

45 **Keywords:** Common Bean, Copy Number Variation (CNV), Genome-wide SNPs calling,
46 Genotyping-by-Sequencing (GBS), Next-generation Sequencing

47

48 **Introduction**

49 Common bean (*Phaseolus vulgaris* L.) is an important legume crop for human nutrition, being an
50 important source of protein, complex carbohydrates, fiber, and beneficial minerals for millions of
51 individuals worldwide (Broughton et al. 2003; Gepts et al. 2008). The species belongs to a large and
52 diverse genus that comprises 70-80 species, five of which have been domesticated (Freytag and
53 Debouck 2002). Among these domesticated species, common bean is the one with the broadest
54 geographic distribution and the highest agronomic, nutritional and economic value (Gepts 2014). It
55 is a diploid species with a haploid complement of 11 chromosomes and a genome size of ~587 Mb
56 (Schmutz et al., 2014).

57 Repeated experimental evidence highlights the existence of two different and genetically divergent
58 wild gene pools in common bean, called Mesoamerican and Andean gene pools, which underwent
59 domestication independently (Bitocchi et al. 2013; Gepts 1998; Kwak and Gepts 2009; Schmutz et
60 al. 2014) and diversified into distinct eco-geographic races (Singh et al. 1991; Chacón et al., 2007).
61 Indeed, the Andean gene pool is generally adapted to relatively higher altitudes and lower
62 temperature, while the Mesoamerican gene pool is adapted to lower altitudes and higher
63 temperatures (Beebe et al. 2011). A range of molecular markers have been developed and employed
64 in beans for the analysis of genetic diversity (domestication, gene pool divergence, and population
65 structure), linkage mapping and association studies, and marker-assisted selection (MAS) in
66 breeding programs (Blair et al. 2009; Kwak and Gepts 2009; Miklas et al. 2006; Talukder et al.
67 2010). However, marker development and use remain relatively expensive and the coverage of
68 available markers in the genome is still modest (Varshney et al. 2014).

69 Next Generation Sequencing (NGS) technologies are revolutionizing genetic studies and molecular
70 markers development by exponentially increasing the number of genetic variants that can be
71 discovered in a single experiment (Stapley et al. 2010). With these technologies, single nucleotide

polymorphism (SNP) and insertion-deletion (InDel) detection and genotyping have become feasible on a whole-genome scale and are widely applied to diversity and association studies in plants (Thudi et al. 2012; Varshney et al. 2014). Nevertheless, in spite of the reduced cost of sequencing technologies and the increased throughput and multiplexing, the cost of sequencing and genotyping large numbers of individuals is still prohibitive in plants with complex and repetitive genomes (Davey et al. 2011; Deschamps and Campbell 2010).

Several complexity reduction approaches that couple restriction enzyme (RE) genome digestion with NGS and SNP calling have been developed in the last years for high-throughput molecular marker discovery in different organisms (Davey et al. 2011). These approaches include reduced-representation libraries (RRLs) (Altshuler et al. 2000), restriction-site-associated DNA sequencing (RAD-Seq) (Baird et al. 2008), restriction enzyme sequence comparative analysis (RESCAN) (Monson-Miller et al. 2012), and GBS (Elshire et al. 2011).

GBS is a robust, high-throughput, cost-effective, and simple technique for obtaining thousands of markers from large numbers of individuals. It has been applied in genetic diversity studies to both plants and animal species (De Donato et al. 2013; Elshire et al., 2011; Glaubitz et al. 2014). In addition, in spite of the high percentage of missing data (Glaubitz et al. 2014; Beissinger et al. 2013), GBS technology has demonstrated its usefulness in the identification of quantitative trait loci (QTLs) in several crops like barley, soybean, chickpea, wheat, and common bean (Hart and Griffiths 2015; Iquiria et al. 2015; Li et al. 2015; Liu et al. 2014; Jaganathan et al. 2015). Despite its several advantages, GBS requires a species-specific optimization regarding the RE used to avoid repetitive regions of the genome and to determine marker number, distribution, and depth (Beissinger et al. 2013). For example, Hart and Griffiths (2015) found good SNP coverage in common bean using *ApeKI*, but there was uneven density distribution, probably because *ApeKI* is a methylation-sensitive enzyme. On the other hand, Zou et al. (2014) employed a methylation-

insensitive enzyme (*Hae*III) in common bean, but detected a high proportion of the SNPs (~73%) in repetitive regions. In the research reported here, an *in silico* analysis of different RE was performed to identify suitable enzymes for GBS in common beans, based on the availability of a *P. vulgaris* reference genome sequence (Schmutz et al. 2014). We then tested the GBS method with a panel of 18 wild and domesticated *P. vulgaris* accessions. Results are considered in light of read mapability among genotypes, marker distribution, and sequence depth. We evaluate also the possibility of using GBS with *Cvi*AII for identifying copy number variations (CNVs) across different genotypes. The information reported here will be useful for planning other GBS experiments in common bean using a larger number of genotypes, for both diversity and association studies.

105

106 **Materials and Methods**

107 ***In silico* digestion, library preparation and sequencing**

Thanks to the availability of the *P. vulgaris* whole-genome sequence (Schmutz et al. 2014), a survey of different restriction enzymes (RE) and their relative cutting sites could be performed. Using the Biopython suite (Cock et al. 2009), we selected enzymes that create a 'sticky' end after cleaving, cut only once for each recognition site, and do not recreate the restriction site after digestion. Elshire *et al.* (2011) suggested a methylation-sensitive enzyme to avoid repetitive elements of the genome when using GBS with maize, a plant with a large genome composed mainly of transposable elements (Schnable et al. 2009). In contrast, common bean has a relative small genome, with only 50% of the genome belonging to pericentromeric regions, which contain 26% of the genes (Schmutz et al. 2014). In addition, because of possible genotype-dependent differences in DNA methylation (Grativol et al. 2012), which could bias genotyping, we followed another approach. For each selected enzyme, we counted the number of recognition sites in the masked (where all the repetitive sequences are converted into string of Ns) and unmasked genome sequences, and kept

120 those enzymes that preferentially cut in the non-repetitive part of the genome, based on a binomial
121 test. In this sub-set of enzymes, we selected *Cvi*AII (recognition site C'ATG), because this enzyme
122 showed the higher restriction site count and displayed a preferential localization in the non-
123 repetitive part of the genome. Since *Ape*KI has been recently applied in common bean (Hart and
124 Griffiths 2015), we also compared the *in silico* distribution of digested fragments suitable for
125 sequencing (50 to 350 bp length) between *Ape*KI and *Cvi*AII across the genome (Supplementary
126 File S1).

127 In order to check the applicability of the GBS protocol using *Cvi*AII, a test experiment was
128 performed with 17 wild and domesticated *P. vulgaris* genotypes belonging to both Andean and
129 Mesoamerican gene pools. In addition, a representative of the wild ancestral gene pool from
130 northern Peru, G21245, was also included (Supplementary File S2). As internal control for our
131 analysis, we included also the common bean genotype used for generating the genome reference
132 sequence (G19833; Schmutz et al. 2014). Specific barcodes and adapters for *Cvi*AII were designed
133 with the GBS barcoded adapter generator (<http://www.deenabio.com/services/gbs-adapters>)
134 (Supplementary File S2).

135 DNA was extracted from freeze-dried bean leaves of greenhouse-grown plants using a modified
136 protocol of Pallotta et al. (2003) with an extra step consisting in re-suspension with 4 µl of RNase
137 and incubation for 30 minutes at 37°C. DNA quality was checked with NanoDrop Lite (Thermo
138 Fisher Scientific) and by 1% agarose gel electrophoresis. DNA with an absorbance ratio
139 (A260/A280) > 1.7 and with no visible degradation on agarose gel was used for subsequent library
140 preparation. Genomic DNA and library adapters were quantified with QUBIT dsDNA HS assay kit
141 (Thermo Fisher Scientific/Invitrogen, Grand Island, NY). GBS libraries and adapters were prepared
142 following the protocol of Elshire *et al.* (2011), using *Cvi*AII (New England Biolabs, Ipswich, MA)
143 for DNA digestion and a 1:4 dilution of adapter mix (common and barcoded adapter) at a final

144 concentration of 4.5 ng per reaction. In the ligation step, we reduced the ligation buffer
145 concentration to 0.6x per reaction, instead of the suggested 1x. During the fragment enrichment
146 step, four separate PCR amplifications were performed and the different reactions were then pooled
147 for PCR purification. The presence of adapter dimers in the sequencing libraries was checked with
148 the Experion DNA analysis kit (Biorad, Berkeley, CA). Genomic libraries were sequenced in a
149 single lane of Illumina HiSeq2000 flowcell, using the 50bp cycle protocol, in the QB3 Vincent J.
150 Coates Genomics Sequencing Laboratory at the University of California, Berkeley, CA. The raw
151 sequencing reads have been deposited in the NCBI Sequence Read Archive
152 (<http://www.ncbi.nlm.nih.gov/sra>) under the accession number SRX1308469.

153 **Sequencing pre-processing, alignment and SNP calling**

154 Recently, TASSEL-GBS (Glaubitz et al. 2014), a specific algorithm for analysis and SNP-calling of
155 GBS datasets, was released. The software was specifically implemented for calling the maximum
156 number of SNPs in low coverage and highly multiplexed datasets, favoring allelic redundancy over
157 quality score (Glaubitz et al. 2014). Since our dataset contained few lines at high coverage, we
158 preferred to follow a different, more robust, and accepted pipeline for bioinformatic analysis
159 (Altmann et al 2012). In particular, we used SAMtools for SNP calling since different studies
160 indicate that it is more conservative in variant calling compared to other algorithms, also in datasets
161 obtained from reduced representation libraries (Altmann et al 2012, Greminger et al. 2014).

162 Reads were quality trimmed at the 3'-end using sickle (<https://github.com/najoshi/sickle>), keeping
163 only reads with no more than 2 Ns and a minimum length after trimming of 30bp. Then, the reads
164 that recreated the *Cvi*AI cutting site (possible chimeras, partial digestion or sequencing errors) or
165 that contained the common adapter sequence (short fragments) were trimmed and only those reads
166 longer than 30bp after this second trimming step were retained. The last filtering step kept only the
167 reads that contained, after the barcode sequence, the overhang sequence of *Cvi*AI digestion (i.e.,

168 ATG). The resulting reads were then demultiplexed using sabre (<https://github.com/najoshi/sabre>)
169 allowing one mismatch for each barcode.

170 Read alignment was performed on the *P. vulgaris* unmasked genome sequence
171 (<http://www.phytozome.net/commonbean>) using BWA (Li and Durbin 2009). After the alignment,
172 only the reads with a minimum mapping quality of 10 were used for downstream application. Base
173 call recalibration was performed with the R package (www.r-project.org) ReQON (Cabanski et al.
174 2012). After quality score recalibration, variants were called with SAMtools considering only loci
175 covered by more than 30% of the lines analyzed (6 lines). The resulting variants were filtered with
176 VCFtools (Danacek et al. 2011); only those with a Minor Allele Frequency (MAF) higher than 0.05,
177 a minimum quality more than 10, and a mean read depth, across all lines, from 5 to 1000 (--maf
178 0.05 --minQ 10 --min-meanDP 5 --max-meanDP 1000) were considered for downstream analysis.
179 SNP and InDel statistics were performed with VCFtools; SNP density and transition to transversion
180 ratio (Ts/Tv) were calculated for non-overlapping bins of 1Mb.

181 **Identification of repetitive regions and phylogenetic analysis**

182 SNPs located in repeated regions were removed with VCFtools using the annotation of *P. vulgaris*
183 repeats available in Phytozome (Goodstein et al. 2012).

184 For phylogenetic analysis, only the variants located in annotated coding DNA sequences (CDS)
185 were used, since these regions are generally subjected to higher evolutionary pressure than non-
186 coding DNA sequences. A FASTA multiple alignment file was created for subsequent phylogenetic
187 analysis by concatenating the extracted variants at each position for each genotype analyzed. During
188 the creation of the multiple alignment file, individual genotypes with a quality below 10 or missing
189 genotypes were treated as missing data. Due to the self-pollinating nature of *P. vulgaris*, the
190 heterozygous calls were also treated as missing data, since they could be sequencing or SNP calling
191 errors. The resulting multiple alignment file was then analyzed using the seaview toolkit (Gouy et

192 al. 2010). A phylogenetic tree was built using the Neighbor-Joining (NJ) clustering approach, with
193 the Kimura two-parameter (Kimura, 1980) nucleotide substitution model and 1000 bootstrap
194 replicates using the seaview toolkit (Gouy et al. 2010).

195 **CNV identification and annotation**

196 CNVs were identified using the reference genotype G19833 as baseline for identifying coverage
197 shifts, as a proxy of CNV, in the other sequenced genotypes. First, we calculated the number of
198 reads in 100Kb non-overlapping genomic bins in each genotype. Then, we normalized the read
199 counts in each bin by dividing the count by the total number of reads mapped in each genotype, and
200 calculating the relative read coverage (RRC) as a ratio between the normalized read counts of the
201 genotype of interest and the reference genotype (G19833). The RRC should be normally distributed
202 with a mean ~ 1 . For this analysis, we removed the genomic bins without mapped reads in the
203 G19833 genotype. We selected as putative CNV the genomic bins with a $RRC < 0.1$ or > 1.9 ; the
204 genes located in these genomic bins were then subjected to Gene Ontology (GO) enrichment
205 analysis using the Blast2Go tool (Conesa et al. 2005).

206

207 **Results and Discussion**

208 ***In silico* genome digestion and analysis of high-throughput sequencing raw data**

209 Comparison of *in silico* genome digestion between *Cvi*AII and *Ape*KI showed that *Cvi*AII would
210 produce more fragments suitable for sequencing but that it will require a higher sequencing
211 coverage than *Ape*KI. On the other hand, by using *Cvi*AII, we would be able to tag 97% of the
212 genes present in *P. vulgaris* genome, 30% more than when using *Ape*KI (Supplementary File S1).

213 Sequencing on a HiSeq2000 (Illumina, San Diego, CA) generated 137,026,622 50bp single-end
214 reads of which 127,384,853 (93%) passed the initial sickle quality trimming. Among these ~ 127 M

reads, 3,002,729 (2.4%) were removed because they were shorter than 30bp after the trimming of reads containing the RE recognition site or adapter contaminants, or because they did not contain the overhang RE sequence after the barcode sequence. As expected from the library preparation strategy, there was a high level of duplicated reads, with only 13,278,501 unique reads in the dataset, suggesting a mean 10x redundancy for each read tag. Nevertheless, these data suggest that the overall library quality was high and consistent with the experimental approach.

After de-multiplexing, alignment, and filtering of the low-quality aligned reads, the number of reads was almost equally distributed among the different genotypes, with > 90% of annotated genes (~25,000) being tagged by at least one read (Table 1). In particular, almost 50% of the reads in each line could be aligned to the reference genome; and 50% of the aligned reads tagged gene sequences. The total number of reads per gene in each line ranged from 36 to 84, with a mean of 52 reads per gene in each line. These results are consistent with the *in silico* digestion of *P. vulgaris* genome, and showed a homogeneous read mapping rate among wild and domesticated races belonging to different gene pools (Table 1).

Analysis of identified SNPs and InDels

A total of 77,595 SNPs and InDels were identified after keeping variants with a Minor Allele Frequency (MAF) higher than 0.05 (--maf 0.05), a minimum calling quality higher than 10 (-minQ 10) and a mean read depth per sites between 5 and 1000 (--min-meanDP 5, --max-meanDP 1000). Among the variants identified, 73,656 (95%) were SNPs, 2,088 (3%) were deletions and 1,851 (2%) were insertions. The InDels ranged from 1 to 8 bp, with the majority of them being mononucleotide insertions and deletions. Due to the repetitive nature of most plant genomes and the resulting miscalls of SNPs and InDels in repetitive regions, all the variants that were located in these regions were removed. The remaining number of variants were 47,838 (61%), divided between 44,875

239 (94%) SNPs, 1,940 (3%) deletions and 1,693 (3%) insertions. This ratio is similar to the occurrence
 240 of *Cvi*AI recognition sites in non-repetitive vs. repetitive regions of the genome, highlighting the
 241 reliability of *in silico* digestion-based approaches. In addition, the percentage of variants located in
 242 non-repetitive regions was three times higher than the variants identified by Zou et al. (2014) in
 243 common bean. For further analysis, only these non-repetitive SNPs were considered.

244 The SNPs and InDels distributions were significantly highly correlated with chromosome length
 245 ($r=0.79$, $p=0.004$) (Supplementary File S3), with a mean of ~4,328 and a median of 4,312 variants
 246 per chromosome, and a median of 79 variants per Mb. These results exceeded markedly the ones,
 247 obtained after *Ape*KI digestion, of Hart and Griffiths (2015). In particular, they found a correlation
 248 of 0.45 between SNPs density and chromosome length using the *Ape*KI restriction enzyme in
 249 common bean. The highest number of variants were observed on chromosome 2 (5,311) and the
 250 lowest on chromosome 10 (3,314). On the other hand, no significant correlation was found between
 251 mean SNP density (in 1Mb non-overlapping bins) and chromosome length ($r=-0.35$, $p=0.28$)
 252 (Supplementary File S3). The variant mean read depth for each line ranged from 5 to 12 reads per
 253 site, with a mean and median of ~8 reads for SNPs. The variant coverage, averaged across all the
 254 lines, ranged from 5 to 439, with a mean and median of 8 and 7, respectively. A plot of variant
 255 density in 1Mb non-overlapping bins closely resembled the density of annotated genes in the *P.*
 256 *vulgaris* chromosomes (Fig. 1), with a Pearson's correlation coefficient (r) of 0.89 ($p < 2.2e^{-16}$).

257 SNPs were classified into transitions (Ts) and transversions (Tv), based on the type of nucleotide
 258 substitution, using VCFtools (Supplementary File S4). The number of C/T and A/G transitions was
 259 similar (~13,000); the A/C and G/T transversions had a similar frequency, while A/T and C/G
 260 transversions were slightly higher or lower, respectively, compared to A/C and G/T transversions.
 261 The Ts/Tv ratio in our dataset was 1.56 for the SNPs localized in non-repetitive regions, slightly
 262 higher than previously reported in common beans using a RRLs approach (Zou et al. 2014).

263

264 **Characterization of SNP and InDel distribution and phylogenetic analysis**

265 The total number of SNPs and InDels per line ranged from 3,512 to 21,415, with the lower number
266 of SNPs and InDels identified in genotypes G19833 (3,512), UC0801 (5,354), CAL143 (5,479), and
267 Midas (9,033) (Table 2). All these genotypes were domesticated beans belonging to the Andean
268 gene pool, as does the genotype used for the reference sequence (G19833), which was also the one
269 with the fewest SNPs in our analysis. SNPs and InDels in Mesoamerican entries ranged from
270 17,308 (accession PI417653) to 19,664 (PI311859 or G35101). PI311859 is a domesticated bean
271 with black, shiny seed (seed weight of 28 g/100 seed), which could potentially have been subjected
272 to introgression from *P. dumosus* or *P. coccineus*. However, further research is needed to clarify the
273 status of this accession. The genotype with the highest number of variant sites was G21245, a wild
274 bean from the ancestral gene pool originating in northern Peru (Kami et al. 1995), with 21,416
275 variants detected.

276 Of the 47,838 SNPs and InDels identified, 23,273 (49%) were located in genic sequences, with
277 11,163 in CDS, 2,285 in untranslated regions (UTRs), and 9,825 in introns (Table 2). For all the
278 genotypes analyzed, 45-49% of the SNPs and InDels were located in genic sequences; among them
279 ~50% were located in CDS, ~40% in introns, and ~10% in UTRs. The 23,273 SNPs and InDels
280 located in genic sequences identified 11,027 different genes (or 40% of genes identified in the
281 whole-genome reference sequence), with an average of 2 variants per gene.

282 The phylogenetic analysis based on the identified SNPs and InDels was clearly consistent with the
283 division in different gene pools and domesticated/wild lines, and was also significantly supported
284 by high bootstrapping values (Fig. 2). The Andean and Mesoamerican gene pools were clearly
285 divided with a bootstrap support > 95. In particular, both domesticated groups of Andean and
286 Mesoamerican genotypes were strongly supported by a bootstrap value of 100, confirming the

major bottleneck that occurred during each of the two independent domestications of common bean (Bitocchi et al. 2013; Gepts 1998; Schmutz et al. 2014). In addition, the phylogenetic tree automatically was rooted with the ancestral genotype G21245 from northern Peru (Kami et al. 1995). Overall, the phylogenetic analysis of the variants identified using GBS with *CviAII* correctly identified genetic relationships among the accessions included in this study, and the level of genetic diversity of the respective gene pools based on previous information about this species (Bitocchi et al. 2013; Gepts 1998; Kwak and Gepts 2009; Schmutz et al. 2014).

CNV identification and annotation

CviAII, having a 4bp recognition sites, is a frequent-cutting enzyme and shows a diffuse read coverage across the genome (Supplementary File S5). Thus, this enzyme could be suitable for identifying CNVs across different genotypes with GBS, and could also represent a cost-effective approach for identifying this kind of variation in different bean genotypes. Indeed, CNVs are extremely important in plant genome evolution, but also affect plant phenotypes and resistance to both biotic and abiotic stresses (Żmieńko et al. 2014). The approach used in our study showed a RRC normally distributed, with a mean approximately equal to 1 (Supplementary File S6), suggestive of the reliability of this approach for the identification of CNV in common bean. Analysis of RRC showed 162 genomic bins, containing 343 genes, which could contain potential CNVs in the genotypes analyzed, with some of them shared across different genotypes (Supplementary File S7). GO enrichment analysis of these genes highlight a significant enrichment in genes involved in the apoptotic process, innate immune response, transmembrane signaling receptor activity, signal transduction, ATP binding and protein binding (Fig. 3). A large number of these genes are annotated as Leucine-rich repeat proteins and transmembrane kinases, NB-ARC domain-containing disease resistance protein, TIR-NBS-LRR class proteins, and cysteine-rich

311 receptor-like kinases (Supplementary File S8). These observations suggest that the majority of
312 putative CNVs segments identified in these genotypes contain genes involved in biotic stress
313 response. This result is in agreement with previous studies in several plants that identify regions
314 harboring CNVs as enriched in biotic stress-response genes (Cook et al. 2012; deBolt 2010;
315 McHale et al. 2012; Żmieńko et al. 2014), further highlighting the feasibility of CNVs identification
316 using GBS with a frequent-cutting enzyme.

317

318 **Conclusions**

319 GBS is a simple, cost-effective, and highly multiplexed protocol for plant genotyping using NGS
320 technologies. Using this protocol, we were able to identify 47,838 variants in 18 wild and
321 domesticated bean genotypes. Even though the use of a frequent-cutting, methylation-insensitive
322 enzyme will require a higher genome sequencing coverage, the small genome size of common bean
323 and the results presented in this study clearly show the advantages of using *CviAII* for GBS in
324 common bean. We identified thousands of evenly spaced markers across the entire common bean
325 genome, with a high density that closely resembles genes distribution. This high density could help
326 in narrowing QTL regions in mapping experiments, and facilitating a more precise location of
327 recombination events. In addition, 50% of the variants identified lay in genic sequences, while the
328 others were situated in the non-coding part of the genome. The variants in genic sequences reliably
329 identified known phylogenetic subdivisions in common bean. They could also be useful in Genome
330 Wide Association Studies (GWAS) for identifying candidate genes responsible for traits of interest.
331 On the other hand, the variants in the non-coding parts of the genome could be useful - as
332 predominantly neutral markers - for ecological studies in this species, in particular for population
333 modeling and for inferring demographic history in wild common bean. Our approach also allowed
334 us to identify several putative CNVs that could be involved in pathogen response and resistance in

335 different common bean genotypes. Last but not least, the increased throughput and reduced cost of
336 sequencing technology will soon leverage the cost and depth of sequencing required when using
337 GBS with different restriction enzymes such as 4bp-recognizing, methylation-insensitive enzymes,
338 especially for plants with small genomes like common bean.

339

340

341

342

343

344

345

346

347 **References**

348 Altmann A, Weber P, Bader D, Preuss M, Binder EB, Müller-Myhsok B (2012) A beginners guide
349 to SNP calling from high-throughput DNA-sequencing data. *Hum Genet* 131:1451-1454

350 Altshuler D, Pollare VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Landes ES (2000) An
351 SNP map of the human genome generated by reduced representation shotgun sequencing.
352 *Nature* 407:513-516.

353 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA,
354 Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers.
355 *PLoS One* 3:e3376

356 Beebe S, Ramirez J, Jarvis A, Rao MI, Mosquera G, Bueno JM, Blair MW (2011) Genetic
357 improvement of common beans and the challenges of climate change. In: Yadav SS, Redden RJ,
358 Hatfield JL, Lotze-Campen H, Hall AE (ed) *Crop adaption to climate change*. Wiley-Blackwell,

359 Oxford, pp 356-369
 360 Beissinger TM, Hirsch CN, Sekhon RS, Foester JM, Johnson JM, Muttoni G, Vaillancourt B, Buell
 361 CR, Kaeppler SM, de Leon N (2013) Marker density and read depth for genotyping populations
 362 using genotyping-by-sequencing. *Genetics* 193:1073-1081
 363 Bitocchi E, Bellucci E, Giardini A, Rau D, Rodriguez M, Biagetti E, Santilocchi R, Spagnoletti
 364 Zeuli P, Gioia T, Logozzo G, Attene G, Nanni L, Papa R (2013). Molecular analysis of the
 365 parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the
 366 Andes. *New Phytol* 197:300-313.
 367 Blair MW, Diaz LM, Buendia HF, Duque MC (2009) Genetic diversity, seed size associations and
 368 population structure of a core collection of common beans (*Phaseolus vulgaris* L). *Theor Appl*
 369 *Genet* 119:955-972
 370 Broughton WJ, Hernandez G, Blair M, Beebe S, Gepts P, Vanderleyden J (2003) Beans (*Phaseolus*
 371 spp.)-model food legumes. *Plant Soil* 252:55-128
 372 Cabanski CR Cavin K, Bizon C, Wilkerson MD Parker JS Wilhelmsen, Perou CM, Marron JS,
 373 Hayes DN (2012) ReQON: a Bioconductor package for recalibrating quality scores from next-
 374 generation sequencing data. *BMC Bioinformatics*, 13:221
 375 Chacón, S., M.I., B. Pickersgill, D.G. Debouck, and J.S. Arias. 2007. Phylogeographic analysis of
 376 the chloroplast DNA variation in wild common bean (*Phaseolus vulgaris* L.) in the Americas.
 377 *Plant Syst. Evol.* 266:175-195.
 378 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
 379 Wilczyński B, de Hoon MJL (2009) Biopython: freely available Python tools for computational
 380 molecular biology and bioinformatics. *Bioinformatics* 25:1422-1423.
 381 Conesa A, Götz S, García-Gómez JM, et al (2005) Blast2GO: a universal tool for annotation,
 382 visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676.

383 Cook DE, Lee TG, Guo X, et al (2012) Copy number variation of multiple genes at Rhg1 mediates
 384 nematode resistance in soybean. *Science* 338:1206–1209.

385 Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide
 386 genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*
 387 12:499-510

388 De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG (2013) Genotyping-by-sequencing
 389 (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation
 390 sequencing. *PLoS One* 8:e62137

391 DeBolt S (2010) Copy Number Variation Shapes Genome Diversity in Arabidopsis Over
 392 Immediate Family Generational Scales. *Genome Biol Evol* 2:441–453.

393 Descham S Campbell MA (2010) Utilization of next-generation sequencing platforms in plant
 394 genomics and genetic variants discovery. *Mol Breed* 25:553-570

395 Elshire RJ, Glaubitz JC, Sun Q, Polanf JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A
 396 robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*
 397 6:e19379.

398 Ender M Terpstra K, Kelly JD (2008) Marker assisted selection for white mold resistance in
 399 common bean. *Mol Breed* 2:149-157

400 Freytag GF, Debouck DG (2002) Taxonomy, distribution, and ecology of the genus *Phaseolus*
 401 (*Leguminosae-Papilionoideae*) in North America, Mexico and Central America. *BRIT*

402 Gepts P (1998) Origin and evolution of common bean: past events and recent trends. *HortScience*
 403 33:1124-1130.

404 Gepts P (2014) Beans: Origins and Development. In: C. Smith (ed.), *Encyclopedia of Global*
 405 *Archaeology*. Springer, pp. 822-827.

406 Gepts P, Aragão F, de Barros E, Blair MW, Brondani R, Broughton W, Galasso I, Hernández G,

407 Kami J, Lariguet P, McClean P, Melotto M, Miklas P, Pauls P, Pedrosa-Harand A, Porch T,
 408 Sánchez F, Sparvoli F, Yu K (2008) Genomics of *Phaseolus* beans, a major source of dietary
 409 protein and micronutrients in the tropics. In: Moore PH, Ming R (ed) Genomics of tropical crop
 410 plants. Springer, Berlin, pp 113-143
 411 Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-
 412 GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS One 9: e90346
 413 Goodstein DM, Shu S, Howson R, et al (2012) Phytozome: a comparative platform for green plant
 414 genomics. Nucleic Acids Res 40:D1178–D1186.
 415 Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user
 416 interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221-224.
 417 Grativol C, Hemerly AS, Ferreira PCG (2012) Genetic and epigenetic regulation of stress
 418 responses in natural plant populations. Biochim Biophys Acta 1819:176–185.
 419 Greminger MP, Stölting KN, Nater A, Goossens B, Arora N, Bruggmann R, Patrignani A,
 420 Nussberger B, Sharma R, Kraus RH, Ambu LN, Singleton I, Chikhi L, van Schaik CP, Krützen
 421 M (2014) Generation of SNP datasets for orangutan population genomics using improved
 422 reduced-representation sequencing and direct comparisons of SNP calling algorithms. BMC
 423 Genomics 15: 16.
 424 Grisi MC, Blair MW, Gepts P Brondani C, Pereire PA, Brondani RP (2007) Genetic mapping of a
 425 new set of microsatellite markers in a reference common bean (*Phaseolus vulgaris*) population
 426 BAT93 x JaloEEP558. Genet Mol Res 6:691-706
 427 Hart JP, Griffiths PD (2015) Genotyping-by-Sequencing Enabled Mapping and Marker
 428 Development for the Potyvirus Resistance Allele in Common Bean. Plant Genome. doi:
 429 10.3835/plantgenome2014.09.0058
 430 Hyten DL, Song Q, Fickus EW Quigley CV, Lim JS, Choi IY, Hwang EY, Pastor-Corrales M,

431 Cregan PB (2010) High-throughput SNP discovery and assay development in common bean.
 432 BMC Genomics 11:475.

433 Iquira E, Humira S, François B (2015) Association mapping of QTLs for sclerotinia stem rot
 434 resistance in a collection of soybean plant introductions using a genotyping by sequencing
 435 (GBS) approach. BMC Plant Biol 15:5.

436 Jaganathan D, Thudi M, Kale S, et al (2015) Genotyping-by-sequencing based intra-specific
 437 genetic map refines a QTL-hotspot region for drought tolerance in chickpea. Mol Genet
 438 Genomics MGG 290:559–571.

439 Kami J, Velásquez VB, Debouck DG, Gepts P (1995) Identification of presumed ancestral DNA
 440 sequences of phaseolin in *Phaseolus vulgaris*. Proc Natl Acad Sci 92:1101–1104.

441 Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through
 442 comparative studies of nucleotide sequences. J Mole Evol 16:111-120

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

445 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.
 446 Bioinformatics 25:1754-1760.

447 Li H, Vikram P, Singh RP, et al (2015) A high density GBS map of bread wheat and its application
 448 for dissecting complex disease resistance traits. BMC Genomics 16:216.

449 Liu H, Bayer M, Druka A, Russel JR, Hackett CA, Poland J, Ramsay L, Hedley PE, Waugh R
 450 (2014) An evaluation of genotyping by sequencing (GBS) to map the *Breviarisatum-e* (ari-e)
 451 locus in cultivated barley. BMC Genomics 15:104

452 McHale LK, Haun WJ, Xu WW, et al (2012) Structural variants in the soybean genome localize to
 453 clusters of biotic stress-response genes. Plant Physiol 159:1295–1308.

454 Miklas PN, Kelly JD, Beede SE, Blair MW (2006) Common bean breeding for resistance against

455 biotic and abiotic stresses:from classical to MAS breeding. *Euphytica* 145:105-131
 456 Monson-Miller J, Sanchez-Mendez D, Fass J, Henry IM, Tai TH, Comai L (2012) Reference
 457 genome-independent assessment of mutation density using restriction enzyme-phased
 458 sequencing. *BMS Genomics* 13:72
 459 Pallotta, M. A., Warner, P., Fox, R. L., Kuchel, H., Jefferies, S. J., & Langridge, P. (2003). Marker
 460 assisted wheat breeding in the southern region of Australia. In *Proc. 10th Int. Wheat Genet.*
 461 *Symp., Paestum, Italy* (pp. 1-6).
 Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and
 genome-wide analysis of dual domestications. *Nat Genet* 46:707–713.
 462 Schnable PS, Ware D, Fulton RS, et al (2009) The B73 maize genome: complexity, diversity, and
 463 dynamics. *Science* 326:1112–1115.
 464 Schmutz J, McClean PE, Mamidi S, We GA, Cannon SB et al. (2014) A reference genome for
 465 common bean and genome-wide analysis of dual domestications. *Nat Genet.* 46:707-713
 466 Singh SP, Gepts P, Debouck DG (1991) Races of common bean (*Phaseolus vulgaris* L., Fabaceae).
 467 *Econ Bot* 45:379-396
 468 Stapley J, Reger J, Feulner PG, smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman
 469 AP, Slate J (2010) Adaptation genomics: the next generation. *Trends Ecol Evol* 25:705-712
 470 Talukder ZI, Anderson E, Miklas PN, Blair MW, Osorno J, Dilawari M, Hossain KG (2010)
 471 Genetic diversity and selection of genotypes to enhance Zn and Fe content in common bean.
 472 *Can J Plant Science* 90:49-60
 473 Thudi M, Li Y, Jackson SA, May GD, Varshney RK (2012) Current state-of-art of sequencing
 474 technologies for plant genomics research. *Brief Funct Genomics* 11:3-11
 475 Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics:
 476 applying genome sequencing technologies to crop breeding. *PLoS Biol* 12:e1001883

477 Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M (2014) Copy number polymorphism in plant
478 genomes. TAG Theor Appl Genet Theor Angew Genet 127:1–18.

479 Zou X, Shi S, Austin RS, Merico D, Munholland S, Marsolaris F, Navabi A, Crosby WL, Pauls KP,
480 Yu K, Cui Y (2014) Genome-wide single nucleotide polymorphism and insertion-deletion
481 discovery through next-generation sequencing of reduced representation libraries in common
482 bean. Mol Breed 33:769-778

483 **Table 1** Distribution of de-multiplexed reads among different individuals.

Genotype	Total reads	Aligned reads*	Aligned reads (%)	Reads aligned to gene sequences	Tagged genes
G21245	8,742,974	4,244,092	48.54	1,813,606	25,299
CAL143	9,421,387	4,829,345	51.26	1,834,224	25,625
G19833	4,905,688	2,570,710	52.40	951,376	25,357
UC0801	7,642,501	3,886,102	50.89	1,467,374	25,419
Midas	5,096,265	2,469,232	48.45	953,307	25,114
PI417653	4,791,423	2,402,640	50.14	995,149	25,147
PI319441	4,423,056	2,172,861	49.13	926,544	25,113
PI343950	8,545,592	4,022,666	47.07	1,693,329	25,494
G12873	5,577,279	2,505,178	44.92	1,040,117	25,010
SEA5	8,044,529	3,724,263	46.29	1,500,884	25,255
Pinto San Rafael	8,533,643	4,053,508	47.50	1,631,999	25,380
Flor de Mayo	5,748,661	2,621,742	45.61	1,063,173	25,077
SER118	6,108,084	2,834,653	46.41	1,123,882	25,199
Matterhorn	4,938,106	2,397,027	48.54	939,353	25,047
UCD9634	11,235,426	5,389,721	47.97	2,141,599	25,434
L88-63	7,657,785	3,633,907	47.45	1,466,989	25,360
Victor	5,591,787	2,624,902	46.94	1,050,803	25,087
PI311859	7,212,192	3,399,587	47.17	1,396,867	25,266

*Only reads with a mapping quality (Q) higher than 10.

484
485
486
487
488
489
490

Table 2 SNPs and InDels distributions among different genotypes and genomic features.

Genotype	Total SNPs	Genic*	Tagged genes**	CDSs	Introns	UTRs
G21245	21,416	10,327	6,574	4,899	4,404	1,024
CAL143	5,479	2,618	1,769	1,477	897	244
G19833	3,512	1,578	1,308	836	604	138
UC0801	5,354	2,464	1,744	1,300	928	236
Midas	9,033	4,196	2,860	2,167	1,618	411
PI417653	17,308	8,515	5,516	4,128	3,542	845
PI319441	17,741	8,737	5,706	4,240	3,677	820
PI343950	18,955	9,251	5,932	4,455	3,912	884
G12873	18,799	9,102	5,928	4,400	3,796	906
SEA5	18,532	8,929	5,660	4,354	3,693	882
Pinto San Rafael	18,586	8,924	5,638	4,371	3,706	847
Flor de Mayo	18,782	9,029	5,733	4,414	3,728	887
SER118	18,047	8,835	5,579	4,277	3,690	868
Matterhorn	17,525	8,553	5,532	4,165	3,566	822
UCD9634	18,570	9,025	5,718	4,424	3,721	880
L88-63	18,550	8,946	5,698	4,361	3,689	896
Victor	18,712	9,021	5,763	4,382	3,762	877
PI311859	19,664	9,531	5,941	4,603	3,980	948
All Genotypes	47,838	23,273	11,027	11,163	9,825	2,285

*SNPs and InDels located in genic loci. ** Genes identified by at least one SNPs or InDels

492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513

514

515 **Figure Legends**

516

517 **Fig. 1** Distribution of variants and genes with the relative density in 1Mb non-overlapping bins in
518 the 11 *P. vulgaris* chromosomes.

519

520 **Fig. 2** Neighbor-Joining (NJ) phylogenetic tree based on variants located in genic sequences of the
521 different bean lines. Bootstrap values and gene pools of the different lines are shown. PhI: Ancestral
522 wild; DA: Domesticated Andean; WM: Wild Mesoamerican; DM: Domesticated Mesoamerican.

523

524 **Fig. 3** Significant GO terms ($FDR < 0.05$) enriched in the genes located in putative CNVs. Test Set
525 is the set of the up-regulated genes, Reference Set is the background of the *P. vulgaris* GO terms
526 mapping.

527

528 **Supplementary material**

529

530 **Supplementary File S1** Comparison *P. vulgaris* genome *in silico* digestion and distribution of
531 fragment suitable for sequencing between *Cvi*AI and *Ape*KI. The number of genes tagged by the
532 fragments produced by the two restriction enzymes is shown.

533 **Supplementary File S2** Bean genotypes analyzed in this study with the barcode used for
534 multiplexed sequencing

535 **Supplementary File S3** Correlation between SNP distribution (Total SNPs) and density on a 1Mb
536 non-overlapping bin (SNPs/Mb) with chromosome length. Regression lines and Pearson regression
537 coefficient (r) are shown.

538 **Supplementary File S4** Transition and Transversion counts for the identified SNPs

539 **Supplementary File S5** Read coverage in 1Mb non-overlapping bins across the 11 chromosomes
540 for the G19833 reference genotype.

541 **Supplementary File S6** RRC in the analyzed genotypes.

542 **Supplementary File S7** Regions harboring putative CNVs in the different genotypes. The
543 coordinates of the genomic bins in the different chromosomes are reported in BED format.

544 **Supplementary File S8** Annotation, together with the best *Arabidopsis* hit, of the genes located in
545 putative CNVs. When available the best *Arabidopsis* hit common name is used.

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

[Click here to access/download](#)
Supplementary Material
S1.pdf

Click here to access/download
Supplementary Material
S2.pdf

Click here to access/download
Supplementary Material
S3.pdf

[Click here to access/download](#)
Supplementary Material
S4.pdf

Click here to access/download
Supplementary Material
S5.pdf

Click here to access/download
Supplementary Material
S6.pdf

Click here to access/download
Supplementary Material
S7.pdf

[Click here to access/download](#)
Supplementary Material
S8.pdf