Agronomie 18 (1998) 209–224

Original article

Variability in six pea gene sequences and mapping through PCR-based markers

Nathalie Pavy, Jan Drouaud, Beate Hoffman, Georges Pelletier, Dominique Brunel*

Unité de génétique et amélioration des plantes, Institut national de recherche agronomique, Route de Saint Cyr, 78000 Versailles, France

(Received 22 January 1998; accepted 8 April 1998)

Abstract – Despite the rapid increase in sequence databases, gene sequences are still under-used in the plant breeding and genetic mapping area. This study was conducted to determine whether pea gene sequences contained enough polymorphism to be used as genetic markers. Molecular variability was examined at the DNA sequence level within different lines and wild ecotypes of *Pisum sativum*. The analysis was conducted for several introns, exons and 5'UTR sequences from six nuclear genes (*GAPC*, *PHYA* and *IAA*-related genes). Each region was specifically amplified and polymorphism was identified by electrophoretic mobility and by direct sequencing of PCR products. The observed polymorphism illustrates the possibility of developing molecular markers since all the analyzed loci have been successfully localised. Polymorphism was detected either as DNA conformational polymorphism following non-denaturing polyacrylamide gel electrophoresis or as CAPS (cleaved amplified polymorphism sequence). The noteworthy property of such genetic markers is their ability to establish bridges between different existing pea genetic maps. (© Inra/Elsevier, Paris.)

DNA sequence polymorphism / DSCP / intron / genetic marker / *Pisum sativum*

Résumé – Analyse de la variabilité des séquences de six gènes de pois et cartographie génétique par PCR. Malgré l'accroissement rapide des bases de données, les séquences des gènes sont encore sous-exploitées dans le domaine de l'amélioration des plantes et de la cartographie génétique. Cette étude a été menée pour déterminer si les séquences des gènes du pois renferment suffisamment de polymorphisme pour générer des marqueurs génétiques. La variabilité moléculaire a été examinée au niveau de la séquence de l'ADN entre différentes lignées et écotypes sauvages de *Pisum sativum*. Cette analyse a été effectuée pour plusieurs séquences introniques, exoniques et promotrices de six gènes nucléaires (*GAPC*, *PHYA*, et gènes apparentés aux gènes *IAA*). Chaque région a été amplifiée spécifiquement et le polymorphisme a été identifié par la mobilité électrophorétique et par séquençage direct des produits de PCR. Le polymorphisme observé illustre la possibilité de développer des marqueurs moléculaires puisque tous les loci analysés ont pu

Communicated by Hervé Thiellement (Versailles)

* Correspondence and reprints
E-mail: brunel@versailles.inra.fr

être cartographiés. Le polymorphisme est détecté soit comme un polymorphisme de conformation de l'ADN après une électrophorèse dans un gel d'acrylamide dans des conditions non dénaturantes, soit comme des CAPS (*Cleaved Amplified Polymorphism Sequence*). La propriété remarquable de tels marqueurs génétiques est leur capacité à établir des ponts entre les différentes cartes génétiques existantes chez le pois. (© Inra/Elsevier, Paris.)

polymorphisme de séquence de l'ADN / DSCP / intron / marqueur génétique / *Pisum sativum*

Abbreviation: DSCP: double strand conformation polymorphism

1. INTRODUCTION

The amount of polymorphism which can be revealed is a critical parameter for the evaluation of a genetic marker system. A whole range of molecular tools have proved useful for detecting polymorphism in many species. As far as the genus *Pisum* is concerned, the variability has been well studied at the molecular level by several methods including isozymes [32], separation of RAPD fragments on polyacrylamide gels [26], and RFLPs as well as AFLP or inter SSR-PCR [21]. Thus, several genetic maps of the pea genome based on isozymes [32], RFLPs [9, 11], AFLPs [14] and RAPDs are now available. Increasing the use of sequences which are potential markers of important agronomic traits in mapping experiments would be of interest for plant breeders.

The rapid growth of the sequence databases and the ease of access to this information about genes are attractive reasons for developing genetic markers which enable one to localise gene sequences on genetic maps. Until now, gene mapping in plants has mainly consisted of using RFLP probes from cDNA libraries. The use of these gene sequences has enabled the establishment of expressed-gene maps which have proved important in rice [19] and maize [4]. PCR markers relying on gene sequence polymorphism have so far been of limited use for plant breeding. However, techniques such as CAPS (cleaved amplified polymorphism sequence) [18] or SSCP (single strand conformation polymorphism) [23–28] are available which have proved useful in revealing polymorphism in gene sequences in several species. Indeed, although some features of introns, such as length, splicing sites and some signals involved in gene expression are sometimes strikingly conserved between species, introns exhibit sequence polymorphism which is being used more and more for population genetics studies [24, 25], or systematics [2]. By using primers homologous to splice site junctions for PCR, this kind of variability has also been useful for distinguishing barley varieties [6] or for the identification of commercial yeast strains [7]. More recently, the 5' untranslated regions have been successfully used to develop genetic markers, as for instance for discriminating *Vitis vinifera* genotypes [12]. Some examples of genotype-specific gene sequences have already been described in comparisons of wild type and mutant plants in pea [15].

In this study, we have focused on sequence differences revealing polymorphism in genes from non-mutant phenotypes of pea. We describe genetic markers which were obtained by PCR amplification with primers derived from available data. We investigated exons, introns and 5'UTR sequences in six genes. Sequence polymorphism was shown to be present in all the sequenced genes and these markers are good landmarks for the pea consensus genetic map whose construction is hampered by chromosomal rearrangements and a lack of anchor markers [31]. The development of such molecular markers, which are based on sequences having a known biological role, will help in the construction of a map of known function sequences in pea.

The studied sequences include: *GAPC*, *PHYA* and *IAA*-related sequences. The *GAPC* gene is a nuclear gene encoding the cytosolic form of glyc-

eraldehyde phosphate dehydrogenase, an enzyme taking part in glycolysis. Since plant architecture and development are important selection critera for pea breeders the other genes used in this study were selected on this basis. *Phytochrome A* is involved in photoreception. The *IAA* genes belong to a family of genes whose transcription is affected by auxin.

2. MATERIALS AND METHODS

2.1 Plant material

To evaluate polymorphism, seven ecotypes from diverse geographic origins including cultivars or wild populations were examined: Champagne, Chine, *P. sativum* ssp. *abyssinicum*, *P. sativum* ssp. *transcaucasicum*, *P. sativum* ssp. *palestinicum*, *P. jomardi* and *P. fulvum*. For each ecotype, genomic DNA was extracted from a bulk of three plantlets. Thus, as the plants are supposed to be diploid homozygous, six alleles were analysed in each ecotype for each locus. For the cultivar Champagne, a study was performed with four plants whose DNA was extracted individually in the aim of analysing heterogeneity in that population. The *PHYA* gene is single copy in the pea genome [27]. For the other genes, as a first approximation, we consider that all loci are present in a single copy in the genome. Therefore, fragments which are specifically amplified with primers defined from a particular gene will be considered as allelic forms. This aspect will be discussed later.

Seven lines already used in three mapping programmes were introduced in the study:

– cv. Térèse (a French proteaginous cultivar) and cv. Torsdag (a central European cultivar), used to build a RAPD-based map (C. Rameau, in prep.);

– 661 (a French proteaginous line) and cv. Erygel (a garden pea cultivar), used to build a RFLP-based map [9];

– JI281 (an ethiopian line), JI399 (cv. Cennia) and JI15, lines from the John Innes *Pisum* germplasm collection which were used to build a RFLP-based map [11].

For the mapping analysis, two populations of recombinant inbred lines were used: one consisting of 139 lines from the cv.Térèse × cv.Torsdag cross (provided by C. Rameau, Inra), the other consisting of 71 lines from the JI281 × JI399 cross (provided by M.Ambrose, John Innes Centre).

2.2. Targeted sequences and PCR primer design

Table I lists the six studied loci, the accession numbers in GenBank, the sequences of the primers, the type and the size of the PCR products. Degenerate consensus primers were deduced from the alignment of the pea gene *IAA4/5D* and the *Arabidopsis* gene *At-AUX211*. In all the genotypes, which were analysed in pea, these primers enabled the amplification of three genes: *IAA1300*, *IAA4/5D*, *IAA850* which are, respectively, 1 300, 900 and 850 bp long. These genes as well as *IAA6* belong to the same multigenic family and share sequence similarity in their coding sequences in four domains. The precise positions of the sequences and the structure of the genes studied are shown in *figure 1*. Specific primers of 20–23 bases were designed from the sequences available in GenBank using the OLIGO4S program (National Biosciences Inc.) [25]. The oligonucleotide primers were purchased from Genosys Biotechnologies.

2.3. PCR amplification and electrophoresis

Genomic DNA was extracted according to Doyle and Doyle [10]. PCR amplifications were performed on 100 ng of DNA template per 50 µL reaction volume; each reaction contained 100 ng of each primer, 1.5 mM $MgCl_2$, 0.1 mM of each dNTP and 1.25 units of *Taq* DNA polymerase (Eurobio) in the reaction buffer provided by the supplier. PCR reactions were performed in a thermal cycler apparatus (MJ Research, Inc.), with the following conditions: 94 °C for 4 min, 35 cycles comprising 94 °C for 1 min, 53 °C for 1 min, 72 °C for 2 min, and a last elongation step at 72 °C for 6 min. The presence of PCR products was checked on 1 % agarose gels. Amplified bands were separated by electrophoresis on 5 % polyacrylamide (acrylamide: N-N' methylenbisacrylamide 29:1) non-denaturing gels in 1X TBE, thermostabilised at 20 °C (1.5 mm thickness, 32 cm height). After a migration time of 20 h at 210V the gels were stained with ethidium bromide and visualised under ultraviolet light. A *Hae*III digest of φx174 DNA (Gibco BRL) was used as a molecular weight marker.

Table I. References of the loci plus the position and sequence of the primers for amplifying them.

Locus	Accession number	'Upper' primer 5' → 3'	'Lower' primer 5' → 3'	Nature of the amplified fragment	Size of the PCR product (bp)	Biological function	Reference
*PS-GAPC1**	X73150	cagtgttcacggtcaatgga	gcagctcttccacctctcca	exon 4 to exon 8	1 200	glycolysis	[16]
*PS-PHYA***	M37217	ctttctttctcccacaccttca	gctttgcatccacagtcgtct	intron 1	674		
PS-PHYA	M37217	tgatggcggtcccttctg	accggccttgtttgtctcct	microsatellite	112	light	[27]
PS-PHYA	M37217	agtgacaagtgagatggttagatt	atcaactgatttgggttctgc	intron 2	520	perception	
PS-PHYA	M37217	gcaaaattcttgacgactctgat	agctttaacaacttccgactgat	intron 3 to intron 4	667		
PS-IAA4/5D	X68216	tccattcacatgctcatgtttc	tgtatgattttcttctcctcttcctc	5'UTR	456	auxin	[1]
PS-IAA6	X68218	aagcaaaccacatggcatgtttc	ccacccgacaacttgattcttctta	5'UTR	502	regulated	[1]
PS-IAA1300	AF026532	gtaagatggtgttggaaatggca	cgcctcctaatcccagtttt	intron	625	genes	this study
PS-IAA850	AF026531	gtgagagtgaaagcgaaataacc	tctgatccccctatgctctttg	intron	388		this study

* *PS-GAPC1* PCR product sequence was completed by sequencing with a third internal primer which is located in the exon 5: 5' tgataaggacaaggctgctgct 3'.
** Gene coding for the phytochrome A in pea is called either *PS-PHYA* or *peaphtap* in GenBank.
Degenerate consensus primers enabling the amplification of *IAA* genes both in pea and *Arabidopsis* are: 5' acagagctgagattgggnytncc 3' (upper primer) and 5' taccatcaccatctttrtcytcrta 3' (lower primer).

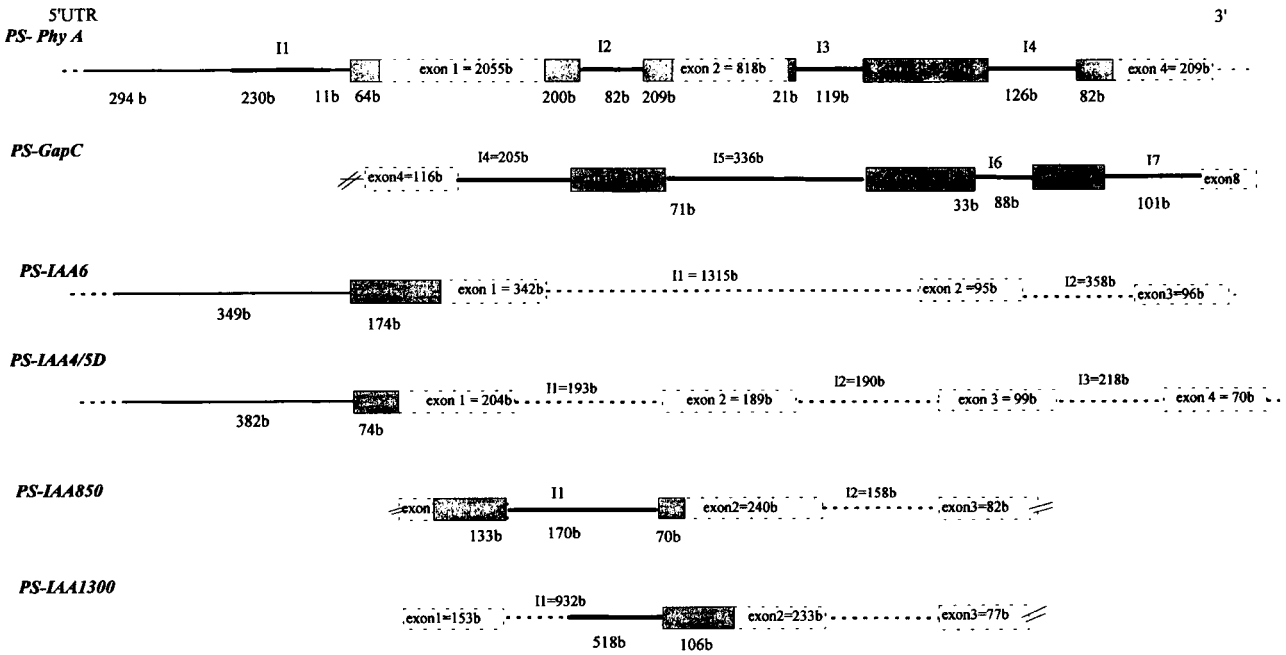

Figure 1. Structure of the genes and positions of the analysed sequences. Exons are represented by boxes and non-coding sequences by lines. Dotted lines indicate regions that were not sequenced. Bold lines show the sequenced introns and grey boxes show the sequenced exons.

2.4. Sequencing

Once the purity of the PCR product was verified, excess primers and salt were eliminated by centrifugation on microconcentrators with a 100 kD cut-off (Amicon, Inc.). Direct sequencing of PCR products was performed on 100 ng of template and 20 ng of primer with the Dye-Terminator kit supplied by Perkin Elmer in a thermal cycler (MJ research, Inc). After cycle sequencing, extension products were precipitated by adding 2 µL sodium acetate (3 M) and 50 µL ethanol (95 %). The pellet was rinsed with 250 µL ethanol (80 %). The sample was resuspended in a formamide: EDTA/blue dextran (in a 5:1 ratio) buffer, heated at 90 °C and loaded on a denaturing gel. Gel analysis was performed using an ABI 373A DNA sequencer. The sequencing method enabled us to obtain the sequence of 500–600 bases in a single reaction.

2.5. Data scoring and analysis

Sequences were analysed using the GCG package [8].

Mutation frequency was calculated using the ratio of the number of mutations to the sequence length. Variation among genotypes was estimated with the number of pairwise nucleotide differences. The k parameter was calculated as follows: $k = \Sigma_i \Sigma_j k_{ij}/[(N(N-1)/2]/m$, where k_{ij} is the number of differences between the sequences i and j and where N is the number of compared sequences ($N = 16$) and m the sequence length.

Multipoint linkage analysis was performed using the MAPMAKER software [20] with a LOD score value of 3.0.

3. RESULTS

3.1 Amplification results

Ten introns, 12 exons, two promoters and a third promoter including an intron, of six nuclear genes were completely or partially amplified and

sequenced in DNA from 14 pea genotypes. To appreciate the polymorphism, direct sequencing of the PCR products was chosen for its rapidity, precision and reliability. During PCR amplification, *Taq* polymerase introduces errors. Nevertheless, data derived from direct sequencing of PCR products have been demonstrated to be more reliable than sequences obtained after a cloning step [30]. PCR product identity was systematically confirmed by sequence comparison with data available in GenBank.

Mostly, a single amplified fragment was observed on a polyacrylamide gel from the DNA of each ecotype. However, in two cases (*PHYA* and *IAA6*), with the DNA from Champagne and from the subspecies *palestinicum*, two bands were amplified and separated after polyacrylamide gel electrophoresis (PAGE). For each gene, the two bands differed in their sequences at only a few sites. Thus we will consider 16 genotypes (the two genotypes from Champagne and *P. sativum* ssp. *palestinicum* will be called Champagne1, Champagne2, *palestinicum*1 and *palestinicum*2).

3.2. Polymorphism analysis

The number of variants which were observed for each sequence was relatively low. Polymorphism in the studied sequences mainly consisted of substitutions, whereas insertion/deletion events are rare: among the 4 617 compared bases, 126 substitutions and 15 insertion/deletions were observed. All substitutions involved two nucleotides except for three sites, where at least three substitutions occur. DNA sequence variability can be evaluated by the number of polymorphic sites in each sequence and by the number of polymorphic genotypes per site. These two parameters are quantified in *table II.*

3.2.1. Insertion/deletion polymorphism

In our sample, the only polymorphism in the *PHYA* gene was an insertion/deletion of three bases in the repeated motif $(AAT)_n$ in the 5'UTR sequence. This microsatellite is polymorphic between JI281 and JI399 and between 661 and Erygel. In order to highlight this difference, primers flanking the motif were used to amplify a small fragment (112 bp). This microsatellite generated little polymorphism among the analysed genotypes, three genotypes share the $(AAT)_5$ allele, the 13 other genotypes share the $(AAT)_4$ allele.

3.2.2. Substitution polymorphism

In some cases, the differences can be detected directly after separation with a 5 % non-denaturing polyacrylamide gel. Examples of the observed variability are illustrated in *figure 3* showing migration of PCR products from the 5'UTR of the *IAA4/5D* gene. The multiple sequence alignment reveals that polymorphism is due only to base substitutions. For instance, there are only four substitutions between the fragments produced from cv. Térèse and cv. Torsdag (*figure 2*). The difference in migration reflects different conformations of the DNA (DSCP: double strand conformation polymorphism). The difference between the JI399 and JI15 samples which is visible on the gel (*figure 2*) involves only a G to C transversion. Whenever polymorphism was not directly revealed after electrophoresis, comparison of sequences and of restriction maps enabled us to choose restriction enzymes distinguishing the genotypes after digestion. The enzymes employed to reveal these CAPS markers are indicated in *table III.*

3.3. Genotype distinction

The ability of the markers to distinguish each genotype was evaluated as the ratio of polymorphic genotype pairs in each gene sequence. Qualitatively, polymorphism distribution among genotypes is an important criterion for evaluating the efficiency of the markers. Considering all the analysed genes, all the genotypes studied can be individually identified. By adding all data concerning one gene, we note that *PHYA* enables the dis-

Table II. Polymorphism rates detected by sequencing PCR products among the genotypes.

Locus	Sequence	Sequence lenght (m) (number of bases)	Substitution number (S)	Substitution frequency ($S/m \times 10^3$)	Indel** number (I)	Total indels length (bp)	Indel frequency ($I/m \times 10^3$)	ks*** substitutions	ki*** indel	No. of polymorphic genotype pairs
				Variation between sites				Variation between genotypes		
PHYA	5'region*	294	0	0	1	3	3	0	0.001	48
	intron 1	230	5	22	1	1	4	0.0045	0.0005	54
	5'UTR	524	5	*9.5*	2	4	4	0.002	0.001	*81*
	exon 1	264	2	7.5	0	0	0	0.0009	0	15
	intron 2	82	1	12	0	0	0	0.001	0	15
	exon 2	209	1	5	0	0	0	0.001	0	28
	intron 3	119	2	17	1	1	8	0.002	0.001	29
	exon 3	251	1	4	0	0	0	0.0005	0	15
	intron 4	136	0	0	0	0	0	0	0	0
	exon 4	82	0	0	0	0	0	0	0	0
GAPC	intron 4	198	10	50	1	1	5	0.009	0.0006	91
	exon 5	100	3	30	0	0	0	0.004	0	29
	intron 5	341	35	**103**	3	13	9	0.021	0.001	**113**
	exon 6	147	6	41	0	0	0	0.008	0	90
	intron 6	88	6	68	1	1	11	0.013	0.001	69
	exon 7	61	1	16	0	0	0	0.002	0	15
IAA6	*5'UTR*	337	7	*21*	2	12	6	0.005	0.001	*103*
	exon 1	177	3	17	0	0	0	0.006	0	82
IAA4/5D	*5'UTR*	382	5	*13*	0	0	0	0.004	0	*103*
	exon 1	74	1	13.5	0	0	0	0.007	0	63
IAA850	exon 1	137	4	29	0	0	0	0.005	0	53
	intron 1	174	16	**92**	4	31	23	0.025	0.003	**105**
	exon 2	80	1	12.5	0	0	0	0.002	0	15
IAA1300	intron	518	16	31	1	1	2	0.011	0.0008	90
	exon	106	0	0	0	0	0	0	0	0

* 5'region in *PHYA* designates the 5'UTR sequence excluding the intron 1.
** Indel: insertion/deletion
*** $k = \Sigma_i \Sigma_j k_{ij}/[(N(N-1)/2]/m$, where k_{ij} is the number of differences between the sequences i and j and where N is the number of compared sequences ($N = 16$) and m the sequence length.

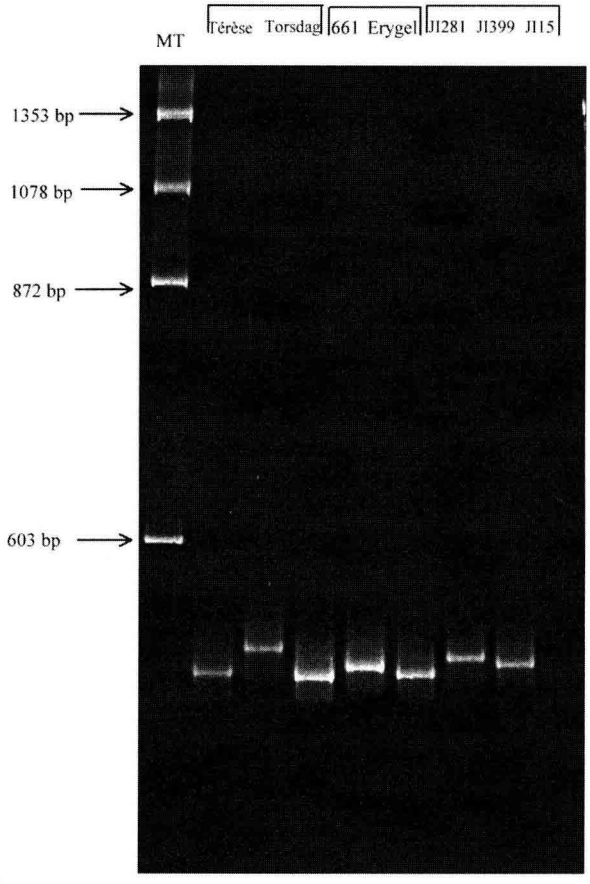

Figure 2. Non-denaturing polyacrylamide gel electrophoresis of the PCR product corresponding to the 5'UTR sequence of the *IAA4/5D* gene. Amplification was performed from genomic DNA of the parents of crosses used in genetic mapping. Parents of each cross are bracketed. MT is the molecular weight marker (φX174 DNA digested by *Hae*III).

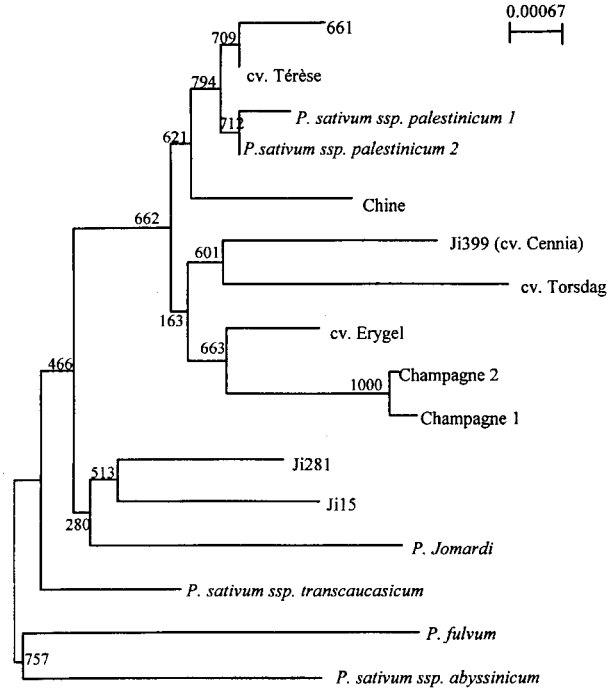

Figure 3. Dendrogram. A tree was derived by the 'Neighbour-Joining' method with the Clustal V program. The distance matrix was calculated according to all data and by considering all mutations (substitutions as well as insertion/deletions). The nodes are labelled with the score of the boostrap analysis with 1 000 replicates and based on the 50 % majority rule.

tinction of five ecotypes, while the *GAPC* sequence discriminates all analysed ecotypes. The number of alleles which were identified for each gene is detailed in *table III.*

The dispersion and clustering of the genotypes are represented by a dendrogram (*figure 3*). It is not our intention to reconstruct the evolutionary history of these genotypes but it serves as a summary of all the observations made from the multiple sequence alignments. *P. fulvum* and the subspecies *abyssinicum* are clearly separated from the other genotypes in the genus *Pisum*. This separation is due to more frequent insertion/deletions in introns and to a higher level of substitutions. Except for these two genotypes, the wild ecotypes which were analysed do not reveal more sequence variation than the cultivated genotypes used for genetic mapping. Lu et al. [21], who analysed genetic diversity in pea with several molecular tools, came to similar conclusions. Varieties form a group in which the subspecies *palestinicum* is included. The genotypes which were selected for genetic mapping represent a large and diverse population since JI281, JI15 and *P. jomardi* form a separate group from JI399 (cv. Cennia) which

Table III. Number of detected alleles, kinds of revealed markers and their position on the pea genetic map. The enzymes used to detect CAPS markers are given in parentheses. Italic characters indicate the technique which was chosen for the mapping. Linkage groups refer to the pea genetic map published by Ellis et al. [11].

Locus	Sequence size (bp)	Number of different alleles observed among all the studied genotypes	Térèse/Torsdag	661/Erygel	JI281/JI399	JI399/JI15	Linkage group	Localization on the genetic map with reference to markers published by Hall et al. [14]
PHYA	1973	5	monomorphic	microsatellite length polymorphism	microsatellite *length polymorphism*	monomorphic	II	between E3/9- and F15/8+
IAA4/5D	456	7	*DSCP*	DSCP	DSCP CAPS(EcoRI)	DSCP	II	between E3/5++ and *a*
IAA6	502	9	undetected mutations	monomorphic	DSCP	DSCP	VII	extremity of group VII
IAA1300	625	7	monomorphic	monomorphic	DSCP *CAPS(Bsp1206I)*	DSCP	I	between F19/14- and cDNA150/1
IAA850	388	10	DSCP CAPS (DdeI or RsaI)	monomorphic	DSCP *CAPS* (*DdeI* or RsaI)	DSCP	II	between *a* and cDNA260
GAPC	1200	14	*CAPS* (*BstYI* or HhaI)	CAPS(HhaI)	CAPS(HhaI)	CAPS (NlaIII or MaeIII)	III	has not been precisely mapped because of a biased segregation ratio

DSCP: double strand conformation polymorphism.

belongs to a cultivated group. The proximity shared by the two proteaginous cultivars Térèse and 661 may reflect the narrow genetic basis of these varieties.

3.4. Genetic mapping

The genes *IAA4/5D*, *IAA6* and *GAPC* were mapped on the RAPDs-based map constructed in our laboratory by Laucou V., Haurogne K., Ellis N, Rameau C. (Theor. Appl. Genet., accepted). *PHYA*, *IAA850* and *IAA1300* were mapped on the reference pea genetic map constructed by Ellis et al. [11]. *PHYA* and *IAA850* are linked to the *legJ* gene and to the morphological marker *a*. The *IAA1300* gene is on the same linkage group (I) as the classical gene *i* and the reference marker cDNA44 which is tightly linked to the *sym2* locus. The linkage groups to which the analysed genes belong are summarised in *table III*. Interestingly, four of the six loci have been localised on at least two genetic maps. The gene, *IAA4/5D* can be mapped on three pea genetic maps (*figure 2*). The markers which are common to several maps are indicated in *table III*.

3.5. Polymorphism description within the different regions of the genes

The amount of polymorphism cannot be easily correlated with the sequence type (intron/exon) but it is generally higher in introns than in exons.

There are 21 synonymous sequence substitutions in all the exons. Fifteen of them appear only once (seven of them concerning *P. fulvum*) and six are observed for at least two genotypes. Only four amino acid changes were found: Leu/Val (in the *IAA850* gene from *P. fulvum* and *P. sativum* ssp. *abyssinicum*), Arg/Met (in *PHYA* in *P. fulvum*), Glu/Ala (in *IAA6* in *P. sativum* ssp. *abyssinicum* and *P. sativum* ssp. *palestinicum*) and Phe/Tyr (in *GAPC* in *P. jomardi*).

Two extreme situations illustrate the relationship between the amount of polymorphism and its distribution. When only one genotype is affected by several mutations, polymorphism may appear high (substitution number/sequence length is high), but it does not generate markers which are useful for the mapping analysis. However, when a single mutation affects several genotypes, the sequence presents little polymorphism (substitution number/sequence length is low) but generates discriminating markers for mapping purposes. These two cases are represented in *figure 4* showing the allele repartition in each sequence according to its size. In each sequence, the most frequent allele is represented by grey boxes. The higher this kind of bar is, the less we find genetic markers. For example, the *IAA1300* exon generates no genetic markers; most exons and introns in *PHYA* and exon 5 in *GAPC* generate only a few genetic markers.

As few alleles are detected, their distribution is crucial for the evaluation of the genetic markers. In the *IAA4/5D* exon, the two alleles enabled us to discriminate 63 of the 120 genotype pairs and generate CAPS markers which can be localised on three genetic maps (*table III*) whereas the two alleles in exons 1 and 3 of *PHYA* separate only *P. fulvum* from all the other genotypes. The more mosaic the bar is (*figure 4*), the easier is the characterisation: intron 5 in *GAPC*, *IAA850* intron and the 5'UTR sequences in the *IAA* genes generate genetic markers which are of interest both for genetic mapping and for the classification of the studied genotypes. The highest level of polymorphism is in *GAPC* intron 5, distinguishing 113 of the 120 genotype pairs with 341 sequenced base pairs. Intron1 in the *IAA850* sequence, which is 174 bp long, enables the distinction of 105/120 genotype pairs. Although the number of polymorphic sites is not higher in 5'UTR sequences than in other analysed sequences, the level of molecular markers found in 5'UTR sequences is high: *IAA4/5D* and *IAA6* 5'UTR sequences discriminate 105/120 genotype pairs. As can be seen in *figure 4*, the number of detected alleles does not result from an experimental bias due to variability in the length of the sequenced DNAs. For example, intron 6 in *GAPC* is shorter but reveals more alleles than intron 3 in *PHYA*.

Examination of multiple sequence alignments shows that polymorphism is not equally distributed

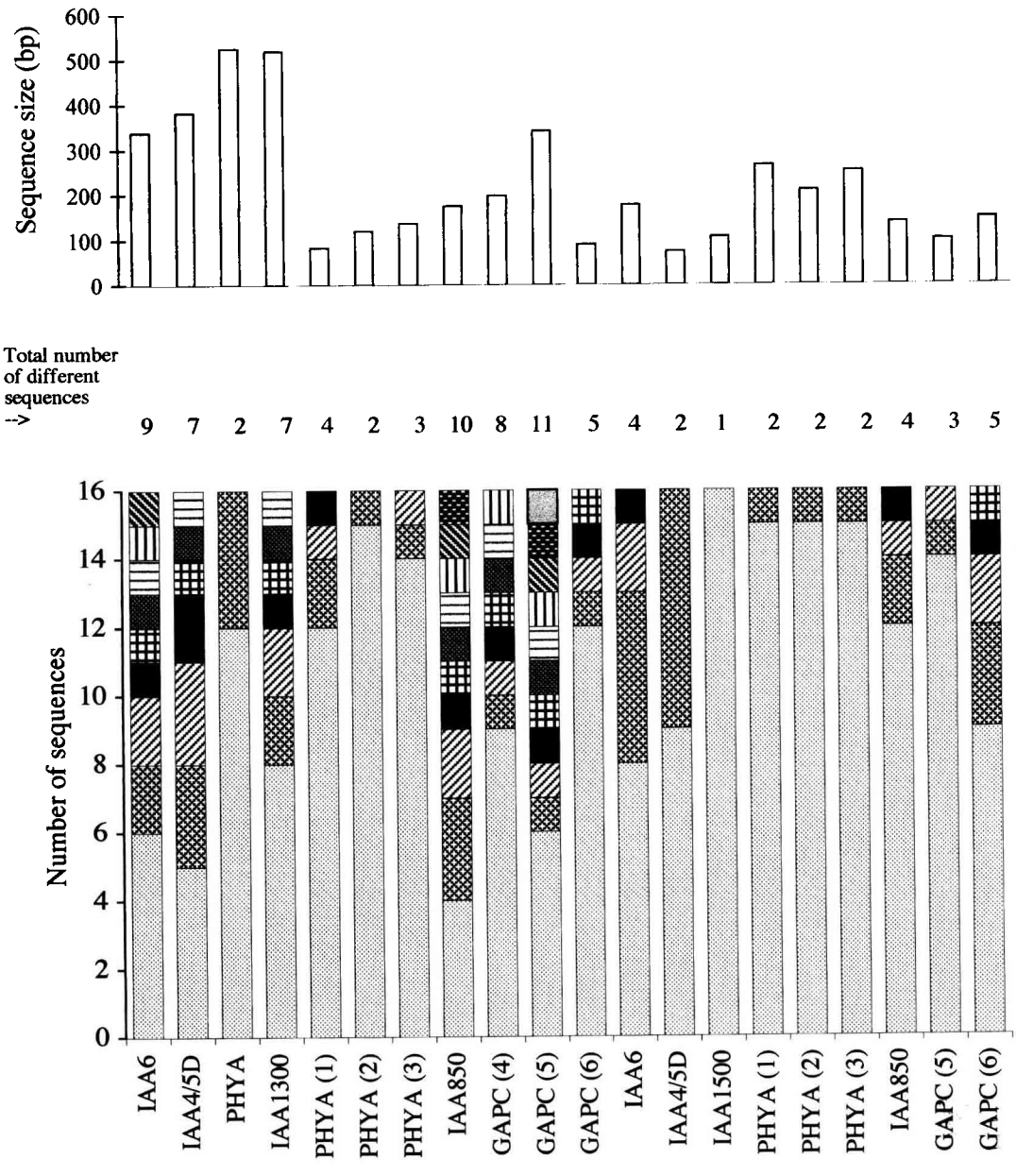

Figure 4. Polymorphism distribution between sequences and between genotypes. The upper graph represents the fragment sizes. In the lower graph, each symbol represents an allele and the height of each box is proportional to the number of sequences of one type. The number in parenthesis indicates the exon or intron in the considered gene. The 5'UTR sequence in *PHYA* includes all sequences situated before the ATG initiation site. We define 16 sequence types at each locus (Térèse, Torsdag, 661, Erygel, JI281, JI399, JI15, Champagne1, Champagne2, Chine, *P. abyssinicum*, *P. palestinicum1*, *P. palestinicum2*, *P. transcaucasicum*, *P. jomardi*, *P. fulvum*). For instance, the sequence which was performed in the *IAA4/5D* exon is 74 bp long and we observed two alleles (nine ecotypes being of one type and eight ecotypes of the other type).

along the sequences. *GAPC* and the *IAA850* sequences both include a region whose mutation frequencies are higher than the average frequency observed for the whole sequence. For example, 12 mutations are observed in a region of 37 bp in the intron of *IAA850*.

In our sample, we note a heterogeneous degree of polymorphism between genes. Two groups of sequences can be distinguished: *PHYA* and *IAA1300* show few differences, while *GAPC* and *IAA850* show much more polymorphism. The *PHYA* gene has been difficult to map because of a lack of polymorphism within its intron sequences. However, a single polymorphic site, consisting of an insertion/deletion, has been found in its 5'UTR sequence. In all the other genes, genetic markers were easily found in introns and even some exons generated variability. With a substitution frequency of 0.1, the intron 5 in *GAPC* and the intron in *IAA850* show more polymorphism than the others. Nevertheless, 23 of the 35 observed mutations in intron 5 in *GAPC* occur only in one genotype and more frequently in *P. jomardi*. Variations in the *IAA850* intron occurred at the borders of a A stretch.

4. DISCUSSION

All analysed genes have been successfully mapped. In order to develop a protocol for designing gene specific PCR markers, the first step should be to target intron sequences. However, we show that exonic polymorphism must not be underestimated.

4.1. Polymorphism rates in pea gene sequences

Our results suggest a variation in the polymorphism rate and distribution between and within genes. This heterogeneity was expected in the exons because of the selection pressure exerted on coding regions, whereas intron polymorphism was expected to be higher and more homogeneous between genes. Further studies including the analysis of more genes are needed to improve this description of sequence polymorphism in pea genes.

The difference between *PHYA* and the other sequences is striking. In this more complicated situation where introns are monomorphic, as well as in genes having no intron, one can ask which will be the technical solution. The frequency of genes which have no introns is unknown in pea, but progress made in the sequencing of the *Arabidopsis* genome is revealing a growing number of such genes. Considering our results, two strategies could be made for the use of the presented technique, based on exons or on 5'UTR sequences.

Although we do not have enough data to estimate the exonic polymorphism in pea, four of six analysed loci can be localised on at least two genetic maps thanks to the polymorphism that has been found in exons. The investigation for enzyme restriction sites which discriminate lines used in genetic mapping reveals that four of the six analysed genes could have been mapped with only the knowledge of the exonic sequence.

There has been one other report of genetic markers based on 5'UTR sequence polymorphism in plants [12]. In their study, the authors show the presence of repeated sequences which were reminiscent of microsatellites in the Stilbene synthase-chalcone synthase gene in *Vitis vinifera*. Interestingly, we encounter the same type of polymorphism in the promoter sequence of the *PHYA* gene in pea. Variability in 5'UTR sequences is attractive in order to create new markers. Nevertheless, their sequences are not available in GenBank for every gene and their utilisation to develop markers would in some cases require to clone the promoter.

4.2. Heterogeneity detected in the ecotypes Champagne and *P. sativum* ssp. *palestinicum*

With the primers amplifying the *PHYA* and *IAA6* genes, the co-amplification of the two fragments

with the DNA from the plants Champagne can be interpreted in two ways. Either they are the two allelic forms of the same gene in a heterozygous plant, or they are forms of two nearly identical and tightly linked duplicated genes (since these genes have been easily localised on the pea genetic map). Concerning *PHYA* (which is single copy in the pea genome according to Sato [27]), the heterogeneity of the PCR products reflects heterozygosity at that locus since we detect, in the Champagne population, plants which are homogeneous for one or the other PCR product. With the primers amplifying the *PHYA* gene, two fragments were amplified from the DNA of two plants and a single fragment was amplified from the DNA of the two other analysed plants. For the *IAA6* gene, a single band is amplified with the DNA from three plants and two bands are amplified with the DNA from the fourth plant. Therefore, a complementary genetic analysis of the selfing progeny of that plant is required to determine whether it is heterozygous at the *IAA6* locus.

4.3. Polymorphism detection tools

The polymorphism described here relies on DNA conformation within a polyacrylamide gel. The different migration rates of DNA fragments of the same size may depend on physical properties of DNA. Phenolic treatment of the PCR products did not affect the differences observed in migration rates, indicating that an interaction between DNA and proteins in the *Taq* polymerase preparation cannot explain the observed conformational polymorphism. These migration anomalies of double strand DNA are well recognised but they are not often used for the development of genetic markers. For example, although the utility of DSCP to detect DNA sequence polymorphism has already been proved by using introns as a source of variability in bovine genes [17], until now no examples have been described in plants.

Even if this kind of polymorphism is particularly interesting for differentiating alleles, the polymorphism detected is underestimated. Indeed, some sequence differences do not generate any differences in migration. In the case of the *IAA6* gene only, the substitutions existing in the 5'UTR sequence between cv. Térèse and cv. Torsdag do not modify the migration of the PCR products in the tested conditions.

On most occasions, a single band was visualised after PAGE. Whenever two alleles were simultaneously amplified, three bands were observed corresponding to homoduplex and heteroduplex DNA. A strategy involving either DSCP (as described here) or SSCP methods enables the detection of the markers with a high level of sensitivity. Simplification of the currently used SSCP protocol (using ethidium bromide staining for example) should make this method attractive for genotyping in pea. The sequences of the different alleles will provide the basis for the development of molecular tools which will enable routine detection of any DNA polymorphism. CAPS, allele specific PCR or the combined chain reaction [3] are three systems which do not require any polyacrylamide gel electrophoresis.

4.4. Advantages of gene-based sequence tagged site markers

All the analysed genes were successfully localised on one map. As a consequence, the markers described could be integrated in the establishment of a map of known function sequences in pea. Gilpin et al. [13] have recently published the mapping of 18 cloned sequences of known function in pea by revealing CAPS or RFLP polymorphism. One of these sequences is related to the *IAA* family. Sequence alignments show a high sequence similarity between this sequence (accession AA427337) and the *IAA850* sequence (accession AF026531) (97.6 % identity in a 254 bp overlap). Both sequences are linked to the *LegJ* gene. They probably correspond to the same locus unless they are two tightly linked (duplicated?) sequences. Concerning the *GAPC* sequence, localisation of the markers seems ambigous between the two genetic maps ([13] compared to our results). The co-locali-

sation of the sequence on other pea genetic maps should be confirmed (and could be validated by using the CAPS marker described which is polymorphic between 661 and Erygel and between JI281 and JI399).

Our study was preliminary and was conducted to develop a protocol for designing genetic markers which are based on gene sequences and PCR. It is now possible to map gene sequences involved in agronomic traits. In the plant materials studied, the mapping population derived from the crosses JI281 × JI399 or JI399 × JI15 are the most appropriate from this perspective.

Sequences which are available in GenBank could be used for genetic mapping even if only the coding sequences are known, since in some cases exon polymorphism is sufficient. Whenever homologous genes have been sequenced in other species, knowledge concerning intron positions in other species can be used as they are often conserved between species.

One noteworthy property of the described system is the ability to transfer markers from one map to another, since each map has been established with mapping populations which were obtained with different parental lines. More than 1 000 markers or polymorphic genes are available, but they are generally mapped with a single cross and integrating these maps to a single consensus map underlined the difficulty in finding common markers [31]. The molecular markers described will be of general interest for integrating all existing maps in pea as they are homologous PCR-based markers.

One extension we can imagine for these markers is an extrapolation to other plant species. The primers designed on the *GAPC* sequence efficiently amplified faba bean DNA and the identity of the PCR product was validated by sequencing. It will be possible in many cases to map the same genes in several species by using species specific primers or degenerate primers. Thus, gene markers identified in pea could be useful for all the leguminous species. Moreover, Strand et al. [29] have already published the possibility of designing universal degenerate consensus PCR primers enabling the amplification of homologous genes in diverse plant taxa (monocots and dicots) thus demonstrating the broad taxonomic usefulness of this kind of marker [29]. Similar PCR-based markers which are common to several species were also developed in animals in order to integrate mammalian genetic maps [22] and called comparative anchor tagged sequence (CATS). The genes which have been sequenced in *Arabidopsis* and rice may provide enough data concerning conserved domains to develop consensus markers. The rapid growth of the sequence databases promises in the near future the development of CATS markers in plants.

Acknowledgements: We would like to thank M. Ambrose (John Innes Centre) and C. Rameau (Inra) for the mapping populations and A. Burghoffer (Inra) for the ecotypes they provided. We are very grateful to N. Ellis (John Innes Centre) for the mapping analysis and for helpful comments. We are indebted to E. Téoulé for her advice and constructive criticisms and to G. Fouilloux for his help in analysing the data. We wish also to thank A. Wilson and I. Small for critical reading of the manuscript.

REFERENCES

[1] Abel S., Nguyen M.D., Theologis A., The *PS-IAA4/5*-like family of early auxin inducible mRNAs in *Arabidopsis thaliana*, J. Mol. Biol. 251 (1995) 533–549.

[2] Adamczyk J.R., Sylvain J.F., Pashley Prowell D., Intra- and interspecific DNA variation in a sodium channel intron in *Sodoptera* (Lepidoptera: Noctuidae), Ann. Entomol. Soc. Am. 89(6) (1996) 812–821.

[3] Bi W., Stambrook P.J., CCR: A rapid and simple approach for mutation detection, Nucleic Acids Res. 25 (1997) 2949–2951.

[4] Causse M., Santoni S., Damerval C., Maurice A., Charcosset A., Deatrick J., de Vienne D., A composit map of expressed sequences in maize, Genome 39 (1996) 418–432.

[5] Côrte-Real H.B.S., Dixon D.R., Holland P.W.H., Intron-targeted PCR: a new approach to survey neutral DNA polymorphism in bivalve populations, Marine Biol. 120 (1994) 407–413.

[6] D'Ovidio R., Tanzarella, Porceddu, Rapid and efficient detection of genetic polymorphism in wheat through amplification by polymerase chain reaction, Plant Mol. Biol. 15 (1990) 169–171.

[7] de Barros Lopes M., Soden A., Henschke P.A., Langridge P., PCR differentiation of commercial yeast strains using intron splice site primers, Appl. Environ. Microbiol. 62 (1996) 4514–4520.

[8] Devereux J., The GCG Sequence Analysis Software Package, Version 6.0, Genetics Computer Group, Inc., University Research Park, 575 Science drive, Suite B, Madison, 53711, USA, 1989.

[9] Dirlewanger E., Isaac P.G., Ranade S., Belajouza M., Cousin R., de Vienne D., Restriction fragment length polymorphism analysis of loci associated with disease resistance genes and developmental traits in *Pisum sativum* L, Theor. Appl. Genet. 88 (1994) 17–27.

[10] Doyle J.J., Doyle J.L., Isolation of plant DNA from fresh tissue, Focus 12 (1988) 13–15.

[11] Ellis T.H.N., Turner L., Hellens R.P., Lee D., Harker C.L., Enard C., Domoney C., Davies D.R., Linkage maps in pea, Genetics 130 (1992) 649–663.

[12] Geuna F., Hartings H., Scienza A., Discrimination between cultivars of *Vitis vinifera* based on molecular variability concerning 5' untranslated regions of the *StSy-CHS* genes, Theor. Appl. Genet. 95 (1997) 375–383.

[13] Gilpin B.J., McCallum J.A., Frew T.J., Timmerman-Vaughan G.M., A linkage map of the pea (*Pisum sativum* L.) genome containing cloned sequences of known function and expressed sequence tags (ESTs), Theor. Appl. Genet. 95 (1997) 1289–1299.

[14] Hall K.J., Parker J.S., Ellis T.H.N., Turner L., Knox M.R., Hofer J.M.I., Lu J., Ferrandiz C., Hunter P.J., Taylor J.D., Baird K., The relationship between genetic and cytogenetic maps of pea. II. Physical maps of linkage mapping populations, Genome 40 (1997) 755–769.

[15] Hofer J., Turner L., Hellens R., Ambrose M., Matthews P., Michael A., Ellis N., *UNIFOLIATA* regulates leaf and flower morphogenesis in pea, Curr. Biol. 7 (1997) 581–587.

[16] Kersanach R., Brinkmann H., Liaud M.F., Zhang D.X., Martin W., Cerff R., Five identical intron positions in ancient duplicated genes of eubacterial origin, Nature 367 (1994) 387–389.

[17] Kirkpatrick B.W., Hart G.L., Conformation polymorphisms and targeted marker development, Anim. Genet. 25 (1994) 77–82.

[18] Konieczny A., Ausubel F.M., A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers, The Plant Journal 4 (1993) 403–410.

[19] Kurata N., Nagamura Y., Yamamoto K., Harushima Y., Sue N., Wu J., Antonio B., Shomura A., Shimizu T., Lin S., Inoue T., Fukuda A., Shimano T., Kuboki Y., Toyama T., Miyamoto Y., Kirihara T., Hayasaka K., Miyao A., Monna L., Zhong H., Tamura Y., Wang Z., Momma T., Umehara Y., Yano M., Sasaki T., Minobe A., 300 kilobase interval genetic map of rice including 883 expressed sequences, Nat. Genet. 8 (1994) 365–372.

[20] Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lincoln S.E., Newburg L., MAPMAKER: An interactive computer package for constructing primary genetic maps of experimental and natural populations, Genetics 116 (1987) 174–191.

[21] Lu J., Knox M.R., Ambrose M.J., Brown J.K.M., Ellis T.H.N., Comparative analysis of genetic diversity in pea assessed by RFLP- and PCR-based methods, Theor. Appl. Genet. 93 (1996) 1103–1111.

[22] Lyons L.A., Laughlin T.F., Copeland N.G., Jenkins N.A., Womack J.E., O'Brien S.J., Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes, Nat. Genet. 15 (1997) 47–56.

[23] Orita M., Suzuki Y., Sekiya T., Hayashi K., Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction, Genomics 5 (1989) 874–879.

[24] Palumbi S.R., Baker C.S., Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales, Mol. Biol. Evol 11(3) (1994) 426–435.

[25] Rychlik W., Rhoads R., A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA, Nucleic Acids Res. 17 (1989) 8543–8551.

[26] Samec P., Nasinec V., Detection of DNA polymorphism among pea cultivars using RAPD technique, Biologia Plantarum 37 (1995) 321–327.

[27] Sato N., Nucleotide sequence and expression of the phytochrome gene in *Pisum sativum*: differential regulation by light of multiple transcripts, Plant Mol. Biol. 11 (1988) 697–710.

[28] Slabaugh M.B., Huestis G.M., Leonard J., Holloway J.L., Rosato C., Hongtrakul V., Martini N., Toepfer R., Voetz M., Schell J., Knapp S.J., Sequence-based genetic markers for genes and gene families: single-strand conformational polymorphisms for the fatty acid synthesis genes of *Cuphea*, Theor. Appl. Genet. 94 (1997) 400–408.

[29] Strand A.E., Leebens-Mack J., Milligan B.G., Nuclear DNA-based markers for plant evolutionary biology, Mol. Ecol. 6 (1997) 113–118.

[30] Thomas W.K., Kocher T.D., Sequencing of polymerase chain reaction- amplified DNAs, Meth. Enzymol. 224 (1993) 391–399.

[31] Weeden N.F., Swiecicki W.K., Timmerman-Vaughan G.M., Ellis T.H.N., Ambrose M., The current pea linkage map, Pisum Genetics 28 (1996) 1–4.

[32] Weeden N.F., Wolko B., Linkage map for the garden pea (*Pisum sativum*), in: S.J. O'Brien Cold Spring Harbor Laboratory (Ed.) Genetic Maps. Locus Maps of Complex Genomes, ed. 5, Cold Spring Harbor, N.Y., 1990, pp. 6.106–6.112.