Plant Mol Biol Rep (2017) 35:61–71
DOI 10.1007/s11105-016-1003-1

ORIGINAL PAPER

Development of genomic simple sequence repeat markers in faba bean by next-generation sequencing

Mazen A. Abuzayed[1] · Mehmet Goktay[1] · Jens Allmer[1] · Sami Doganlar[1] · Anne Frary[1]

Published online: 19 August 2016

Abstract Faba bean (*Vicia faba* L.) is an important food legume crop with a huge genome. Development of genetic markers for faba bean is important to study diversity and for molecular breeding. In this study, we used Next Generation Sequencing (NGS) technology for the development of genomic simple sequence repeat (SSR) markers. A total of 14,027,500 sequence reads were obtained comprising 4,208 Mb. From these reads, 56,063 contigs were assembled (16,367 Mb) and 2138 SSRs were identified. Mono and dinucleotides were the most abundant, accounting for 57.5 % and 20.9 % of all SSR repeats, respectively. A total of 430 primer pairs were designed from contigs larger than 350 nucleotides and 50 primers pairs were tested for validation of SSR locus amplification. Nearly all (96 %) of the markers were found to produce clear amplicons and to be reproducible. Thirty-nine SSR markers were then applied to 46 faba bean accessions from worldwide origins, resulting in 161 alleles with 87.5 % polymorphism, and an average of 4.1 alleles per marker. Gene diversity (GD) of the markers ranged from 0 to 0.48 with an average of 0.27. Testing of the markers showed that they were useful in determining genetic relationships and population structure in faba bean accessions.

Keywords Illumina sequencing · Genomic SSRs · Genetic diversity · Population structure · *Vicia faba* L

Electronic supplementary material The online version of this article (doi:10.1007/s11105-016-1003-1) contains supplementary material, which is available to authorized users.

✉ Anne Frary
annefrary@iyte.edu.tr

[1] Department of Molecular Biology and Genetics, Izmir Institute of Technology, Urla Izmir 35430, Turkey

Introduction

Faba bean (*Vicia faba* L.) is a member of the Fabaceae family and is sometimes referred to as horse, broad or field bean (DUC 1997). It is the most important legume after chickpea and pea and is believed to have originated in the Near East (Torres et al. 2006). However, its origin and domestication are still debated because no wild progenitor has been found and interspecific hybridization with other *Vicia* species has failed (DUC 1997). Annual world production is 4.34 million tonnes, with Asia and Africa accounting for 72 % of production (FAOSTAT 2014). Faba bean is planted in warm-temperate and subtropical countries in the winter and in northern latitudes in the spring (DUC 1997). The crop is widely used in human food, and its seeds are consumed dried, fresh, or canned, and also as livestock feed (Torres et al. 2006). It is an important rotational crop because its roots fix atmospheric nitrogen by a symbiotic relationship with soil bacteria (Suresh et al. 2013).

Faba bean is a diploid with 2n = 2× = 12 chromosomes and has the largest known genome (13,000 Mb) among legumes (Ellwood et al. 2008). More than 38,000 accessions are found in 37 germplasm collections throughout the world and are very important sources for diversity and breeding studies (Duc et al. 2010). Such studies require appropriate molecular marker systems. Several types of molecular markers have been used to characterize and elucidate genetic diversity in faba bean, including Restriction Fragment Length Polymorphism (RFLP) markers (Van de Ven et al. 1990), Random Amplified Polymorphic DNA (RAPD) markers (Link et al. 1995), Amplified Fragment Length Polymorphism (AFLP) markers (Zeid et al. 2003; Zong et al. 2009; Gresta et al. 2010), Inter Simple Sequence Repeat (ISSR) markers (Terzopoulos and Bebeli 2008; Wang et al. 2012), Sequence-Specific Amplification Polymorphism (SSAP) markers (Ouji et al.

2012), and Single Nucleotide Polymorphism (SNP) markers (Kaur et al. 2014).

Simple sequence repeats (SSRs), which were first described by Hamada et al. (1982), are short tandem repeat motifs that occur frequently in all prokaryotic and eukaryotic genomes (Zane et al. 2002). SSRs occur in coding and non-coding regions and are more informative and variable than RAPD, RFLP and AFLP markers (Senan et al. 2014). SSRs are PCR-based, highly abundant in plant genomes, multiallelic and codominant. Because of these characteristics, SSRs are among the most used molecular markers (Akash and Mayers 2012; Zalapa et al. 2012). Moreover, SSRs are the best marker for revealing intervarietal polymorphisms and offer a wide range of applications in the preparation of genome-wide genetic and comparative maps (Senan et al. 2014).

To date, most faba bean SSR markers were developed by mining of public databases for genic SSRs, cDNA sequencing or SSR-enriched library methods (Pozarkova et al. 2002; Zeid et al. 2009; Gong et al. 2010, 2011; Ma et al. 2011; Akash and Mayers 2012; Kaur et al. 2012; Yang et al. 2012; Suresh et al. 2013; El-Rodeny et al. 2014). All of these studies resulted in the development of approximately 33,117 SSR markers with most (86 %) of these from the work of Yang et al. (2012). A total of 1133 of the identified markers were used to study genetic diversity in faba bean with 62 % (707) of the markers showing polymorphism. This number of validated SSR markers is insufficient because faba bean has a very large genome, about 26 times larger than the legume model species, *Medicago truncatula*. Therefore, the goal of this work was to develop genomic SSR markers in faba bean using Illumina sequencing technology. Next-generation sequencing (NGS) methods are high-throughput technologies that can produce millions of short sequences in parallel and are easier, faster and more economical than Sanger sequencing (Shendure and Ji 2008). NGS, along with bioinformatics tools, can be used for large-scale, rapid development of genome-wide and gene-based microsatellite loci (Abdelkrim et al. 2009). A subset of the genomic SSR markers were tested on cultivated faba bean to validate amplification efficiency and assess polymorphism. The SSR markers were also used to analyze the molecular genetic diversity and population structure of 46 accessions and cultivars from 17 countries.

Materials and Methods

Plant Material and DNA Isolation

A total of 46 faba bean accessions from 17 countries were used as plant material (Table 1). Eight accessions were from the Netherlands Gene Bank (NGB); 18 from the Centre for Genetic Resources, the Netherlands (CGN); 11 from the Aegean Agricultural Research Institute (AARI, Turkey); 5 from the University of Adelaide (Australia); and 4 from the International Center for Agricultural Research in the Dry Areas (ICARDA, Syria). Seeds were planted and grown in the growth chamber at 24–25 °C, with a 16 h photoperiod. Total DNA from the youngest leaves was extracted using CTAB extraction buffer according to Doyle and Doyle (1990). DNA quality was checked by agarose gel electrophoresis and quantification by a spectrophotometer (Multiskan GO; Thermo Scientific).

DNA Sequencing

For SSR identification, faba bean cultivar Filiz-99 was provided by AARI. Total genomic DNA was extracted using the Wizard Magnetic 96 Plant System (Promega, Madison, WI, USA) and the Beckman Coulter Biomek NX Workstation. Sequencing of genomic DNA was done by next generation sequencing (Illumina Mi-Seq Technology) by the Biotechnology Center at the University of Wisconsin-Madison, USA (https://www.biotech.wisc.edu/). This technology produced 300-nucleotide-long, paired-end reads. Raw data and further information can be found at the SRA database of NCBI (SRA id : SRP076364).

Data Pre-Processing

Adapter and linker sequences were removed from reads with cutadapt (Martin 2011) v.1.8.3 software using default settings. Any trimmed reads smaller than 50 nucleotides were removed from the dataset since they disrupt mapping and assembly processes. Reads were then mapped with Bowtie v.2.1.0 (Langmead and Salzberg 2012) software using default settings against the human genome to remove possible human contaminants. Contamination may occur during DNA extraction or next generation sequencing and possible contaminant reads were excluded from the dataset.

Sequence Assembly

ABySS v.1.3.6 (Simpson et al. 2009), a de novo, parallel, paired-end sequence assembler, was employed for sequence assembly. More than 100 runs were performed with different settings such as changing kmer (all possible substrings of length k contained in reads) and required number of reads to make a contig. Assembly quality was based on various parameters, such as the weighted median of contig lengths (N50), a commonly used measure. The best assembly was identified according to the N50 value, assembly nucleotide length (closeness to the estimated size of the *V. faba* genome), length of largest contig and contig number. The settings that were finally chosen to create contigs were: kmer ($k = 25$) and number of reads ($n = 2$).

Table 1 Faba bean accessions used in the study. Cluster assignments of 46 faba bean accessions according to Structure and DARwin analyses

No.	Sample Name	Source	Origin	Inferred ancestry subpopulation		Subpopulation assignment[a]	Cluster[b]
				1	2		
1	8642	NGB	Finland	0.887	0.113	1	A
2	1547.1	NGB	Finland	0.928	0.072	1	A
3	Mikko	NGB	Finland	0.944	0.056	1	A
4	Witkiem manida	NGB	Germany	0.911	0.089	1	A
5	Ukko	NGB	Germany	0.915	0.085	1	A
6	Kontu	NGB	Germany	0.899	0.101	1	A
7	1542.1	NGB	Finland	0.924	0.076	1	A
8	1548.2	NGB	Finland	0.901	0.099	1	A
9	7874	CGN	Spain	0.591	0.409	Admixed	A
10	15563	CGN	Syria	0.838	0.162	1	A
11	15619	CGN	Egypt	0.084	0.916	2	B
12	13485	CGN	Pakistan	0.850	0.150	1	C
13	13464	CGN	UK	0.841	0.159	1	A
14	10391	CGN	Egypt	0.140	0.860	2	B
15	7826	CGN	Greece	0.470	0.530	Admixed	B
16	7716	CGN	Italy	0.734	0.266	1	A
17	7844	CGN	Jordan	0.097	0.903	2	B
18	7781	CGN	Netherland	0.647	0.353	Admixed	A
19	15641	CGN	Netherland	0.153	0.847	2	B
20	10382	CGN	Turkey	0.486	0.514	Admixed	A
21	10371	CGN	Algeria	0.694	0.306	Admixed	B
22	18892	CGN	Netherland	0.750	0.250	1	A
23	07875	CGN	India	0.572	0.428	Admixed	C
24	10385	CGN	Turkey	0.757	0.243	1	A
25	10374	CGN	Syria	0.340	0.660	Admixed	A
26	10325	CGN	Syria	0.698	0.302	Admixed	A
27	TR23018	AARI	Turkey	0.216	0.784	2	C
28	TR31590	AARI	Turkey	0.158	0.842	2	B
29	TR33140	AARI	Turkey	0.583	0.417	Admixed	C
30	TR37255	AARI	Turkey	0.763	0.237	1	A
31	TR44876	AARI	Turkey	0.153	0.847	2	C
32	TR44928	AARI	Turkey	0.785	0.215	1	A
33	TR49380	AARI	Turkey	0.088	0.912	2	B
34	TR53748	AARI	Turkey	0.673	0.327	Admixed	A
35	TR61267	AARI	Turkey	0.713	0.287	1	A
36	Ascot	Adelaide Univ.	Australia	0.076	0.924	2	A
37	Manafest	Adelaide Univ.	Australia	0.043	0.957	2	B
38	Fiord	Adelaide Univ.	Australia	0.050	0.950	2	B
39	Fiesta	Adelaide Univ.	Australia	0.220	0.780	2	A
40	Aquadulce	Adelaide Univ.	Australia	0.128	0.872	2	C
41	Filiz-99	AARI	Turkey	0.512	0.488	Admixed	B
42	Salkım	AARI	Turkey	0.291	0.709	2	C
43	26139	ICARDA	.Colombia	0.220	0.780	2	B
44	26145	ICARDA	Egypt	0.615	0.385	Admixed	A
45	ILB938/2	ICARDA	Unknown	0.096	0.904	2	B
46	Melodie/2	ICARDA	France	0.702	0.298	1	A

[a] Accessions were assigned to subpopulations based on the proportion of inferred ancestry with a threshold of ≥0.70

[b] Cluster assignments based on the neighbor-joining dendrogram

SSR Detection, Annotation and Primer Design

Only contigs larger than 200 nucleotides were analyzed for SSR detection using our in-house tool SiSeer (http://bioinformatics.iyte.edu.tr/index.php?n=Softwares.SiSeeR). The minimum number of repeats required to identify perfect SSRs was 10 for mononucleotides, 4 for dinucleotides, and 3 for motifs comprised of three or more nucleotides. Identified SSR sequences were extracted with their genomic context (padded with 100 nucleotides) and were converted to FASTA formatted sequences. These queries were searched against the Uniprot (http://www.uniprot.org/uniprot/?query=taxonomy%3A%22Viridiplantae+%5B33090%5D%22&sort=score) non-redundant plant protein database (Taxonomy = Viridiplante) with BLASTX ve.2.2.30. Primer design was performed on contigs larger than 350 nucleotides with Primer3 (primer_core) v.2.3.6 (Koressaar and Remm 2007) using default parameters (Primer task = generic, primer optimum size = 20, primer minimum sixe = 18, primer maximum size = 24, primer product size = 100–300, primer minimum TM = 50, primer maximum TM = 60, and primer optimum TM = 55).

Validation of Genomic SSR Markers

For SSR validation and to ensure that the expected SSRs were amplified by the primers, two faba bean samples (ILB938/2 and Melodie/2) were used as template for PCR with four of the SSR markers. Amplified products were analysed using the dye-terminator sequencing method. First, PCR products were purified with the DNA Clean & Concentrator – 5 Kit (Zymo Research) and were used as template in sequencing reactions prepared using GenomeLab DTCS Quick Start Kit (Beckman Coulter). The thermal cycling conditions were 30 cycles of 96 °C for 20 s, 50 °C for 20 s, 60 °C for 4 min. ZR DNA Sequencing Clean-up Kit (Zymo Research) was used for purification of the reaction mixture for each SSR amplicon which was then resuspended in 30 μL sample loading solution (Beckman Coulter) and run on a Beckman CEQ8800 capillary electrophoresis device using the LFR-c method (injection voltage 2.0 kV for 10–15 s, separation temperature 60 °C, separation voltage 7.4 kV, separation time 45 min).

SSR Amplification

SSR amplification for each primer pair was carried out in a final volume of 20 μl and contained 30 ng DNA, 1× PCR buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 1 pmol forward and reverse primers, 1 U Taq Polymerase. PCR conditions were 95 °C for 4 min for one cycle, followed by 35 cycles of 45 s at 95 °C for denaturation, 1 min at 55 °C for annealing and 1 min at 72 C° for extension, the final extension cycle was at 72 °C for 5 min. PCR reactions were performed in a Veriti 96-Well Thermal Cycler (Applied Biosystems). PCR products were separated using capillary electrophoresis instrument (Fragment Analyzer Automated CE System; Advanced Analytical) using the DNF-900 dsDNA Reagent Kit (Advanced Analytical), and SSR alleles were visualized and scored using PROSize 2.0 software v.1.2.1.1 (Advanced Analytical).

SSR Data Analysis

SSR alleles were scored for presence (1) or absence (0). Codominant scoring was not possible because most of the markers amplified more than two fragments and alleles could not be identified. Fragments that were observed at less than 10 % (low frequency) in faba bean accessions were excluded from all analyses because such products can be unreliable. Gene diversity (Nei 1973) was calculated depending on the frequency of the allele for each SSR marker and the calculations were performed with the GDdom online computer program (http://plantmolgen.iyte.edu.tr/GDdom/) using the formula of Roldan-Ruiz et al. (2000):

$$GD_i = 2f_i(1 - f_i)$$

Where GD_i is the gene diversity of marker 'i', f_i is the frequency of the amplified allele (band presence), and 1 - f_i is the frequency of the null allele. Marker data were used to infer population structure of the 46 faba bean accessions with the Structure computer program (Pritchard et al. 2000), and models with 1 to 10 subpopulations (K) were tested for 20 iterations. Burn-in period was 100,000 and the number of Monte Carlo Markov Chain repeats was 500,000. Structure Harvester computer program (Earl and VonHolt 2012) was used to calculate ΔK values for each model based on posterior probabilities. The model with the highest ΔK was selected as the best. Inferred ancestry threshold was set as ≥0.70 to assign the accessions to subpopulations. Accessions with lower probabilities were assigned to the admixed group. To study genetic diversity, the binary presence/absence data were used to generate a dissimilarity matrix using the Dice coefficient as implemented by Darwin5 computer program (http://darwin.cirad.fr/product.php). The distance data were used to construct a dendrogram of the 46 faba bean accessions using unweighted neighbor-joining method.

Results

Sequence Assembly and Simple Sequence Repeat Identification

Illumina sequencing of faba bean Filiz-99 resulted in 14,027,500 sequence reads comprising more than 4,208 Mb. Removal of adaptor and linker from sequence reads resulted in 4182 Mb with an average cleaned sequence length of 298.2 nucleotides (nt). Any sequences larger than 200 nt were assumed to be a contig. As a result, 56,063 contigs were retrieved which encompassed 16.37 Mb, representing 0.13 % of the genome (Table S1). Contigs were mined for SSRs, resulting in the identification of 2138 SSRs. SSR length ranged between 6 and 32 nt, with an average length of 13.2 nt. Among the 2138 SSRs identified, mononucleotide repeats were the most abundant, representing 57.5 % of all SSRs. Dinucleotide repeats were the second most common type and accounted for 20.9 % of all SSRs. Trinucleotides represented 6.5 % of SSRs (Fig. 1). The most common motifs were the A/T repeat (98.9 %) for mononucleotides, and AG/CT (26.6 %) for dinucleotides (Table 2), followed by GA/TC (23.2 %). Among trinucleotides, the most frequent repeats were A/T-rich, AAT/ATT repeats were the most abundant (23.1 %) followed by ATA/TAT repeats (19.0 %).

Primer Design and SSR Validation

Of the 2138 SSRs identified in contigs, 430 met the requirements for primer design. We tested 50 of the designed primers

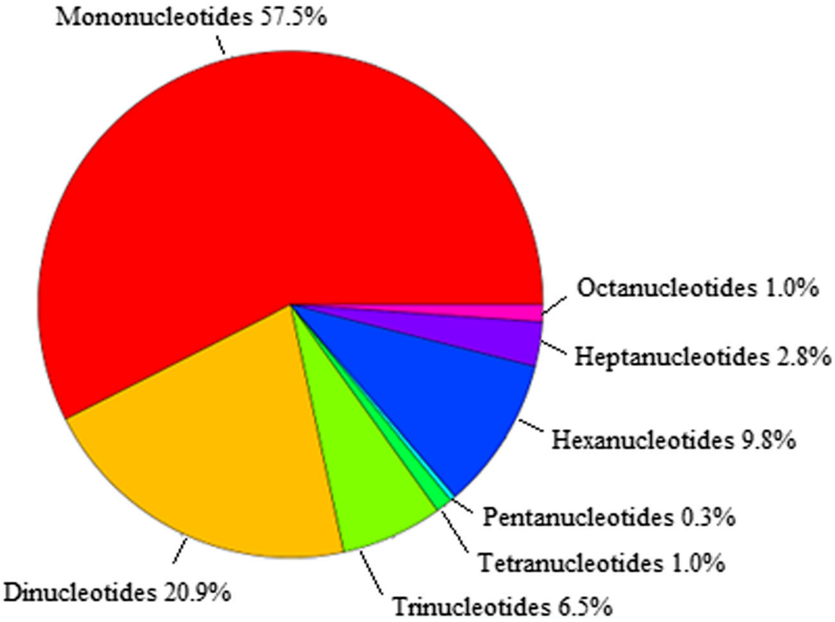

Fig. 1 Simple sequence repeat types in faba bean genome

for their amplification efficiency on two faba bean genotypes (ILB938/2 and Melodie/2). Primer sequences are found in Tables 3 and S2. Of these primers, 48 (96 %) amplified products. For SSR validation and to ensure that the expected SSRs were amplified by the primers, two faba bean samples were used for amplification of the SSR regions followed by sequencing. Genotype ILB938/2 was used as a DNA template for markers FbgSSR26 (TTGAAT), FbgSSR37 (AG), and FbgSSR309 (AG) while Melodie/2 was used as a DNA template for FbgSSR293 (ACCAAA) (Table 3). All four sequences contained the expected motifs (Supplementary Table S3), proving that these primers amplified the expected SSRs. The details of the designed SSR primers, with the corresponding SSR motif, motif lengths, repeat motif annotation, and primer sequences are available online (http://plantmolgen.iyte.edu.tr/data/).

Table 2 Most abundant simple sequence repeat (SSR) motifs[a] in faba bean genomic sequence

SSR Motif	Number of SSRs	Motif frequency (%)
A/T	1216	98.9
TA	85	21.2
AT	77	17.3
GA/TC	103	23.2
AG/CT	119	26.6
AAT/ATT	28	23.1
ATA/TAT	23	19.0
TTA/TAA	18	14.8
TTC/GAA	13	10.7
AAAT/ATTT	4	26.6
AATA/TATT	4	26.6
ATAA/TTAT	3	13.3

[a] Motif frequencies are relative to SSR type. Only motif with a frequency ≥10 % and number of SSRs >1 are listed

Gene Diversity, Population Structure and Genetic Diversity Analyses Using SSR Markers

A total of 39 genomic SSR markers which showed clear amplification were applied to 46 faba bean accessions from throughout the world. Turkey was represented by the most accessions (13) with 5 accessions each from Finland and Australia. Germany, the Netherlands, Syria and Egypt were represented by 3 accessions with the remaining 11 accessions from other countries (Table 1). Thirty-one of 39 markers (79 %) were polymorphic. The SSR primers generated 161 alleles, 141 (87.6 %) of which were polymorphic (Table 3). The average number of amplified fragments per genomic SSR marker was 4.1, with a range of 1–10 alleles. The average gene diversity value (also called polymorphism information content, PIC) of the markers (based on a calculation that ranges from 0 to 0.5) was 0.27, with the highest value calculated for FbgSSR545 (0.48 ± 0.01). The lowest value was zero for monomorphic markers (Table 3).

The SSR data were used to study the population structure and genetic diversity of the 46 faba bean accessions. Population structure assigned the accessions to two subpopulations (Table 1 and Fig. S1), the first subpopulation included 18 accessions, the second included 16 accessions while 12 accessions were assigned as admixed (Table 1). All accession from Germany and Finland were assigned to subpopulation 1. Australian accessions were assigned to subpopulation 2 while

Table 3 Simple sequence repeat (SSR) markers used for the molecular genetic analysis of faba bean

SSR primers	Sequence	Repeat motif	Number of polymorphic alleles/total alleles	Gene diversity[a] (GD)
FbgSSR11-F	GAGTGAGGACAAATCAAGGT	(TTTCTG/CAGAAA)$_3$	0/1	0
FbgSSR11-R	AGGCAAACCTCTTGTTACAA			
FbgSSR26-F	GGTTGTGTCACTTTTCTTGG	(ATTCAA/TTGAAT)$_3$	1/2	0.12 ± 0.12
FbgSSR26-R	AATAAGACCTTAACTTTATTAACC			
FbgSSR29-F	ACTTCCAAAAATTTCAGAATCTC	(AAATTG/CAATTT)$_4$	7/7	0.35 ± 0.06
FbgSSR29-R	CCCAACTGAAGAAAAGGGTA			
FbgSSR30-F	TCCAAAAATTTCAGAATCTCCA	(AAATTG/CAATTT)$_4$	9/10	0.29 ± 0.05
FbgSSR30-R	CCCAACTGAAGAAAAGGGTA			
FbgSSR37-F	ATGCACGTTACAAGACATTG	(AG/CT)$_9$	9/9	0.38 ± 0.04
FbgSSR37-R	CTTTCCTCGCAAAAGGATTG			
FbgSSR109-F	CATGTCTCCTCACCATTTCA	(ATTG/CAAT)$_5$	0/1	0
FbgSSR109-R	TGTAGCGGAACTCAAATGAA			
FbgSSR140-F	TTCAAATGTAAACAGGCGTG	(AC/GT)$_6$	7/7	0.31 ± 0.05
FbgSSR140-R	ACCGTTGAGAGTAAAAGGAA			
FbgSSR198-F	TGAGACAAATCAGCATTCCA	(AGTTTTGA/TCAAAACT)$_3$	1/1	0
FbgSSR198-R	GCATTTGCATTCACATTTGG			
FbgSSR229-F	TTCTAGAATTGGTGCTCCTG	(TC/GA)$_7$	5/5	0.39 ± 0.04
FbgSSR229-R	TGCTTGAATATTGAGAGAAGT			
FbgSSR293-F	TGAGTGGAGATCTGCTAAGA	(ACCAAA/TTTGGT)$_3$	4/4	0.40 ± 0.03
FbgSSR293-R	AGCAATTGCATTCTAAAGCC			
FbgSSR306-F	CCACTCATTACCTTGAACCA	(GA/TC)$_8$	5/7	0.20 ± 0.06
FbgSSR306-R	CAACATCATCAGAAGCAACC			
FbgSSR309-F	GAACTATGAAGAGCAGCAGT	(CT/AG)$_9$	5/5	0.42 ± 0.06
FbgSSR309-R	AGTTGTTTACATGGACGTGT			
FbgSSR319-F	CTTCCGTCTTCTTTCCGTAT	(TGCAAG/CTTGCA)$_3$	0/1	0
FbgSSR319-R	ATAACTAATAGCAGCACCGG			
FbgSSR322-F	AAGGTGGTGGTGATTCAATT	(AAAATG/CATTTT)$_3$	5/5	0.26 ± 0.07
FbgSSR322-R	ATTTTATCTTGCCCATGGGT			
FbgSSR375-F	TTCAACCGGTAAAGAGAAGG	(CTTAGG/CCTAAG)$_3$	0/2	0
FbgSSR375-R	ACCAAAACTCTGATGGTGAA			
FbgSSR382-F	TGAGAAAGTTGAGTGACTGG	(GAATTG/CAATTC)$_3$	4/4	0.30 ± 0.05
FbgSSR382-R	ACCTTTGATAAATTGGAATAGA			
FbgSSR443-F	AAAACATCAATTTTGACTCAT	(TATTTAT/ATAAATA)$_3$	2/3	0.33 ± 0.16
FbgSSR443-R	TGAAGCAAATAAAATAACAGCAAG			

Table 3 (continued)

SSR primers	Sequence	Repeat motif	Number of polymorphic alleles/total alleles	Gene diversity[a] (GD)
FbgSSR444-F	GCACCTGGCAAAATGATTTA	(AATTCTG/CAGAATT)$_3$	0/1	0
FbgSSR444-R	GCGTTTCAGCATTTTCAAAC			
FbgSSR451-F	GAACGACTTGAGAGAGAGTC	(TC/GA)$_6$	7/7	0.38 ± 0.06
FbgSSR451-R	TTTTAAACCCTAAGGACGGG			
FbgSSR518-F	AGTTCTCAAAGCGTTCTTCT	(AT/AT)$_6$	4/4	0.38 ± 0.03
FbgSSR518-R	GCTTGTATATTGTGTGAAGTCT			
FbgSSR520-F	GCTTGCAAGTAAGTGTGTTT	(AG/CT)$_6$	7/7	0.33 ± 0.04
FbgSSR520-R	GAAAGGTTGTGGTTGATTGG			
FbgSSR525-F	GGACACATCTCAATCATCCA	(CAGTCA/TGACTG)$_3$	4/5	0.23 ± 0.08
FbgSSR525-R	ACACATCTCTTGTTACAGCA			
FbgSSR545-F	TGAATTCTCTTCTCACGTGG	(TC/GA)$_6$	2/2	0.48 ± 0.01
FbgSSR545-R	CGAGTCAATTTGCACAAACT			
FbgSSR563-F	TTTATGAATTGGCGTTGTGG	(CT/AG)$_7$	9/9	0.32 ± 0.03
FbgSSR563-R	AACAAAACTCACCTTTCAATT			
FbgSSR564-F	TCCCTTTTGCTTGTTTATGA	(CAT/ATG)$_6$	2/3	0.19 ± 0.11
FbgSSR564-R	CCTCCGTGTTATCAAACAGT			
FbgSSR566-F	GCAAGAAGCAACATCCATTT	(AC/GT)$_6$	0/1	0
FbgSSR566-R	TTGCTTCAATCCTTCGAAGA			
FbgSSR599-F	TGTTTGGGACCTTTCTTTGA	(TC/GA)$_7$	0/1	0
FbgSSR599-R	GCAAGTCACCATCAAACAAA			
FbgSSR604-F	CGTTTTGGCTCATAATGCTT	(TTCCTC/GAGGAA)$_3$	4/4	0.33 ± 0.02
FbgSSR604-R	TTTTAGCCATGTACTGTGCT			
FbgSSR617-F	ATAGATGCCTCTCTCCATGT	(GA/TC)$_6$	0/1	0
FbgSSR617-R	GAAGGAGGACTAGACTGACT			
FbgSSR619-F	TATTTTAGTGGCCAGATGCA	(TA/TA)$_6$	4/5	0.34 ± 0.09
FbgSSR619-R	TGGAGAGGTGTTTCAACAAA			
FbgSSR623-F	AAAACCCATTTCTGGTACGA	(AAACTA/TAGTTT)$_3$	0/1	0
FbgSSR623-R	AGACAACCAACGTCGAATAA			
FbgSSR631-F	AATGTGATAAGCGCAACATG	(ACTCTCA/TGAGAGT)$_3$	2/2	0.23 ± 0.04
FbgSSR631-R	TTGGTATTTATCGCTTGTCT			
FbgSSR633-F	CTCCAAAACCAGAGTCTGTT	(TCATCG/CGATGA)$_3$	2/2	0.29 ± 0.09
FbgSSR633-R	TTTATCTGTAGAGGCATCGC			
FbgSSR643-F	GGCAAAAGATGGAGTCCTTA	(ACAAAACT/AGTTTTGT)$_3$	3/3	0.19 ± 0.07
FbgSSR643-R	TAATTTTTGGGCATTGGGAC			

Table 3 (continued)

SSR primers	Sequence	Repeat motif	Number of polymorphic alleles/total alleles	Gene diversity[a] (GD)
FbgSSR663-F	ACTCGAAATCCATCAAGCAT	$(AG/CT)_6$	8/8	0.36 ± 0.04
FbgSSR663-R	GCTTTGTGCACCAACAATAT			
FbgSSR675-F	ATTGGGGAACTGCCTAATTC	$(GA/TC)_7$	6/6	0.40 ± 0.04
FbgSSR675-R	GCAATTTATCAAACACTTGGTG			
FbgSSR679-F	TGGATTGCATGCATGGTATA	$(TAGT/ACTA)_5$	4/5	0.31 ± 0.08
FbgSSR679-R	TCCAAAAGTCAGCTTGATGA			
FbgSSR695-F	GTTCTGTAAACACTAGGGCA	$(GA/TC)_8$	9/9	0.41 ± 0.04
FbgSSR695-R	TGTTGACGGTGATTTGTTTG			
FbgSSR734-F	CTCTTCTACAACGTCCCAAA	$(GTTGGT/ACCAAC)_3$	0/1	0
FbgSSR734-R	TCTCCTGAAAACGGTAAAGG			

[a] For each marker, average gene diversity ± standard error is presented

the Turkish accessions were distributed between subpopulations 1, 2 and admixed.

A dendrogram was drawn using the Dice coefficient and the unweighted neighbor-joining algorithm (Fig. 2). Average pairwise dissimilarity among the 46 faba bean accessions was 0.29, with the highest value, 0.46 (54 % similarity), calculated between accessions from Turkey (TR37255) and Greece (7826). The lowest dissimilarity was 0.16 (84 % similarity) calculated between Australian accessions (Manafest and Fiord). The faba bean accessions grouped into three clusters (A, B, and C) in the dendrogram (Fig. 2). Cluster A included 26 accessions and dissimilarity ranged from 0.16 to 0.40 with an average was 0.28. Cluster A had four subclusters, one of these subclusters had all the accessions from Finland and two

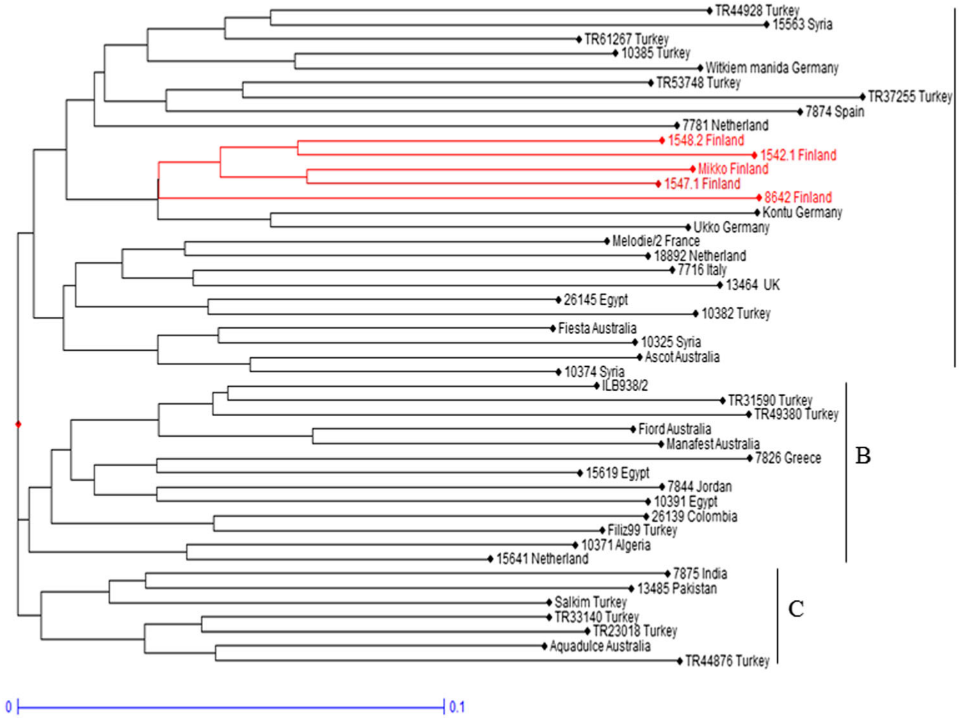

Fig. 2 Unweighted neighbor-joining dendrogram of 46 faba bean accessions based on 161 simple sequence repeat (SSR) alleles. Accession names and origins are provided

accessions from Germany. Cluster B had 13 accessions and cluster C had 7 accessions. The Turkish accessions were distributed to all three clusters A, B, C (Table 1; Fig. 2). Dendrogram and population structure analyses showed high correspondence. All of subpopulation 1 coincided with cluster A, except one accession from Pakistan which grouped in cluster C. Also, all of subpopulation 2 coincided with cluster B, except for four accessions which grouped in cluster C (Table 1).

Discussion

SSR Markers Developed by NGS

NGS has become a common method for discovering SSRs in plants because it can be easily performed on non-model organisms. Moreover, it is rapid and more cost-effective than traditional SSR development methods and Sanger sequencing (Zalapa et al. 2012), and allows sequencing of millions of bases. In this study, NGS did not provide good genome coverage (0.13 %), but was sufficient for the development of 2138 nonredundant SSR markers for the faba bean genome, allowing detection of one SSR marker every 7.6 kb (on average) in the 16.37 Mb of sequenced contigs. Cardle et al. (2000) reported one SSR every 6.8 Kb in genomic DNA of many plants, and one SSR every 6.04 kb for *Arabidopsis* genomic DNA compared to 14 kb for ESTs. Akash and Mayers (2012) developed mono-, di-, tri- and tetranucleotide EST-SSR markers from publicly available faba bean ESTs and reported one SSR repeat every 6.13 kb. When only the same types of repeats (mono-, di-, tri- and tetranucleotide) are compared, we found one SSR repeat every 8.9 Kb. El-Rodeny et al. (2014) examined di-, tri-, and tetranucleotides derived from faba bean ESTs with one SSR every 34.4 kb, while we observed one genomic SSR every 26.9 kb when mononucleotides were excluded. The variable frequency of genic and genomic SSRs may reflect a difference in their distribution in coding sequences compared to the entire genome. In addition, Leclercq et al. (2007) reported that the variable number of genomic SSRs identified in genomes is due to the algorithm tools used for mining and their parameter settings which can determine how many SSRs are detected.

Among the selected genomic SSRs, mononucleotide repeats were the most abundant (57.5 %) and outnumbered di- and trinucleotide repeats. Dinucleotide repeats (20.9 %) were more abundant than trinucleotide repeats (6.5 %). This result agrees with the fact that mono- and dinucleotide repeats outnumber trinucleotide repeats in eukaryotic intergenic and intron regions (Toth et al. 2000). Reports of genic SSR development in faba bean demonstrated that trinucleotides were most abundant in coding regions. For example, Akash and Mayers (2012) identified 38 % trinucleotide repeats followed by mononucleotides (36 %) and dinucleotides (22 %). El-Rodeny et al. (2014) also reported that trinucleotides were the most abundant and accounted for 72.7 % of SSRs, followed by dinucleotides (21.9 %). The motif length frequency differences between genomic and genic SSRs is most probably due to selection pressure on genic SSRs which reduces the fixation of mutations leading to frameshifts.

Among mononucleotidess, A/T repeats (Table 2) were the most frequent (98.9 %), agreeing with the observation that the most common SSR repeats in plants are A/T (Cardle et al. 2000). Among dinucleotide repeats, AG/CT was most frequently observed (26.6 %), followed by GA/TC (23.2 %) which is in agreement with El-Rodeny et al. (2014), who reported that AG/CT was the most frequent genic dinucleotide (57.4 %) in faba bean. Gong et al. (2011) reported that AG/CT and GA/TC were the most common (33.3 %) genic dinucleotides while we observed that GA/TC (23.2 %) was the second most common genomic dinucleotide. Among trinucleotides, AAT/ATT and ATA/TAT repeats were the most abundant accounting for 23.1 and 19.0 %, respectively. Cordoba et al. (2010) also reported that ATA/TAT repeats were the most abundant (46 %) in common bean while Cardle et al. (2000) reported that AAT/ATT was the most common trinucleotide in other plants.

Population Structure and Genetic Diversity Assessed with Genomic SSR Markers

A total of 39 SSR markers were selected based on their amplification efficiency and were applied to 46 faba bean accessions from 17 countries. All markers produced clear, reproducible fragments. Rare alleles were excluded from genetic analysis. The number of alleles ranged from 1 to 10 with an average of 4.1 alleles. Abid et al. (2015) reported up to 10 alleles with an average of 5.9 alleles per locus when genetic diversity of 46 faba bean accessions was analyzed using 17 SSR markers. In our study, population structure analysis assigned faba bean accessions into two distinct subpopulations and a group of admixed accessions. A dendrogram was constructed to understand the genetic relationship of faba bean accessions from different origins. Accessions from the same origin did not form exclusive clusters. Turkish accessions were distributed through all three clusters reflecting the genetic diversity of these accessions. This result was expected because most Turkish faba beans were introduced from different countries and have started to replace the few original Turkish accessions (Baloch et al. 2014). In contrast, all Finnish accessions grouped in one subcluster indicating the narrow genetic basis of these accessions. Population structure analysis coincided with dendrogram clustering with a few exceptions.

Conclusion

In conclusion, this is the first report of development of genomic SSR markers for faba bean using NGS. Sequencing and mining of the faba bean genome allowed identification of 2138 SSR markers. A subset of these SSR markers was used to test efficiency of PCR amplification and study genetic diversity and population structure within faba bean. Thus, the markers were found to be a useful tool for further studies of genetic diversity and population structure and in genetic mapping and breeding of faba bean.

Acknowledgments This study was supported by grant 0424.STZ.2013-2 from The Republic of Turkey's Ministry of Science, Industry and Technology to AF with contributions from Polen Seed Co. Additional support was received from the Scientific and Technological Research Council of Turkey grant 113E326 to JA. We are grateful to Ali Uncu and Ayse Ozger Uncu for SSR validation by sequencing.

References

Abdelkrim J, Robertson BC, Stanton JA, Gemmell NJ (2009) Fast cost-efective development of species-specific microsatalite markers by genomic sequencing. Biotechniques 46(3):185–192. doi:10.2144/000113084

Abid G, Mingeot D, Udupa SM, Muhovski Y, Watillon B, Sassi K, M'hamdi M, Souissi F et al (2015) Genetic relationship and diversity analysis of faba bean (*Vicia faba* L. var. *Minor*) genetic resources using morphological and microsatellite molecular markers. Plant Mol Biol Rep 33(6):1755–1767. doi:10.1007/s11105-015-0871-0

Akash MW, Mayers GO (2012) The development of faba bean expressed sequence tag–simple sequence repeats (EST-SSRs) and their validity in diversity analysis. Plant Breed 131(4):522–530. doi:10.1111/j.1439-0523.2012.01969.x

Baloch FS, Karakoy T, Demirbas A, Toklu F, Ozkan H, Hatipoglu R (2014) Variation of some seed mineral contents in open pollinated faba bean (*Vicia faba* L.) landraces from Turkey. Turk J Agric For 38:591–602. doi:10.3906/tar-1311-31

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156(2):847–854

Cordoba 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 Genomics 11:436. doi:10.1186/1471-2164-11-436

Doyle JJ, Doyle JE (1990) Isolation of plant DNA from fresh plant tissue. Focus 12:13–15

Duc G (1997) Faba bean (*Vicia faba* L.). Field Crop Res 53:99–109

Duc G, Bao S, Baum M, Redden B, Sadiki S, Suso MJ, Vishniakova M, Zong X (2010) Diversity maintenance and use of *Vicia faba* L. genetic resources. Field Crop Res 115:270–278. doi:10.1016/j.fcr.2008.10.003

Ellwood SR, Phan HTT, Jordan M, Hane J, Torres AM, Avila CM, Cruz-Izquierdo S, Oliver RP (2008) Construction of a comparative genetic map in faba bean (*Vicia faba* L.); conservation of genome structure with Lens culinaris. BMC Genomics 9:380. doi:10.1186/1471-2164-9-380

Earl DA, VonHolt BM (2012) Structure Harvester: a website and program for visualizing structure output and implementing the Evanno method. Conserv Genet Resour 4:359–361. doi:10.1007/s12686-011-9548-7

El-Rodeny W, Kimura M, Hirakawa H, Sabah A, Shirasawa K, Sato S, Tabata S, Sasamoto S et al (2014) Development of EST-SSR markers and construction of a linkage map in faba bean (*Vicia faba*). Breed Sci 64(3):252–263. doi:10.1270/jsbbs.64.252

FAOSTAT 2014. Crops. Available online: http://faostat3.fao.org

Gong YM, Xu SC, Mao WH, Hu QZ, Zhang GW, Ding J, Li ZY (2010) Generation and characterization of 11 novel est derived microsatellites from *Vicia faba* (Fabaceae). Am J Bot 97:e69–e71. doi:10.3732/ajb.1000166

Gong YM, Xu SC, Mao WH, Lize Y, Hu QZ, Zhang GW, Ding J (2011) Genetic diversity analysis of faba bean (*Vicia faba* L.) based on EST-SSR markers. Agric Sci China 10:838–844. doi:10.1016/S1671-2927(11)60069-2

Gresta F, Giovanni A, Emidio A, Lorenzo R, Valerio A (2010) A study of variability in the Sicilian faba bean landrace 'Larga di Leonforte'. Genet Resour Crop Evol 57(4):523–531. doi:10.1007/s10722-009-9490-7

Hamada H, Petrino MG, Kakunaga T (1982) A novel repeated element with Z-DNA forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc Natl Acad Sci U S A 79(21):6465–6469

Kaur S, Cogan NOI, Forster JW, Paull JG (2014) Assessment of genetic diversity in faba bean based on single nucleotide polymorphism. Diversity 6(1):88–101. doi:10.3390/d6010088

Kaur S, Pembleton LW, Cogan NOI, Savin KW, Leonforte T, Paull J, Materne M, Forster JW (2012) Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. BMC Genomics 13:104. doi:10.1186/1471-2164-13-104

Koressaar T, Remm M (2007) Enhancement and modifications of primer design program Primer3. Bioinformatics 23(10):1289–1291. doi:10.1093/bioinformatics/btm091

Langmead B, Salzberg SL (2012) Fast gapped – read alignment with Bowtie2. Nat Methods 9(4):357–359. doi:10.1038/nmeth.1923

Leclercq S, Rivals E, Jarne P (2007) Detecting microsatellites within genomes: significant variation among algorithms. BMC Bioinf 8:125. doi:10.1186/1471-105-8-125

Link W, Dixkens C, Singh M, Schwall M, Melchinger AE (1995) Genetic diversity in European and Mediterranean faba bean germ plasm revealed by RAPD markers. Theor Appl Genet 90:27–32. doi:10.1007/BF00220992

Ma Y, Yang T, Guan J, Wang S, Wang H, Sun X, Zong X (2011) Development and characterization of 21 EST-derived microsatellite markers in *Vicia faba* (fava bean). Am J Bot 98:e22–e24. doi:10.3732/ajb.1000407

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetJ 17(1):10–12, doi:10.14806/ej.17.1.200

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci U S A 70(12):3321–3323. doi:10.1073/pnas.70.12.3321

Ouji A, El-Bok S, Syed NH, Abdellaoui R, Rouaissi M, Flavell AJ, El-Gazzah M (2012) Genetic diversity of faba bean (*Vicia faba* L.) populations revealed by sequence specific amplified polymorphism (SSAP) markers. Afr J Biotechnol 11:2162–2168. doi:10.5897/AJB11.2991

Pozarkova D, Koblizkova A, Roman B, Torres AM, Lucretti S, Lysak M, Dolezel J, Macas J (2002) Development and characterization of microsatellite markers from chromosome 1-specifi c DNA libraries of Vicia faba. Biol Plant 45:337–345

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

Roldan-Ruiz I, Dendauw J, Bockstaele EV, Depicker A, Loose MD (2000) AFLP markers reveal high polymorphic rates in ryegrasses (*Lolium* spp.). Mol Breed 6:125–134. doi:10.1023/A:1009680614564

Senan S, Kizhakayil D, Sasikumar B, Sheeja TE (2014) Methods for development of microsatellite markers: an overview. Not Sci Biol 6(1):1–13. doi:10.15835/nsb.6.1.9199

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145. doi:10.1038/nbt1486

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123. doi:10.1101/gr.089532.108

Suresh S, Park JH, Cho GT, Lee HS, Baek HG, Lee SY, Chung JW (2013) Development and molecular characterization of 55 novel polymorphic cDNA-SSR markers in faba bean (*Vicia faba* L.) using 454 pyrosequencing. Molecules 18:1844–1856. doi:10.3390/molecules18021844

Terzopoulos PJ, Bebeli PJ (2008) Genetic diversity analysis of Mediterranean faba bean (*Vicia faba* L.) with ISSR markers. Field Crop Res 108:39–44. doi:10.1016/j.fcr.2008.02.015

Torres AM, Roman B, Avila CM, Satovic Z, Rubiales D, Sillero JC, Cubero JI, Moreno MT (2006) Faba bean breeding for resistance against biotic stresses: towards application of marker technology. Euphytica 147(1):67–80. doi:10.1007/s10681-006-4057-6

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

Van de Ven M, Powell W, Ramsay G, Waugh R (1990) Restriction fragment length polymorphisms as genetic markers in *Vicia*. Heredity 65:329–342. doi:10.1038/hdy.1990.102

Wang HF, Zong XX, Guan JP, Yang T, Sun XL, Ma Y, Redden R (2012) Genetic diversity and relationship of global faba bean (*Vicia faba* L.) germplasm revealed by ISSR markers. Theor Appl Genet 124(5): 789–797. doi:10.1007/s00122-011-1750-1

Yang T, Bao SY, Ford R, Jia TJ, Guan JP, He YH, Sun XL, Jiang JY et al (2012) High-throughput novel microsatellite marker of faba bean via next generation sequencing. BMC Genomics 13:602. doi:10.1186/1471-2164-13-602

Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot 99(2):193–208. doi:10.3732/ajb.1100394

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Mol Ecol 11(1):1–16. doi:10.1046/j.0962-1083.2001.01418.x

Zeid M, Schon CC, Link W (2003) Genetic diversity in recent elite faba bean lines using AFLP markers. Theor Appl Genet 107(7):1304–1314. doi:10.1007/s00122-003-1350-9

Zeid M, Mitchell S, Link W, Carter M, Nawar A, Fulton T, Kresovich S (2009) Simple sequence repeats (SSRs) in faba bean: new loci from Orobanche-resistant cultivar Giza 402. Plant Breed 128(2):149–155. doi:10.1111/j.1439-0523.2008.01584.x

Zong X, Liu X, Guan J, Wang S, Liu Q, Paull JG, Redden R (2009) Molecular variation among Chinese and global winter faba bean germplasm. Theor Appl Genet 118(5):971–978. doi:10.1007/s00122-008-0954-5