Mol Breeding (2016) 36:6
DOI 10.1007/s11032-015-0431-1

Optimization of genotyping by sequencing (GBS) data in common bean (*Phaseolus vulgaris* L.)

Stephan Schröder · Sujan Mamidi · Rian Lee · Michael R. McKain · Phillip E. McClean · Juan M. Osorno

Received: 10 June 2015 / Accepted: 29 December 2015 / Published online: 5 January 2016

Abstract Genotyping by sequencing (GBS) is a technique to discover large numbers of single nucleotide polymorphisms (SNPs) within a sample pool. The standard version of the GBS method uses a pool of relatively large DNA fragments and is typically sequenced at low coverage (1×). This often results in mis-scoring of heterozygotes as homozygotes and a high rate (≥30 %) of missing data points. The purpose of this study was to improve the quality and the coverage of GBS data in common bean (*Phaseolus vulgaris* L.) in order to increase the number of SNPs available for genome-wide association studies (GWAS). An improved and *Phaseolus*-specific GBS method was developed, which utilizes an *in silico* digest of the bean genome to predict the best fragment density and length, a double digest with the restriction enzymes *Mse*I and *Taq*αI, and size selection after library preparation to achieve a reduced fragment pool that can be sequenced at higher coverage. The study consisted of 25 diverse common bean genotypes belonging to the Mesoamerican gene pool and compared libraries using *Ape*KI fragments with *Mse*I/*Taq*αI double-digest fragments. The new improved bean-specific GBS library provided a 3.8- to 12.5-fold increase in SNPs, based on a minimum coverage (3×, 5× and 8×). These results provide insight for future GBS library constructs, and how to achieve a higher SNP density for GWAS studies.

Keywords Genotyping by sequencing (GBS) · *In silico* digest · Single nucleotide polymorphism (SNP) · *Phaseolus vulgaris* L

Introduction

With consumer sales of $2 billion in the USA from 2006 to 2008 (ers.usda.gov), common bean is the most important edible legume in the Americas, Africa and Europe (Osorno and McClean 2014). Beans are rich in protein and fiber, and an excellent source of minerals, such as potassium and iron, and vitamins like thiamine, vitamin B_6, and folate (Garden-Robinson and McNeal 2013; Bennink and Rondini 2008). Considering the importance of dry edible beans for the human diet, as well as their economic impact, the development of genomic resources, such as high-resolution

Electronic supplementary material The online version of this article (doi:10.1007/s11032-015-0431-1) contains supplementary material, which is available to authorized users.

S. Schröder (✉) · S. Mamidi · R. Lee · P. E. McClean · J. M. Osorno
Department of Plant Sciences, North Dakota State University, 1360 Albrecht Blvd. - Loftsgard Hall 474-F, NDSU Dept. 7670, P.O. Box 6050, Fargo, ND 58108-6050, USA
e-mail: stephan.schroder@ndsu.edu

M. R. McKain
Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, MO 63132, USA

genetic maps and markers, plays a pivotal role in dry bean breeding to help breeders to improve their germplasm.

In recent years, there have been vast improvements in dry bean genetics, such as the development of the customized 6000-Golden Gate iSelect BeadChip panel (Hyten et al. 2010), which drastically improved mapping quality, providing more than 5000 SNPs. Moghaddam et al. (2014) developed 2687 market class specific InDel (Insertion/Deletion) markers, distributed across the genome, which can be used for mapping, as well as for phylogeny. In addition, the release of the *Phaseolus vulgaris* genome by Schmutz et al. (2014) significantly helps to facilitate genomics research.

Another helpful tool to improve genetic analyses is GBS (Elshire et al. 2011), which is based on next-generation sequencing (NGS), that captures SNP data using a reduced-representation library (RRL). It has become an important tool to analyze genomes and generally provides improved genomic data in terms of marker distribution and density. However, the quality of the GBS results depends on two main factors: genome size and library preparation (Hamblin and Rabbi 2014). Preparing a GBS library consists of a digestion step, adapter annealing, PCR amplification, and sample-pooling. The pooled samples are then sequenced using the Illumina platform, whereby a specific fraction, rather than the entire genome, is sequenced.

The original GBS methodology by Elshire et al. (2011) uses a single restriction enzyme (e.g. *Ape*KI) to digest DNA samples. Then, the barcoded adapter is ligated to one side, and a common adapter to other side of the restricted DNA. However, since both cut-sites are identical, 25 % of the fragments will have common adapters on both sides, and another 25 % of the fragments will have two barcoded adapters on both sides. In both cases, bridge amplification in the Illumina flow cell is not possible (Illumina.com) and results in a loss of data. The use of two enzymes and Y-adapters as a common adapter, as described by Poland et al. (2012), helps to prevent these issues. Additionally, the enzyme choice is highly important, since it influences the DNA fragment size and the number of fragments represented in the GBS library. For example, a frequent cutter produces many small DNA fragments, resulting in a GBS library with a low coverage per read. Ideal fragment sizes ranges from 150 bp to 300 bp for single-end reads, and from 250 to 500 bp for paired-end reads (support.illumina.com; Hamblin and Rabbi 2014).

Thanks to the availability of the *P. vulgaris* genome sequence (Schmutz et al. 2014), GBS optimizations such as an *in silico* digest analyses can facilitate the optimization of DNA fragment size distribution for the library. The chromosome scale assembly of the common bean genome is 521 Mb and consists of 41 % repetitive DNA. The gene models are organized in gene islands and can also be found in heterochromatic regions (Schmutz et al. 2014). For a uniform distribution of markers, non-methylation-sensitive enzymes are favorable to use for a *Phaseolus* GBS library. Methylation-sensitive enzymes, in contrast, as used by Huang et al. (2014) for studies on green foxtail [*Setaria viridis* (L.) P. Beauv.], using a *Pst*I and *Msp*I double digest, generated 39,416 SNPs from 252 genotypes. This SNP abundance increased coverage in particular loci, mainly in non-methylated regions, for the purpose of a better detection of heterozygotes. Increased coverage in particular loci was also the reason to develop a double-digest library to study blackcurrant (*Ribes nigrum* L.; Russell et al. 2014). While the use of methylation-sensitive enzymes would lose information about the gene islands existing in *Phaseolus* in methylated regions of the chromosomes, it would increase coverage at certain loci in the non-methylated areas of the chromosomes. Sonah et al. (2013) used single digests of *Ape*KI, *Msp*I, and *Pst*I for *in silico* digestion of the soybean [*Glycine max* (L.) Merr.] and found *Ape*KI the most suitable candidate for GBS library construction and Illumina sequencing, producing 800,000 DNA fragments between 100 and 400 bp, and 10,120 SNPs from eight genotypes. In common bean, Hart and Griffiths (2015) also made a comparison of digestion enzymes, such as *Ape*KI and *Pst*I, and used eight different adapter concentrations for each enzyme in order to optimize GBS results. This approach resulted in 7530 high-quality SNPs, after imputation and selection for minor allele frequency of $\geq$0.05, from a 96-plex *Ape*KI GBS library, which they found superior to the *Pst*I library. For this study a RIL population of 84 lines and 12 parental checks were used.

The objective of this study was to improve GBS quality and SNP density for dry edible bean libraries by comparing various library preparation methods.

Materials and methods

A priori genome analysis

The following factors were considered in order to optimize the GBS protocol for dry beans: (1) the genome size and structure of *P. vulgaris*; (2) DNA methylation and restriction sites; (3) restriction fragment size selection; (4) the Illumina sequencing method (HiSeq 200 rapid single-end run); and (5) number of samples.

Due to the *Phaseolus* genome structure, which contains gene islands within the heterochromatic regions (Schmutz et al. 2014) a uniform distribution of markers within both, the euchromatic and heterochromatic regions is preferable for the purpose of mapping. To optimize coverage, the sequencing method as well as the number of DNA fragments going into the GBS library plays an important role. A HiSeq 200 rapid single-end run has an approximate output of 130 million 200-bp reads (Illumina.com). Commonly, 96 samples are run at a time, providing more than 1.3 million reads per sample. Depending on the number of DNA fragments used for sequencing, the theoretical coverage can be calculated by reads per sample divided by the number of fragments. Also of benefit are double-digest libraries, constructed with a Y-adapter (Poland et al. 2012), to prevent data loss caused by unassigned reads.

In order to evaluate different enzymes and enzyme combinations and to better estimate how certain combinations may benefit GBS library construction, the reference genome was digested *in silico*, using an in-house software (https://github.com/mrmckain/REDFreq; Table 1; Fig. 1a–d). After digesting the *P. vulgaris* genome (phytozome.net) *in silico* with five enzyme combinations (Table 1), *Taq*αI and *Mse*I double digestion appeared to be the best fit for library construction, considering the five criteria described above. Neither of these enzymes are methylation sensitive. The combination of *Taq*αI and *Mse*I provides coverage across all the chromosomes, including the heterochromatic regions, and an optimized amount of fragments between 300 and 800 bp (Fig. 1) for bridge amplification.

DNA extraction and GBS library preparation

Twenty-five dry bean genotypes (supplemental table S1) out of a pool of 96 samples were chosen according to DNA quality (260/280 nm absorbance ratio >1.8) and digested with *Mse*I/*Taq*αI to develop a GBS library. A second library was developed using DNA digest with *Ape*KI. All samples are part of the Mesoamerican Diversity Panel or MDP (www.beancap.org). The plants were grown in the greenhouse at 22 °C and additional light (600 W high-pressure sodium lamps) from 6:00 am to 8:00 pm, to the first trifoliate leaf stage until sampling. High molecular weight DNA of each individual was extracted from young leaves using a CTAB protocol (Doyle and Doyle 1987).

Both GBS libraries (*Ape*KI and *Mse*I/*Taq*αI) were constructed based on a modified protocol of Poland et al. (2012). The only differences in library construction were the enzymes themselves, and the corresponding adapters for ligation. Both libraries were size selected for ideal bridge amplification. After ligation, DNA fragments <300 bp were removed from individual samples using 0.7 volumes of Sera-Mag^{TM} Magnetic SpeedBeads prepared according to Rohland and Reich (2012). Individual samples (4 µl of sample solution) were checked via PCR, using a 34 s extension time and visualized on a 3 % agarose gel to estimate product size range and quantity. Barcoded samples were pooled and used for library construction. The PCR extension time during the library preparation step was limited to 17 s to reduce amplification of DNA fragments larger than 800 bp. The pools were sequenced by the HudsonAlpha Genome Sequencing Center, Huntsville, AL, USA, as 200 bp single-end reads on one lane of an Illumina Hi-Seq 2500 using the high-output run mode or two lanes (on-board clustering) of a HiSeq using the rapid run mode, respectively.

Data processing and handling

The read quality was checked with FastQC 0.11.2 (bioinformatics.babraham.ac.uk). Raw fastq reads of all accessions were split into separate fastq files, based on their barcodes, using either an in-house barcode splitter (for *Mse*I/*Taq*αI) or Stacks 1.30 (for *Ape*KI) (Catchen et al. 2013). All reads were trimmed to 190 bp at the 3′ end based on the Phred scale quality scores >20. The trimmed sequences were aligned to the non-masked (repetitive sequences not masked with "N's") reference genome of *P. vulgaris* (phytozome.net) using bowtie2 (Langmead and Salzberg 2012). SNPs were called using VarScan (Koboldt et al.

Table 1 Enzymes and enzyme combinations used for *in silico* digestion of the *P. vulgaris* L. genome

Enzyme/enzyme combination	Methylation sensitivity	Genome mask	Note	Number of fragments for bridge amplification
*Ape*KI	Some CpG methylation blocked (Elshire et al. 2011)	Unmasked	Used as baseline	185,000
*Ape*KI	Some CpG methylation blocked (Elshire et al. 2011)	Hardmasked	Size selected	60,000
*Pst*I/*Msp*I	Methylation sensitive/partial sensitivity to methylation (Matthes et al. 2001)	Hardmasked	Size selected	20,000
*Pst*I/*Ape*KI	Methylation sensitive (Matthes et al. 2001)/some CpG methylation blocked (Elshire et al. 2011)	Hardmasked	Size selected	12,000
*Msp*I/*Mse*I	Partial sensitivity to methylation/not sensitive (Matthes et al. 2001)	Hardmasked	Size selected	19,000
*Mse*I/*Taq*αI	Not sensitive (Matthes et al. 2001)	Unmasked	Size selected	35,000
*Mse*I/*Taq*αI	Not sensitive (Matthes et al. 2001)	Hardmasked	To estimate fragments in heterochromatic regions	27,000

2012) (supplemental figure S1). Several filters were applied to minimize the number of false positives SNPs using VCFtools 0.1.12b (Danecek et al. 2011): Only those SNPs with all of the following characteristics were retained for analysis: (1) with missing data less than 50 %; (2) with only one alternative allele; (3) a minor allele frequency of more than 5 %; (4) mapped to one of the 11 pseudo-chromosomes; (5) with Phred scale mapping quality greater than 25; and (6) total read depth >100×.

Individual genotypes were regarded as low quality if individual read depth was smaller than 3×, 5× and 8×, respectively. In order to make both runs comparable (total number of SNPs, average and maximum SNP distance), *Ape*KI and *Mse*I/*Taq*αI HiSeq runs were normalized to an average of 1,000,000 reads per sample after mapping to the reference genome, to exclude sequencing effects, such as number of reads.

Results

This study describes an optimized GBS method, using *in silico* digestion of the *Phaseolus* genome for fragment size optimization. This analysis compared both single-enzyme (Elshire et al. 2011) and double-enzyme digests (Poland et al. 2012).

The *Mse*I/*Taq*αI enzyme combination covers both, the euchromatic and the heterochromatic chromosome regions, and creates approximately 35,000 fragments in the desired length range from 300 to 800 bp. In comparison, a GBS library constructed using *Ape*KI, which is widely used for GBS, produced more than 60,000 fragments, even after size selecting for DNA fragments >300 bp. This would result in a theoretical coverage of approximately 20×, considering 1.3 million reads per sample. Unfortunately, the number of DNA fragments going into the *Ape*KI GBS library is hard to predict, due to the partial methylation sensitivity of the enzyme. For this reason, the unmasked *P. vulgaris* genome was also used for *in silico* digestion. The unmasked genome digest with *Ape*KI showed 185,000 DNA fragments in the same size range, lowering the theoretical coverage drastically. Due to the increased number of DNA fragments, the resulting theoretical coverage is therefore less than half of a *Taq*αI/*Mse*I GBS library, since more fragments have to be covered by only a limited amount of reads.

The use of Y-adapter ensures that all sequenced fragments will be flanked by one barcoded adapter and one common adapter. The fragments are also size selected to narrow the fragment pool for sequencing. The upper size limit is determined by the length of the PCR elongation step, and the lower size limit by removal of small fragments using magnetic beads. Size selection was adjusted for genome size, and project interests (mapping), and the Illumina sequencing technique used (HiSeq 2500 rapid single-end run). The pool of size selected fragments was multiplexed, and a library was prepared for sequencing. To analyze the data obtained from sequencing, an in-house computing pipeline was used (supplemental figure S1).

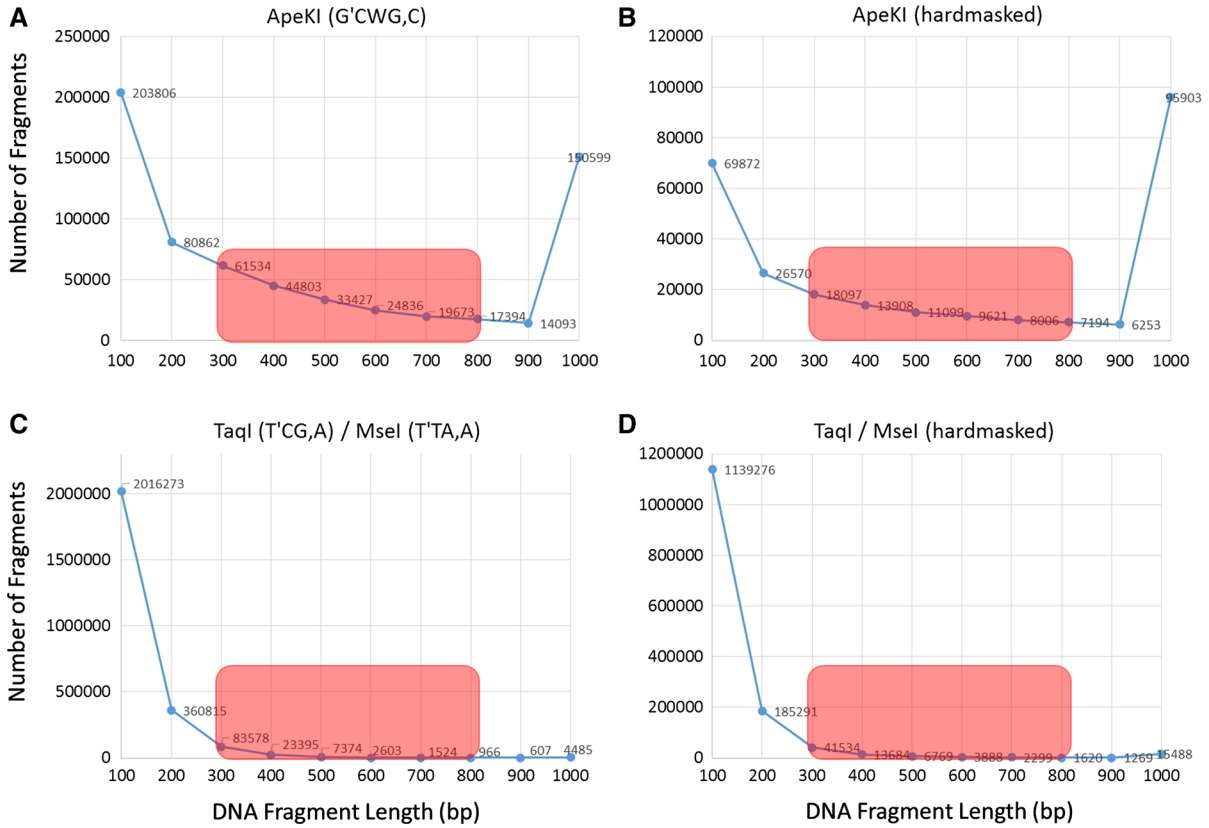

Fig. 1 *In silico* digestion of the *P. vulgaris* L. genome with *Ape*KI, using the unmasked genome (**a**), which includes all repetitive genome sequences, and (**b**) the hardmasked genome, which hides repetitive sequences by using "N's" instead and therefore simulating methylation in this region. *In silico* digestion with *Taq*αI/*Mse*I, (**c**) unmasked and (**d**) hardmasked. The *red rectangle* represents the number and length of DNA fragments for size selection. (Color figure online)

From the two GBS libraries, we obtained a total of 17,784,641 (*Ape*KI) and 63,633,785 (*Mse*I/*Taq*αI) raw reads, respectively. The read count of the *Ape*KI library varied from 23,530 to 3,473,482 reads per sample with an average of 711,385 reads per sample. The number of reads of the *Mse*I/*Taq*αI library varied from 1,837,775 to 4,998,047 reads per sample, averaging 2,545,351 reads (supplemental table S1). With an average of about 50 % of mapped reads generated by the *Ape*KI library, compared to almost 67 % mapped reads generated by the *Mse*I/*Taq*αI library, 4.76 times more reads mapped to the reference genome, using a *Mse*I/*Taq*αI GBS library compared to the *Ape*KI library. After filtering, the average coverage for the *Ape*KI library was 22.4×, and for the *Mse*I/*Taq*αI library it was 16.1× (Table 2). Despite the lower coverage of mapped reads, the *Mse*I/*Taq*αI library shows significantly more mapped read sites (247,482) compared to *Ape*KI (2444) where read mapping is distorted along the chromosomes (Fig. 2a, b).

The SNP call is strongly correlated with raw read counts as well as to the percentage of mapped reads. For 3× coverage, 6779 SNPs out of 112,513 sites (6.0 %) were kept after filtering the data generated by the *Ape*KI library, contrasting with 121,740 SNPs out of 523,605 sites (23.3 %) obtained by the *Mse*I/*Taq*αI library. After filtering the data using 5× coverage, 4080 SNPs out of 112,516 sites (3.6 %) of the *Ape*KI library, and 97,329 SNPs out of 523,602 sites (18.6 %) of the *Mse*I/*Taq*αI library were kept. In total, 937 SNPs out of 69,733 sites (1.3 %) were kept of the *Ape*KI library applying filters and 8× coverage, and 55,752 SNPs out of 360,482 sites (15.5 %) of the *Mse*I/*Taq*αI library. After normalization to an average of 1,000,000 reads per library and sample, 18,981 (3×), 11,424 (5×) and 2624 (8×) SNPs, respectively,

Table 2 Number of sites with mapped reads per sample and library construction and their corresponding average coverage. Total number of sites and average coverage per library

Genotypes (BIC-entry number)	*Ape*KI						*Taq*αI/*Mse*I					
	3×		5×		8×		3×		5×		8×	
	Number of sites	Mean depth	Number of sites	Mean depth	Number of sites	Mean depth	Number of sites	Mean depth	Number of sites	Mean depth	Number of sites	Mean depth
Bill Z (16)	220	11.90	117	19.32	54	35.19	82,575	12.89	65,421	15.02	36,615	20.40
T-39 (92)	6450	11.76	3893	13.16	879	24.91	90,099	13.11	72,299	15.18	40,963	20.63
Sedona (94)	468	10.54	244	16.48	85	36.74	80,741	10.90	63,652	12.75	36,233	17.25
Topaz (110)	6741	29.33	4064	33.68	929	61.56	102,400	14.06	83,339	16.13	48,896	21.62
Buckskin (111)	809	16.84	412	24.19	187	43.76	81,752	11.32	65,719	13.08	37,275	17.79
Medalist (133)	6710	18.74	4061	21.68	932	39.54	66,013	9.52	47,674	11.62	24,412	16.46
Navigator (134)	5449	8.99	3243	10.44	767	20.66	94,562	12.25	74,080	14.26	41,445	19.52
Beryl R (137)	6099	10.04	3674	11.67	850	23.86	90,253	11.79	71,583	13.67	41,122	18.42
Midnight (145)	6463	12.23	3907	14.07	908	27.35	82,179	11.05	65,041	12.89	36,935	17.45
UI-537 (160)	6707	16.75	4040	19.24	924	35.98	74,328	9.66	57,257	11.40	31,088	15.66
Common Pinto (161)	6206	10.08	3711	11.21	871	21.02	108,666	14.91	89,160	17.03	52,200	22.87
Common Red Mexican (162)	6727	46.42	4062	53.13	927	87.33	109,082	14.62	88,212	16.75	51,222	22.49
Kimberly (164)	54	8.17	14	21.43	14	24.57	101,410	13.04	80,665	15.03	46,116	20.16
Sawtooth (165)	352	10.93	167	18.01	86	29.69	100,544	13.18	81,628	15.11	47,893	20.14
UI-239 (170)	1930	8.46	722	11.39	199	25.56	88,162	11.15	70,117	12.93	39,809	17.46
I9365-31 (216)	6122	10.15	3657	11.49	825	22.31	77,195	12.04	61,864	13.91	34,959	18.93
Quincy (222)	13	21.77	9	30.00	6	42.00	83,265	11.36	65,664	13.29	36,940	18.20
TARS-VCI-4B (224)	24	15.63	15	22.87	10	31.40	96,844	12.19	78,050	14.02	45,762	18.60
PT7-2 (234)	367	11.44	187	18.34	74	37.73	93,917	11.30	74,384	13.13	43,111	17.57
Victor (267)	2319	7.76	941	9.93	192	24.90	104,125	14.30	84,160	16.44	49,417	21.99
Viva (278)	6741	20.41	4068	23.11	928	39.61	115,226	19.28	92,808	22.07	53,735	29.52
Harold (280)	6637	13.02	4013	14.86	900	28.58	108,066	15.49	86,205	17.75	49,413	23.94
SEA 10 (291)	1230	9.58	400	12.36	77	39.42	98,172	12.79	77,773	14.76	44,046	19.96
Maverick (299)	6272	10.59	3785	11.66	845	22.00	73,488	10.30	53,910	12.40	28,189	17.31
Stampede R	6440	11.49	3848	13.40	910	26.27	114,201	20.98	91,526	23.92	52,577	31.79
Average	3902	14.52	2290	18.68	535	34.08	92,691	12.94	73,688	14.98	42,015	20.24
Average deviation	2811	5.58	1757	6.54	392	9.92	11,379	1.89	9924	2.05	6417	2.64

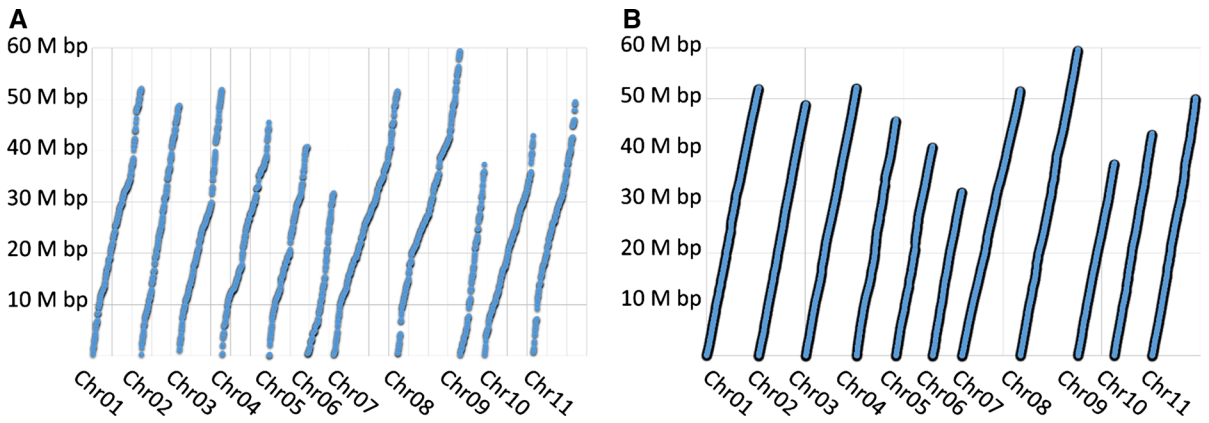

Fig. 2 Mapped read distribution across the chromosomes. The *x*-axis represents the number of tags along the chromosome, while the *y*-axis shows the tag location on the chromosome in bp. (**a**) *Ape*KI library; (**b**) *Mse*I/*Taq*αI library

were obtained from the *Ape*KI library, and 71,827 (3×), 57,424 (5×) and 32,894 (8×) from the *Mse*I/*Taq*αI library.

SNPs could be detected on average every 69,559 bp in the *Ape*KI library for 3×, and every 126,295 bp for 5× coverage. The maximum distance between adjacent SNPs was 4.5 and 6.2 Mbp, respectively. For 8× coverage, SNPs were detected averaging every 541,330 bp, and the maximum distance between two SNPs was 13.2 Mbp. SNP distribution in the *Mse*I/*Taq*αI library is far denser, detecting SNPs 4,307 bp apart from each other on average, and 487,180 bp at maximum for 3×, and 5,399 bp in average, and 610,250 bp at maximum for 5× coverage, respectively. The distance between two SNPs in the *Mse*I/*Taq*αI library, filtered for 8× coverage, was 9,449 bp, with a maximum distance between two adjacent SNPs of 1.8 Mbp.

Discussion

Based on the *in silico* digests, an increased coverage with double-digested fragments (*Mse*I/*Taq*αI) compared those generated using a single-digested fragments (*Ape*KI) was expected. However, increasing the number of DNA fragments, going into the GBS library, does not necessarily result in a decrease in detected SNPs. This is due to the fact that a calculated 20× coverage for *Ape*KI (using DNA fragments in the optimal size for bridge amplification) still can be considered useful for SNP detection. Only rare DNA fragments, which were either underrepresented in the GBS library, or which did not amplify well during bridge amplification, would be left undetected.

However, *Ape*KI's methylation sensitivity makes predictions about restriction fragment sizes hard. *In silico* digestion of the unmasked *P. vulgaris* genome (Fig. 1a; Table 1) shows a much increased number of fragments (185,000) that are ideal for bridge amplification, resulting in a decrease in mapped reads with high coverage. The mapped read distribution (Fig. 2) of the *Ape*KI library compared to the *Mse*I/*Taq*αI library shows, that both methods have a similar and uniform distribution of tags along the chromosomes. This indicates that *Ape*KI is not much affected by methylation, resulting in more restriction fragments than originally anticipated, leading to a low coverage in many parts of the genome. Still, the single-digest library also shows a distorted tag distribution, which is caused by low read counts and a low number of mapping reads compared to the *Mse*I/*Taq*αI library.

The SNP distribution was also found to be similar between the two GBS libraries. While *Ape*KI is partially methylation sensitive, it was expected to cut insufficiently in methylated DNA regions such as centromeres and telomeres. However, the distribution of SNPs is concentrated within the centromeric regions of the chromosomes, similar to the enzyme combination *Mse*I/*Taq*αI (Fig. 3a, b).

Size selection via magnetic beads is another method to improve GBS results, not only because DNA fragments have the optimum length for bridge amplification, but it also prevents tedious adapter adjustments. If *Ape*KI was not size selected, approximately 470,000 DNA fragments (Fig. 1) per sample would go

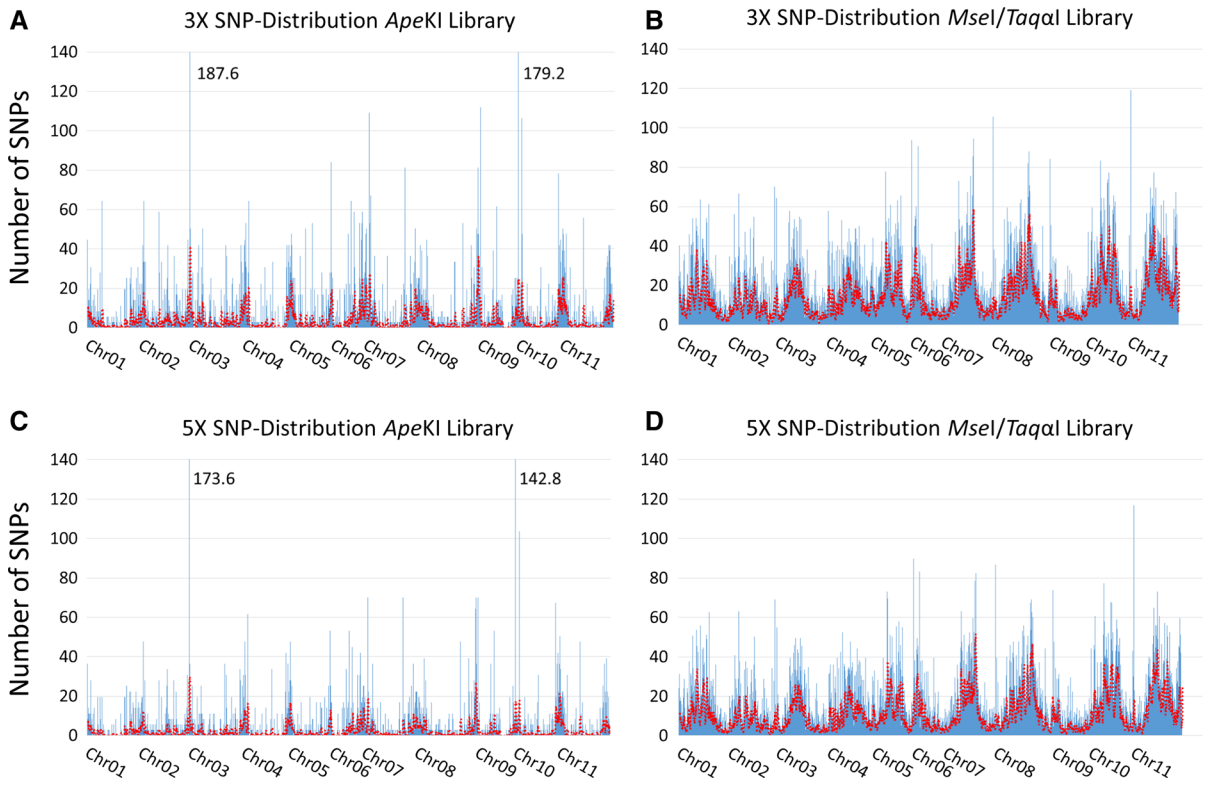

Fig. 3 SNP distribution for 3× (**a**, **b**) and 5× (**c**, **d**) coverage across the chromosomes, normalized to an average of 1 million mapped reads. *Red line* represents moving average trend line. Not shown is the 8× coverage, due to low SNP count for the *Ape*KI library. (Color figure online)

into the GBS library, resulting in a calculated average coverage of 2.8×. For *Ape*KI, this theoretical value was more than doubled, considering 185,000 fragments per sample. Size selection is even more important for creating the *Mse*I/*Taq*αI library. Here, without size selection, 2.5 million fragments per sample would be included in the library and reduce coverage to 0.52×. However, there is no study which suggests what value this theoretical coverage should be in order to achieve good GBS data. In this study, a theoretical coverage of more than 20× was considered to be sufficient. In soybean, *Ape*KI digestion produced approximately 800,000 DNA fragments for library preparation (Sonah et al. 2013). Fragment representation was further reduced using primers with an additional base or bases. The approach of optimizing the digest for an increased number of fragments ideal for bridge amplification, with a latter reduction of fragments for higher coverage, makes the size selection step redundant. However, adapter adjustments for library preparation still remain an obstacle.

Both GBS runs performed very differently during sequencing and generated different amounts of reads. Because comparing an average of about 710,000 reads, mapping to 50 % (≈356,000 of total reads mapped) for the *Ape*KI digest, to an average of 2.5 million reads, mapping to 67 % (≈1.7 million reads of total reads mapped) from the double digest is problematic, both GBS runs were normalized to 1 million reads per sample in average, to eliminate sequencing effects. After normalization and depending on coverage (3×, 5× and 8×), the GBS library generated by the double digest using the enzyme combination *Mse*I/*Taq*αI provided a 3.8–12.5 times more SNPs compared to the single digest using *Ape*KI for library preparation (Fig. 3). Hence, this study introduces an optimized GBS protocol for dry edible beans, which takes several measures such as *in silico* digest of the reference genome, the use of Y-adapters and size selection into account, to provide dense SNP coverage that is useful for QTL mapping and GWAS.

Acknowledgments This project was supported by Agriculture and Food Research Initiative Competitive Grant No. 2013-67014-21367 from the USDA National Institute of Food and Agriculture. Partial funding for this work was also received from the Northarvest Bean Growers Association and the Dry Bean Checkoff Funds.

Compliance with ethical standards

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

References

Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 21 April 2015

Bennink M, Rondini E (2008) Dry beans and human health. The Bean Institute, pp 1–31

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22(11):3124–3140

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker R, Lunter G, Marth G, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2215

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure from small quantities of fresh leaf tissues. Phytochem Bull 19:11–15

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

Garden-Robinson J, McNeal K (2013) All about beans—nutrition, health benefits, preparation and use in menus. FN1643 (Revised), The Bean Institute, NDSU Extension Service, pp 1–16

Hamblin MT, Rabbi IY (2014) The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: a study in cassava (*Manihot esculenta*). Crop Sci 54:1–6

Hart JP, Griffiths PD (2015) Genotyping-by-sequencing enabled mapping and marker development for the by-2 potyvirus resistance allele in common bean. Plant Genome 8(1):1–14

Huang P, Feldman M, Schroder S, Bahri BA, Diao X, Zhi H, Estep M, Baxter I, Devos KM, Kellogg EA (2014) Population genetics of *Setaria viridis*, a new model system. Mol Ecol 23:4912–4925

Hyten DL, Song Q, Fickus EW, Quigley CV, Lim JS, Choi IY, Hwang EY, Pastor-Corrales M, Cregan PB (2010) High-throughput SNP discovery and assay development in common bean. BMC Genom 11:475–482

Illumina.com—Technology Spotlight: Illumina Sequencing Technology http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf. Accessed 17 April 2015

Illumina.com—Support: Questions & Answers http://support.illumina.com/sequencing/sequencing_instruments/cluster_station/questions.html. Accessed 21 April 2015

Koboldt D, Zhang Q, Larson D, Shen D, McLellan M, Lin L, Miller C, Mardis E, Ding L, Wilson R (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. doi:10.1101/gr.129684.111

Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Matthes M, Singh R, Cheah S-C, Karp A (2001) Variation in oil palm (*Elaeis guineensis* Jacq.) tissue culture-derived regenerants revealed by AFLPs with methylation-sensitive enzymes. Theor Appl Genet 102:971–979

Moghaddam SM, Song Q, Mamidi S, Schmutz J, Lee R, Cregan P, Osorno JM, McClean PE (2014) Developing market class specific InDel markers from next generation sequence data in *Phaseolus vulgaris* L. Front Plant Sci 5:185

Osorno JM, McClean PE (2014) Common bean genomics and its applications in breeding programs. In: Gupta S, Nadarajan N, Gupta DS (eds) Legumes in the Omic Era, Springer, New York, pp 185–206

Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res 22:939–946

Russell J, Hackett C, Hedley P, Liu H, Milne L, Bayer M, Marshall D, Jorgensen L, Gordon S, Brennan R (2014) The use of genotyping by sequencing in blackcurrant (*Ribes nigrum*): developing high-resolution linkage maps in species without reference genome sequences. Mol Breeding 33:835–849

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Geffroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MMS, Miklas PN, Osorno JM, Rodrigues J, Thareau V, Urrea CA, Wang M, Yu Y, Zhang M, Wing RA, Cregan PB, Rokhsar DS, Jackson SA (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Brian Boyle B, Normandeau E, Laroche J, Larose S, Jean M, Belzile F (2013) An Improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE 8(1):e54603

USDA Economic Research Service (2012) Vegetables and pulses—dry beans. http://www.ers.usda.gov/topics/crops/vegetables-pulses/dry-beans.aspx. Accessed 21 April 2015