Theor Appl Genet (2008) 117:413–424
DOI 10.1007/s00122-008-0785-4

ORIGINAL PAPER

Genetic diversity and population structure of pea (*Pisum sativum* L.) varieties derived from combined retrotransposon, microsatellite and morphological marker analysis

Petr Smýkal · Miroslav Hýbl · Jukka Corander · Jiří Jarkovský · Andrew J. Flavell · Miroslav Griga

Received: 5 October 2007 / Accepted: 2 May 2008 / Published online: 27 May 2008

Abstract One hundred and sixty-four accessions representing Czech and Slovak pea (*Pisum sativum* L.) varieties bred over the last 50 years were evaluated for genetic diversity using morphological, simple sequence repeat (SSR) and retrotransposon-based insertion polymorphism (RBIP) markers. Polymorphic information content (PIC) values of 10 SSR loci and 31 RBIP markers were on average high at 0.89 and 0.73, respectively. The silhouette method after the Ward clustering produced the most probable cluster estimate, identifying nine clusters from molecular data and five to seven clusters from morphological characters. Principal component analysis of nine qualitative and eight quantitative morphological parameters explain over 90 and 93% of total variability, respectively, in the first three axes. Multidimensional scaling of molecular data revealed a continuous structure for the set. To enable integration and evaluation of all data types, a Bayesian method for clustering was applied. Three clusters identified using morphology data, with clear separation of fodder, dry seed and *afila* types, were resolved by DNA data into 17, 12 and five subclusters, respectively. A core collection of 34 samples was derived from the complete collection by BAPS Bayesian analysis. Values for average gene diversity and allelic richness for molecular marker loci and diversity indexes of phenotypic data were found to be similar between the two collections, showing that this is a useful approach for representative core selection.

Communicated by D. A. Hoisington.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-008-0785-4) contains supplementary material, which is available to authorized users.

P. Smýkal (✉) · M. Hýbl · M. Griga
Plant Biotechnology Department, Agritec Plant Research Ltd,
Zemědělská 2520/16, 787 01 Šumperk, Czech Republic
e-mail: smykal@agritec.cz

M. Hýbl
Grain Legumes Department, Agritec Plant Research Ltd,
Zemědělská 2520/16, 787 01 Šumperk, Czech Republic

J. Corander
Department of Mathematics and Statistics,
Åbo Akademi University, P.O. Box 68,
Turku 20500, Finland

J. Jarkovský
Institute of Biostatistics and Analysis,
Masaryk University, Kamenice 3, 625 00 Brno,
Czech Republic

A. J. Flavell
Plant Research Unit, University of Dundee at SCRI,
Invergowrie, Dundee DD2 5DA, Scotland, UK

Introduction

The demand for productivity and homogeneity in crops has resulted in a limited number of standard, high-yielding varieties, at the price of the loss of heterogeneous traditional local varieties (landraces), a process known as genetic erosion. Landraces and older crop varieties preserve much of this lost diversity and comprise the genetic resources for breeding new crop varieties to cope with environmental and demographic changes (Esquinas-Alacazar 2005). To prevent the extinction of such genotypes, ex situ conservation of germplasm resources was pioneered by Vavilov (1926) and nowadays, germplasm collections hold over 6 million crop plant accessions world-wide.

The study of genetic diversity for both germplasm management and breeding has received much attention,

especially following the introduction of the core collection concept by Frankel and Brown (1984). For legumes, core collections have been defined using various strategies, varying from random and stratified sampling strategies (Erskine and Muehlbauer 1991) to the use of evolutionary, agroecological and/or molecular data (Tohme et al. 1995; Baranger et al. 2004). Morphological descriptors are widely used in defining germplasm groups and remain the only legitimate marker type accepted by the international union for the protection of new varieties of plants (UPOV). Morphological traits represent the action of numerous genes and thus contain high information value but can be unreliable owing to a strong influence of the environment. In contrast, molecular markers accurately represent the underlying genetic variation and now dominate the genetic diversity field. Initially, storage proteins and isozymes (Brown and Weir 1983) were applied to assess diversity but these do not provide sufficient polymorphism and can suffer from tissue and environment influence. Therefore, DNA-based descriptors of genetic diversity have largely superseded protein-based methods. A variety of DNA marker methods have been widely used for diversity analysis in plants, including randomly amplified polymorphic DNA (RAPD) markers (Williams et al. 1990), inter-simple sequence repeat (ISSR; Zietkiewicz et al. 1994), amplified fragment length polymorphism (AFLP; Vos et al. 1995) and simple sequence repeat (SSR; Beckmann and Soller 1990).

For the analysis of pea diversity, SSRs have been popular because of their high polymorphism level and information content, co-dominance and good reproducibility (Burstin et al. 2001; Ford et al. 2002; Baranger et al. 2004; Loridon et al. 2005). A potential problem in using SSRs for characterising highly diverse germplasm is homoplasy, associated with the high mutation rate and the possibility of back-mutation exhibited by this marker type. Marker systems based on retrotransposon insertion polymorphism have also been widely used for phylogeny and genetic relationship studies in pea. Sequence-specific amplification polymorphism (SSAP; Waugh et al. 1997) is a multiplex approach that reveals large numbers of polymorphic insertions in a single gel assay (Ellis et al. 1998). An alternative is inter-retrotransposon amplified polymorphism (IRAP; Kalendar et al. 1999; Smýkal 2006) which requires only a simple PCR, using retrotransposon primer(s), followed by gel analysis. Both methods, however, suffer from the dominant nature of detection and problems with reproducibility between experiments. Retrotransposon-based insertional polymorphism (RBIP) avoids these drawbacks by scoring both presence and absence of individual insertions, with a combination of retrotransposon-specific and flanking host-specific primers (Flavell et al. 1998). RBIP is more accurate for studies of deeper phylogeny in highly diverse germplasm owing to its co-dominant nature and has been adapted to a high throughput microarray format (Flavell et al. 2003; Jing et al. 2005, 2007).

Improvements in marker methods for revealing genetic diversity have been accompanied by corresponding refinements in computational methods to convert raw marker data into useful representations of diversity. Distance-based methods were initially used (Reif et al. 2005). Bayesian approaches offer an alternative for germplasm genetic structure assessment (Pritchard et al. 2000; Falush et al. 2003; Corander et al. 2004, 2007; Maccaferri et al. 2005) and the incorporation of probability, measures of support and complex model and data character processing (Beaumont and Rannala 2004), makes them more attractive.

The aim of this study was to investigate genetic diversity in a selection of pea assessions, which have been used in Czech and Slovak breeding over the last ca. 50 years. We have combined morphological qualitative and quantitative characters, with RBIP and SSR markers and tested a variety of clustering approaches to reveal the diversity of the sample set and suggest the composition of a working core collection to faithfully represent this germplasm.

Material and methods

Plant materials

One hundered and sixty-four *Pisum sativum* sp. *sativum* accessions were obtained from the pea collection of the Czech gene bank held in AGRITEC Ltd. (Electronic supplementary material S1). Cultivars Gotik, Alan, Adept and Bohatyr were included as controls for quantitative traits.

Analysis of morphological descriptors

Plants were grown in field trials in 2004 and 2005 at Šumperk (Czech Republic) on orthic luvisol soils at 315 m altitude, with long term average temperature of 8°C and long term rainfall of 693 mm. The trials used a randomised complete block design with three replicates. Thirty-three characters (for stem, leaflets, stipules, flowers, pods and seeds) were evaluated (ESM S3a, b) according to the descriptor list of genus *Pisum* L. (Pavelková et al. 1986). All plants chosen for DNA extraction were first described morphologically.

DNA isolation

Young leaves from ten randomly chosen plants per accession were bulked together and stored at −80°C until DNA isolation. Genomic DNAs were manually isolated by a modified CTAB method (Smýkal 2006). DNAs obtained

from approximately 100 mg fresh weight leaf material per accession were resuspended in 300 µl of TE buffer at concentration of 50–100 ng/µl and stored at –20°C until use.

DNA marker analysis

SSR primer pairs (Table 1) were selected from Ford et al. (2002) and Loridon et al. (2005). PCRs and gel analysis were performed as described in Smýkal et al. (2007). RBIP analysis was performed according to Flavell et al. (2003), with the exception that BioTools Taq DNA polymerase (BioTools S.A., Madrid, Spain) was used. The following 31 RBIP primers pairs were selected from Jing et al. (2005): Birte-B1, Birte-x5, Birte-x16, Birte-x28, MKRBIP3, MKRBIP4, MKRBIP7, 1006-x19, 1006nr27, 1006nr13, 399-14-9, 45x31, 64x45, 281x5, 281x40, 281x16, 281x44, 2055nr1, 2055nr23, 95x2, 1794x35, 2055nr36, 2055nr51, 1794-2, 399-80-46, 1794-1, 2385x23, 2385x64, 2201Cyc6, 1074Cyc12, 1074cyc29. PCRs, as described in Jing et al. (2005), were resolved by electrophoresis on 1.5 or 2% agarose-TBE gels (Serva, Heidelberg, Germany) using UV-visualised ethidium bromide staining.

Genetic similarity, cluster and structure analysis

SSR and RBIP scores were converted into binary data by presence (1) or absence (0) of the selected fragment. In the case of RBIP analysis, a fourth state, namely complete absence of any PCR product corresponding to primer site mutation (Jing et al. 2005) was added. Genetic similarity coefficients were calculated using the Jaccard index of similarity (Nei 1973, 1978; Reif et al. (2005) using SPSS 12 software (SPSS 2003). Polymorphic information content (PIC) was calculated for each marker using the following formula: $PIC_i = 1 - \sum P^2_{ij}$, where P_{ij} is the frequency of the jth allele in clone (i). For the visualisation of genetic data in factorial space, multidimensional scaling (MDS) based on similarity matrix of Jaccard coefficients, was adopted (Kruskal 1964). Morphological descriptors were analysed using principal component analysis (PCA). In a further analysis, the multivariate space of morphological descriptors was combined with genetic variability described by the multidimensional scaling (MDS). Cluster analysis was performed on the genetic similarity matrices by the method of Ward (1963) using Statistica for Windows 7.1. (StatSoft 2006). The silhouette method was applied for the identification of the optimal number of the most homogeneous clusters (Rousseeuw 1987). The resulting clusters were expressed as dendrograms. Goodness of fit was assessed by Mantel test (Mantel 1967) using NTSYS-pc version 2.2 (Rohlf 2006). The PopGene program (version 1.32.; Yeh and Boyle 1997) was used to calculate the following parameters: allele frequencies at each locus for complete and subdivided populations; gene diversity H value (Nei 1973), expected and observed homozygosities, population genetic distance expressed as Nei unbiased genetic distance (Nei 1978), F-statistic (Wright 1965; Reynolds et al. 1983) and Shannon index (Lewontin 1972).

Bayesian structure analysis

To investigate the genetic structure of the pea collection, the Bayesian method available in the BAPS software (Corander et al. 2004, 2007; Corander and Martiinen 2006) was used. Initial screening of the morphological characters revealed them to be highly informative. Therefore, the sample set was first clustered using the BAPS model for the discrete-valued traits. As a single clustering solution with three clusters was conclusively supported, the molecular data were subsequently analysed separately for each of

Table 1 List of SSR loci used in this study

SSR locus	Linkage group	Position (cM)	Number of alleles	Observed heterozygozity	Size range (bp)	Polymorphic information content (PIC)
AA-67	I	80.3	6	0.012	330–390	0.882
AD-186	II	36.2	8	0.147	220–320	0.961
AD-270	III	254.3	7	0.0	230–290	0.964
A-278	III	154.9	3	0.062	130–170	0.827
A-9	IV	62.1	3	0.073	330–380	0.886
AA-163	V	100.3	5	0.120	250–320	0.869
AD-141	VI	70.1	7	0.185	210–330	0.973
B-14	VII	113.9	4	0.017	430–470	0.929
AD-237	VII	152.1	7	0.073 1	220–360	0.934
AB-65	VII	94.1	3	0.0	1 140–180	0.697
Mean			5.3	0.069		0.892

Indicated linkage groups and map positions according to Loridon et al. (2005)

these three clusters. In all analyses the clustering was done using the model for non-linked markers and the estimation was performed using 30 replicate runs of the algorithm, with the a priori upper boundary for the number of clusters ranging between 10 and 40.

Core set analysis:

To test the utility of BAPS-based core set selection, genetic diversity parameters of selected accessions were analysed using FSTAT v2.9.3.2 (Goudet 1995) and PopGene v1.32. (Yeh and Boyle 1997). Gene diversity, allelic richness and fixation indices were computed. To compare morphological diversity, Shannon–Weaver Diversity Indexes were computed for each trait separately (Shannon and Weaver 1962) according to Nersting et al. (2006).

Results

Marker analysis and allelic richness of SSR and RBIP data

We surveyed 164 pea accessions using 10 SSR loci (Table 1). A total of 53 alleles were identified with a minimum three and maximum eight alleles per locus. Twelve rare alleles (22%) with frequencies below 0.05 were found at six SSR loci. Calculated PIC values were high, ranging from 0.697 to 0.964, with an average of 0.89. Heterogeneity, associated with the use of ten bulked plants per accession, was detected in 60 out of 164 accessions (37%) at eight loci, averaging 0.069. Analysis of individual plants in the case of 15 such accessions, indicated heterogeneity between plants rather than heterozygosity (data not shown).

The same sample set was then analysed with 31 retrotransposon RBIP markers. Sixteen of these detected polymorphism in the investigated germplasm set (Table 2), identifying 42 alleles. Ten RBIP loci repeatedly produced occasional zero scores (frequencies 0.011–0.35). Fourteen of the informative RBIPs detected residual heterogeneity, varying from 0.006 to 0.335 in 93 accessions (57%). Calculated PIC values ranged from 0.484 to 0.888, with an average of 0.730. Most RBIP loci displayed a balanced distribution across the 164 accessions, apart from 2201Cyc6, 1074Cyc12, 95x2 and MKRBIP4, where the occupied site allele dominated over the empty site (0.84–0.91).

Genetic relationships revealed by SSR and RBIP molecular markers

Pairwise genetic distances were calculated from Jaccard similarity coefficient for combined SSR and retrotransposon data. Ward hierarchical ascendant classification was then performed on the distance matrix and finally a dendrogram was constructed. The silhouette method, adopted after the Ward clustering (Ward 1963), identified nine clusters as the most probable estimate (ESM S2). Cluster I contains mainly fodder type accessions, cluster II contains five fodder and 23 dry-seed, clusters III, IV and VI contain only dry-seed varieties, cluster V contains 17 dry-seed and 4 fodder type, cluster VII contains 25 dry-seed and 1 fodder type, cluster VIII contains 11 dry-seed and 6 fodder type and cluster IX contains 1 dry seed and 16 fodder type varieties. Further inspection

Table 2 Polymorphic RBIP loci used in this study

RBIP-locus	Frequency of occupied site	Frequency of empty site	Null allele	Observed heterozygozity	Polymorphic information content (PIC)
MKRBIP-3	0.686	0.194	0.120	0.335	0.678
MKRBIP-4	0.843	0.157	0	0.421	0.484
MKRBIP-7	0.217	0.771	0.011	0.022	0.782
Birte-B1	0.737	0.263	0	0.137	0.694
Birte-x5	0.517	0.477	0.006	0.022	0.835
Birte-x16	0.906	0.060	0.034	0.034	0.724
1006-x19	0.677	0.294	0.028	0.146	0.819
399-14-9	0.457	0.543	0	0	0.748
45-x31	0.389	0.283	0.348	0.101	0.888
64-x45	0.546	0.437	0.017	0.006	0.836
281-x40	0.080	0.920	0	0.128	0.574
2055-nr51	0.651	0.349	0	0	0.727
95-x2	0.863	0.137	0	0.205	0.618
281-x44	0.280	0.714	0.006	0.165	0.803
2201Cycl-6	0.911	0.049	0.040	0.084	0.722
1074Cycl-12	0.869	0.046	0.086	0.053	0.745
Mean				0.116	0.730

revealed that 33 out of 49 fodder types (67%) are found in clusters I and IX, clusters IV, V and VI contain mostly older varieties (registered up to 1975) and cluster VII contains largely modern varieties bred after the 1980s, including all *afila* type accessions. Based on combined RBIP and SSR data, the Nei genetic distances were 0.0689 and 0.1401, respectively for fodder and dry-seed type groups.

Cluster analysis using only RBIPs placed about one third of fodder pea accessions in the same group as field peas, while combining SSR and RBIP data clustered 67% of fodder pea accessions into two of the nine clusters (data not shown). In the case of RBIP markers, no specific allele is linked to seed type, but 5 RBIP loci showed altered frequencies of occupied/empty sites. RBIP null alleles were detected in seven loci in the case of fodder-types and nine in case of dry-seed types (data not shown). Average detected heterogeneities for RBIP and SSR loci were 0.15 and 0.05, respectively, for fodder and 0.06 and 0.08, respectively, for dry-seed types. SSR-based cluster analysis revealed both quantitative and qualitative differences from the RBIP results, with five alleles (<0.05 frequency) at three SSR loci being specific for dry-seed pea, while three alleles at three SSR loci were specific for fodder pea.

Frequency calculations for all SSR and RBIP marker-based distances of the entire data set resulted in a column graph (Fig. 1a) with a normal-like distribution in the range of 0.2–1.0. To reveal another level of structure for the collected sample set, multidimensional scaling (MDS) was performed on the SSR and RBIP data (Fig. 2). This identified a broad, continuous variation for the pea sample set, with no clear outgroup. Fodder pea types are confined largely to the bottom right sector and dry-seed pea accessions are spread across the plot, indicating exchange of genetic material during the breeding process.

To compare genetic diversity in relation to breeding period, the sample set was divided into three sub-sets: the first consists of 91 older varieties and landraces established before 1950, the second comprises 19 samples from 1960 to 1970 and the third contains 54 modern varieties (since 1980). Although differences in allele frequencies both for RBIP (Table 3) and SSR (Fig. 3) were encountered, no statistically significant differences were observed between the three periods using DNA marker data (not shown).

Morphology-based characterisation of pea diversity

The distribution of 15 qualitative and 18 quantitative morphological characteristics across the pea germplasm set is presented in Supplementary material S3a, b. Calculation of Euclidean distances derived from these data resulted in Fig. 1b. The distribution lies in the range of 1.5–16.5 with two peaks discernible at 6–7 and 12–13.

Correlations between morphological characters were examined next. Numerous traits were strongly correlated,

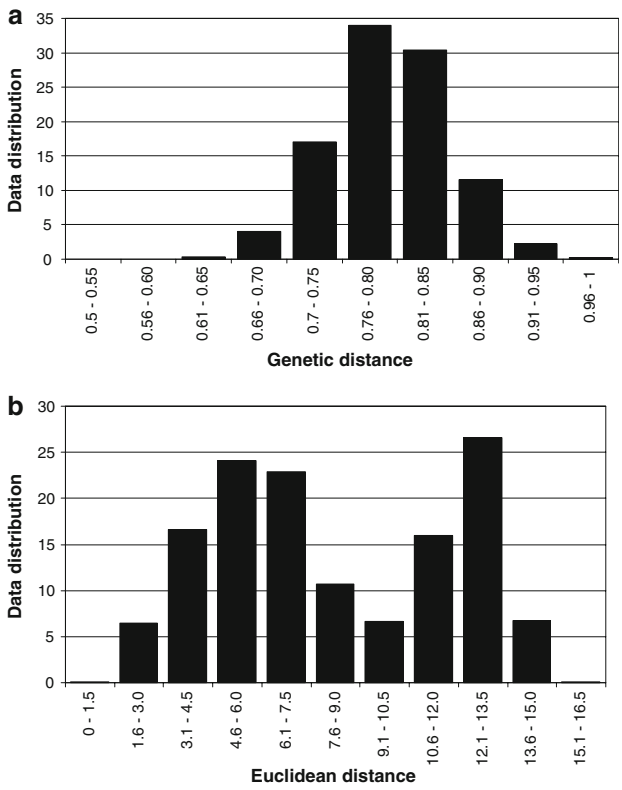

Fig. 1 Frequency calculation of distances for molecular and morphological markers. **a** SSR and polymorphic RBIP markers. Values are expressed as ($\sqrt{}$1-Jaccard similarity coefficient) on the *x* axis. **b** Euclidean distances for qualitative and quantitative morphological characters

for example stipules-character of anthocyan spot with flower-*vexillum* colour ($r = 0.91$), flower-wings colour ($r = 0.94$), seed-colour at full ripeness ($r = 0.65$) and seed-testa colour ($r = 0.64$). Nine out of the fifteen qualitative traits were used to estimate phenotypic diversity by PCA (ESM S4a). More than 90% of the total variation of qualitative traits was explained three PCs (43.85, 36.64, and 5.86%), based on the nine qualitative eigenvectors. The flower characters of anthocyan spotted stipules and colour of flower wings and wexillum, were the eigenvectors with high positive loading for PC1. The leaf characters leaflet colour, shape, shape of leaflet apex and type of leaf, were components of PC2. Colour of seed testa had high positive loading, whilst seed colour at full ripeness had high negative loadings on PC3. Significant correlations were found also between quantitative characters. For example, seed number per plant correlated closely with pod number per plant ($r = 0.90$), thousand seed weight (TSW; $r = -0.65$) and seed weight per plant ($r = 0.70$).

Eight of the 18 quantitative traits were then used for evaluation of morphologic diversity. PCA of quantitative traits included 93% of total variation in four PCs (ESM S4b). The most important eigenvectors for PC1 were seed and pod number per plant, together with length of stem and length of

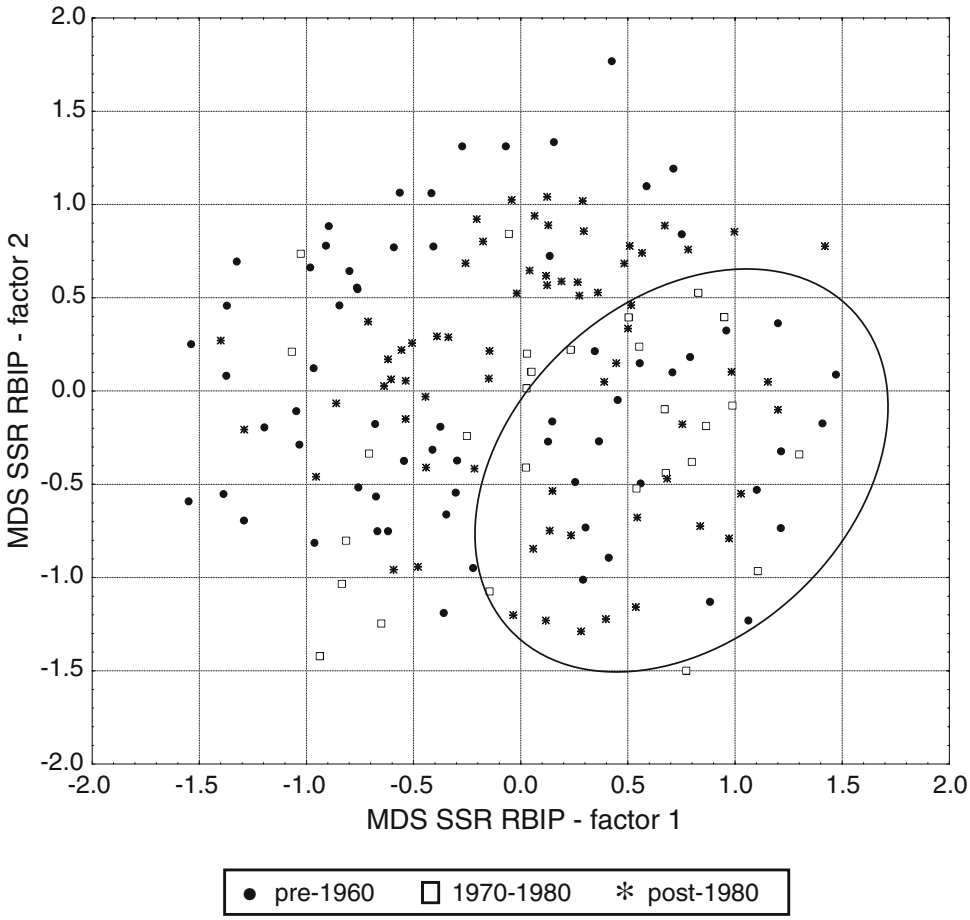

Fig. 2 Multidimensional scaling (MDS) for combined molecular data. Three breeding periods (pre-1960, 1970–1980, post-1980) are shown in *blue circle*, *red square* and *green triangle*. Fodder pea accessions are indicated as a cluster

Table 3 Frequency distributions of RBIP occupied site alleles according to breeding period (pre-1960, 1970–1980, post-1980)

RBIP locus	pre-1960	1970–1980	post-1980
MKRBIP-3	0.597	0.65	0.851
MKRBIP-4	0.792	0.833	0.904
MKRBIP-7	0.149	0.267	0.272
Birte-B1	0.721	0.717	0.779
Birte-x5	0.481	0.417	0.588
Birte-x16	0.851	0.967	0.941
1006-x19	0.643	0.433	0.823
399-14-9	0.506	0.4	0.441
45-x31	0.299	0.3	0.463
64-x45	0.487	0.871	0.471
281-x40	0.13	0.05	0.037
2055-nr51	0.5649	0.517	0.809
281-x44	0.282	0.25	0.265
2201Cycl-6	0.896	0.9	0.934
1074Cycl-12	0.831	0.867	0.911
Mean	0.57	0.57	0.64

stem to the first productive node. PC2 was positively defined by length of internode and stem under the first productive node. Length of stem and TSW and seed and pod numbers per plant were negatively defined for this PC. PC3 was positively influenced by length of internode under the first productive node, while high negative influence was noticed in the number of sterile nodes per stem. Seed weight per plant and TSW had high positive impacts in PC4.

The morphological characteristics were loaded into dummy variables and clustered using simple matching coefficients and Ward method (Ward 1963; ESM S5). The silhouette method revealed four clusters as the most homogeneous solution for morphological parameters, with three, five and six clusters also providing meaningful solutions.

Lastly, to compare the DNA-based and morphological diversity data against each other, the molecular data MDS was plotted against morphological PCA1 factor analysis (Fig. 4). This revealed a clearer separation of fodder pea accessions in the upper right part of the field and, again, no clear distinction between the three breeding periods.

Pea collection structure estimation using a Bayesian model-based approach

Bayesian model-based analysis was first applied to the morphological and molecular data separately. Analysis of molecular data partitioned the sample set into 29 clusters,

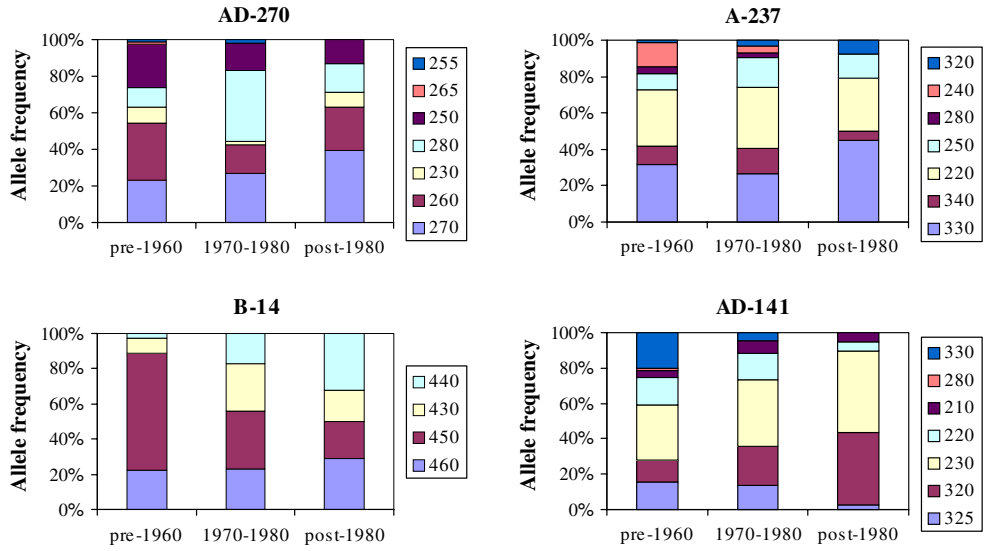

Fig. 3 Temporal changes in SSR allele frequencies in three breeding periods. Breeding periods are shown in Fig. 2. SSR markers used are AD-270, B-14, A-237 and AD-141. Allele sizes are in bp

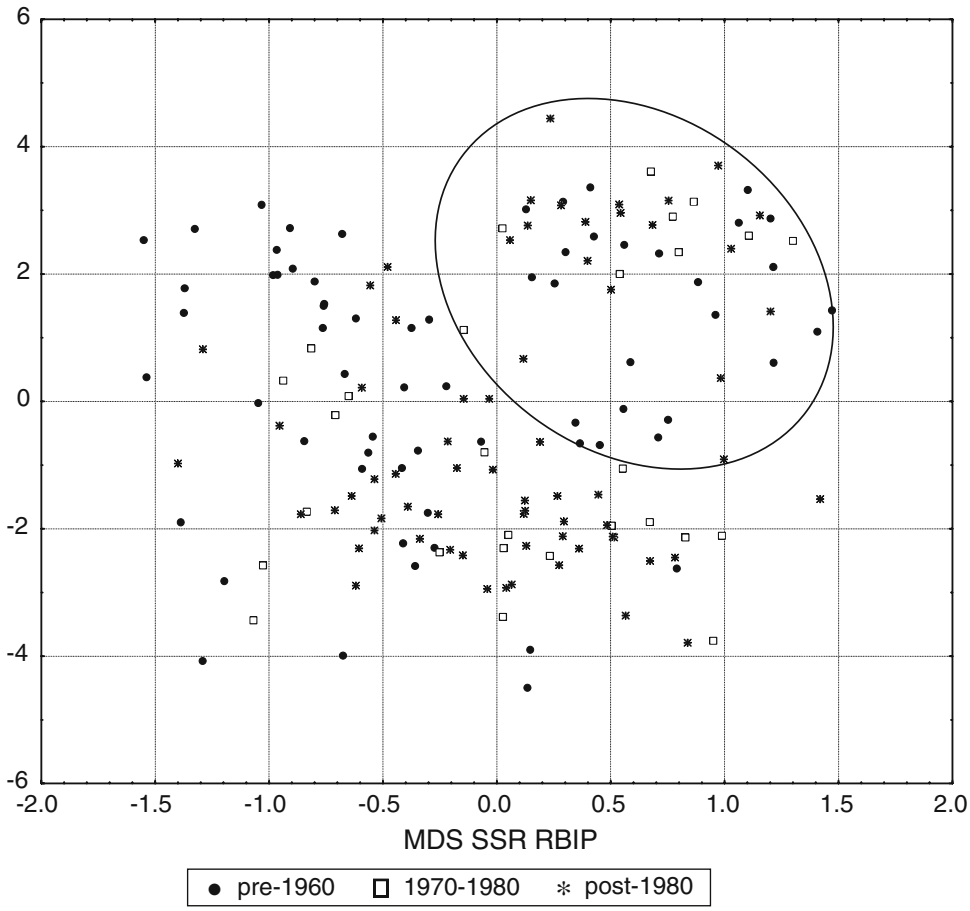

Fig. 4 Plot of molecular data MDS versus morphological PCA1 factor analysis. Breeding periods are indicated in Fig. 2. Fodder pea accessions are indicated as a cluster

with a log marginal likelihood value of optimal partition at −7184.9 and a probability of 0.948, showing high structuring of the set (data not shown). Eleven of these clusters contained nearly exclusively fodder type accessions, four others grouped 23 out of 47 dry-seed accessions collected before 1965, while 17 of the 28 most recent dry-seed varieties released after 1989 were separated into three clusters. The remaining clusters provided no clear

assignment of the accessions to either type or breeding period. In the case of morphological data, three or six clusters were found by optimal partitioning, with log marginal likelihood values of −14971.4 for six clusters and −17237.7 for three clusters. One cluster comprised 105 dry-seed plus two fodder varieties, a second cluster comprised 47 fodder accessions plus two dry-seed varieties and the third cluster comprised eight dry seed varieties of *afila* type. Therefore, partitioning into three clusters was accepted, with a probability of 1.0.

Combined analysis of all morphological descriptors and molecular markers (SSR and RBIP) resulted in three clusters, largely corresponding to those defined by the morphological data alone (data not shown). Consecutive analysis of morphological data, followed by subclustering based on molecular data, yielded 17 sub-clusters in cluster 1, comprising 107 dry-seed pea accessions with a probability of 0.938, 12 sub-clusters in cluster 2, comprising 49 fodder pea accessions with a probability of 0.475 and 5 sub-clusters in cluster 3 comprising 8 *afila* types with a probability of 0.724 (Fig. 5).

Formulation of a core collection based on molecular and morphological data

The final aim of this study was to formulate a core collection for the samples under study, using the combined diversity data. Using the Bayesian BAPS analysis of integrated data approach, a single accession per cluster was selected out of 34 clusters, to form a core set (ESM S1). To determine whether this core set is an adequate representation of the entire collection, the SSR and RBIP allele frequencies were compared with the morphological descriptor data. Due to the different natures of the RBIP and SSR data classes (three possible alleles for the former vs. multiple alleles for the latter), the two marker classes were analysed separately. Table 4 shows that both average gene diversity value and allelic richness per locus are similar for both molecular marker types between the core collection and the complete collection. These data indicate that the core collection represents very well the diversity of the complete collection.

A similar comparison between the core and complete germplasm sets was performed using all 15 qualitative morphological traits. Sixty-three out of 78 trait categories shown by the entire set are present in the core selection. Furthermore, average Shannon–Weaver values for the core set are comparable to the entire set (0.95 vs. 0.97), demonstrating good representation of the morphological diversity in the core set (Table 5).

Discussion

In this study we have examined the genetic diversity captured within 164 pea accessions originating from over 50 years of Czech and Slovak breeding activities, by scoring a combination of morphological and two different DNA-based molecular characters.

Pea germplasm genetic diversity assessment using SSRs and RBIPs

Two different codominant molecular marker methods have been used to assess pea genetic diversity, namely SSR and RBIP. Only 16 of the 28 RBIP insertions studied here display polymorphism in our collection. We suggest that this is due to the lower diversity of the sample set studied, relative to the John Innes collection (JIC), the source of these markers. Our collection comprises mainly breeders' lines and varieties, whereas the JIC contains hundreds of wild and landrace samples, some of which were used in the

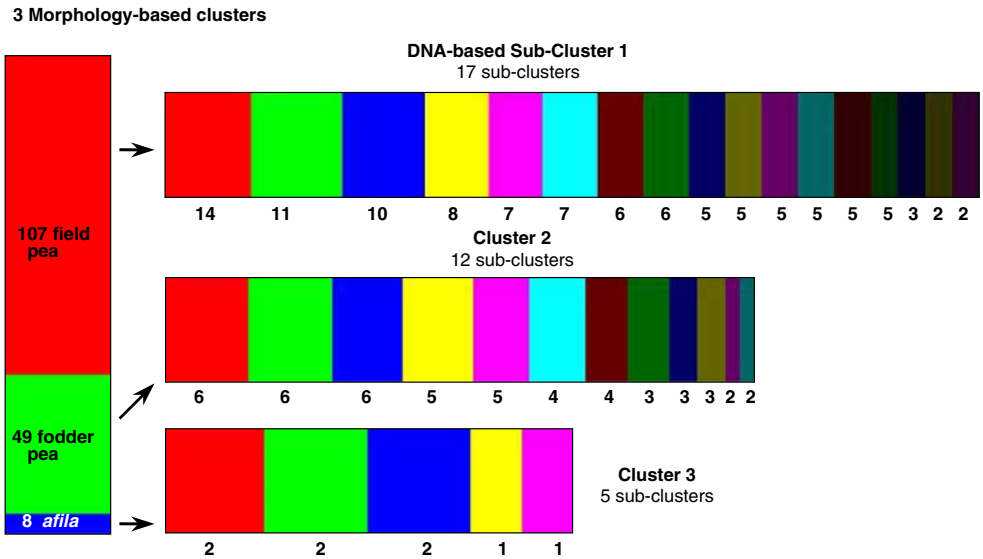

Fig. 5 Bayesian model-based analysis of morphological and molecular marker data. Three morphology-based clusters are further separated into 17, 12 and 5 DNA-based sub-clusters with indicated numbers of accessions

Table 4 Diversity data for entire germplasm set and BAPS-selected core set using SSR and RBIP markers

Entire collection (164)						Core set (34)				
Locus SSR	na	ne	I	Gene diversity	Allelic richness	na	ne	I	Gene diversity	Allelic richness
AD-270	7	4.52	1.59	0.783	6.38	5	3.8	1.46	0.809	6.18
AD-9	3	2.99	1.09	0.670	3.00	3	2.6	1.03	0.674	3.00
B-14	4	3.19	1.26	0.691	4.00	4	3.8	1.36	0.681	4.00
AD-237	7	4.18	1.60	0.764	6.72	5	3.2	1.34	0.730	6.66
AA-278	5	2.25	0.98	0.560	3.85 1	3	1.70	0.67	0.630	3.85
AD-141	8	4.05	1.60	0.757	7.49	5	2.59	1.12	0.723	7.26
AB-65	3	1.17	0.32	0.149	2.96	3	1.59	.32	0.121	2.95
AD-186	6	3.42	1.43	0.722	6.38	5	2.67	1.12	0.671	6.47
AA-67	4	2.10	0.89	0.528	3.62	4	2.65	0.95	0.544	3.47
AA-163	4	2.78	1.12	0.697	3.49	4	2.69	1.08	0.613	3.40
Mean	5.2	3.10	1.20	0.63	4.78	4.1	2.45	1.04	0.61	4.72
RBIP-locus		na	ne	I	Gene diversity	na	ne	I	Gene diversity	
MKRBIP-3		3	1.91	0.83	0.478	3	2.10	0.88	0.536	
MKRBIP-4		2	1.36	0.43	0.264	2	1.57	0.55	0.368	
MKRBIP-7		3	1.57	0.60	0.367	2	1.60	0.56	0.383	
Birte-B1		2	1.62	0.57	0.468	2	1.68	0.59	0.412	
Birte-x5		3	2.02	0.72	0.508	3	2.07	0.78	0.529	
Birte-x16		3	1.79	0.75	0.443	3	2.02	0.84	0.516	
1006-x19		3	1.87	0.74	0.468	3	2.18	0.87	0.553	
399-14-9		2	1.98	0.69	0.499	2	2.00	0.69	0.511	
45-x31		3	2.97	1.09	0.667	3	2.96	1.09	0.676	
64-x45		3	2.09	0.79	0.525	2	1.73	0.61	0.430	
281-x40		2	1.17	0.28	0.147	2	1.25	0.35	0.205	
281-x44		3	1.72	0.63	0.419	3	1.87	0.72	0.473	
2055-nr51		2	1.83	0.65	0.457	2	1.99	0.69	0.509	
95-r2		3	1.31	0.42	0.238	2	1.49	0.51	0.337	
2201Cycl-6		3	1.28,	0.43	0.224	3	1.23	0.38	0.194	
1074Cycl-22		3	1.31	0.47	0.236	3	1.28	0.43	0.227	
Mean		2.69	1.74	0.62	0.41	2.50	1.81	0.66	0.42	

na observed number of alleles; *ne* effective number jof alleles (Kimura and Crow 1964); *I* Shannon's Information index (Lewontin 1972)

RBIP isolation strategy. In contrast, all of the ten SSR markers used display polymorphism in the sample set. This is not surprising in view of the much higher polymorphism level for this marker class and the fact that these ten markers were pre-selected by us partly on the basis of polymorphism level. Nevertheless, the PIC values obtained for the polymorphic RBIP markers used in our study indicate high information value which is comparable to that available from SSRs.

We observe no significant correlation between the genetic distance values derived from SSR and RBIP marker data, indicating that these two marker types sample different fractions of genetic diversity in this germplasm. We suggest that combining these two data types is more informative than using just one alone. The RBIP approach is accurate for studies of deeper phylogeny in diverse germplasm, as insertions occur on the Mya scale (Jing et al. 2005), whereas the SSR approach should provide high resolution discrimination between closely related accessions, because of the high mutability of SSRs but should prove less useful for analysing diverse germplasm because of homoplasy (see "Introduction").

An argument for using RBIPs is the good transferability of results between labs. RBIPs yield presence or absence of single bands of known size. SSRs are less transferable, because precise marker size reading and validation are necessary (Bredemeijer et al. 2002). In particular, the use of different analytical systems between labs for SSRs hinders data comparison. For example, agarose gels were used by Tar'an et al. (2005), sequencing gels with silver staining

Table 5 Shannon-Weaver Diversity Indexes per 15 qualitative morphological traits for entire germplasm set and BAPS-selected core set

Trait	Core	Entire
Stipules-character of anthocyan spot	0.64	0.68
Flower-wings colour	1.43	1.42
Flower-vexillum colour	1.28	1.34
Leaflet-margin shape on the second realleaf	1.35	1.23
Seed-funiculus stability	0.00	0.00
Leaflet-margin shape at the first flowering node	0.87	0.78
Seed-colour at full ripeness	1.65	1.80
Seed-cotyledons colour	1.03	1.16
Leaf-type	0.11	0.25
Seed-hilum colour	0.46	0.47
Leaflet-colour	1.41	1.42
Leaflet-shape (at the first flowering node)	1.60	1.52
Leaflet-appex shape	0.97	1.11
Seed-testa colour	0.66	0.70
Seed-surface	0.82	0.74
Mean index	0.95	0.97

were used by Burstin et al. (2001) and Baranger et al. (2004) and automated DNA sequencer reading was used by Ford et al. (2002).

Mixed molecular marker types have been used previously in pea diversity analysis. Tar'an et al. (2005) used a mixed set of RAPD, ISSR and SSR markers to analyse 65 pea varieties and 21 wild accessions. Unfortunately, no direct comparison is possible between out studies and that of Tar'an et al. because only a single sample was shared (cv. Olivin). The most comprehensive study published to date on pea germplasm (Baranger et al. 2004) used a combination of isozymes, seed-storage proteins, RAPD, and 13 EST derived-SSR markers on 148 pea accessions of mainly Western European origin. Again, there is virtually no sample overlap with our study. Interestingly, Baranger et al. (2004) found the most informative marker type to be EST-derived SSRs, with high allele richness and occurrence of rare alleles. Our test of five of these EST-SSRs on a set of 20 Czech origin varieties did not reveal such high polymorphism (data not shown).

Most previously published experiments on pea germplasm have used either a single plant sample per accession or no information has been provided, preventing the analysis of heterozygosity/heterogeneity in accessions. *Pisum* as a predominant selfer but crosses can be made across the genus and heterozygosity is a possibility. Our use of bulks of ten plants per accession provides a more representative description of the germplasm samples, reduces the possibility of mis-scoring, compared to single plant sampling (Van Hintum 1999) and reveals heterogeneity in accessions. In our study both SSR and RBIP markers reveal accession heterogeneity, and no heterozygosity.

Diversity of pea germplasm used in this study

We observe a broad continuous distribution of genetic distances in the germplasm set using molecular data but morphological traits indicate two separated peaks (Fig. 1a, b). Furthermore, we see no significant correlation between the morphological and molecular data for this germplasm set. We suggest that these differences are due to the very different types of data classes used. The molecular markers used derive from multiple dispersed loci in the *Pisum* genome and represent the spectrum of genetic distances between orthologous genomic regions in the germplasm, whereas the morphological traits are controlled by multiple genes, some of which have probably been subjected to strong direct or indirect selection during the breeding process.

A narrow genetic base or even genetic erosion of Western European commercial pea cultivated germplasm has been claimed for pea cultivars (Baranger et al. 2004; Tar'an et al. 2005; Simioniuc et al. 2002). We find quite high diversity in our collection and our analysis of three temporally divided subsets, spanning the past 50 years, has not revealed significant genetic erosion. Similar experiments for maize and bread wheat using molecular data showed genetic narrowing (Le Clerc et al. 2005; Roussel et al. 2005, 2006) but analyses combining morphological, DNA and protein marker and enzymatic characters revealed no such losses of diversity (Donini et al. 2000; Le Clerc et al. 2006). We suggest that it would be important to consider crosses between varieties in different breeding programs to increase the diversity.

Our cluster analysis using molecular data has not fully separated fodder and pea types, in agreement with Tar'an et al. (2005) but in contrast to the study of Baranger et al. (2004). We presume that this is because only the latter included seed storage protein and isozyme data in their study. We suggest that no global genomic differences exist between the two pea types and this is reflected in the SSR and RBIP markers used here but genes controlling seed characters probably show clear distinction.

The ordination analysis reported here has identified nine qualitative and eight quantitative morphological characters, which together account for >90% of both corresponding total variations observed. The qualitative traits identified here differ from those found by Tar'an et al. (2005) but the two sets of results agree with that of Baranger et al. (2004) that the most important PC is associated with flower colour, leaf type and seed weight.

In contrast to previous studies which used PCA and PCO approaches (Baranger et al. 2004; Tar'an et al. 2005), we have used MDS for analysis of molecular data, because of advantages in the rank order of magnitude preservation of distances between the data points and visualisation of a larger proportion of variability. Our data suggest significant gene flow between fodder and dry seed pea types, despite the existence

of several type-specific SSR alleles. Interestingly, MDS separated several accessions which are in the same Ward cluster. It should be noted that molecular markers display much less of the total variance in the first 2–3 axes in than do morphological traits. Therefore, unless highly distinct accessions are tested, such analysis might not be expected to resolve the germplasm into clearly separated outgroups. Such a clear outgrouping was observed when molecular data for ten accessions of winter fodder type were analysed with unique SSR alleles and differential combinations of RBIP loci (data not shown), but the lack of comparable morphological data excluded them from the presented analysis.

We have used BAPS analysis to combine molecular with the morphological data and to select a core collection from the complete collection. The core set retains the majority of diversity for the complete collection, validating this multifactorial approach. Future studies will attempt a similar approach on the full Czech national pea germplasm collection hold at Agritec. There has been low utilisation of exotic germplasm in pea improvement up to now and the selection of a representative core collection will make the germplasm more accessible to the breeder.

Model-based population structure

Our BAPS analysis has shown that consecutive rather than combined morphological and molecular data computation leads to better interpretable results which essentially agree with Ward cluster analysis of morphological data. No direct computational comparison between distance and model-based population structure has been attempted here, since these methods rely upon different principles. Nevertheless, the utility and complementarity of these approaches has been shown here and previously (Corander et al. 2004; Maccaferri et al. 2005). Bayesian approaches can readily deal with combinations of different data types and STRUCTURE software (Pritchard et al. 2000; Falush et al. 2003) has been particularly popular. In contrast to probabilistic assignment of genotypes into user defined cluster numbers, the partitioning based BAPS software uses an analytical integration strategy combined with stochastic search methods. As shown in our study, BAPS provides a good alternative that requires much less computational time, suits more complex data sets, accommodates spatial models of genetic population and investigates admixture inference. Therefore, we propose it as an attractive approach for future germplasm analysis.

Core and reference collections

The core collection concept is well established but assessment of representativeness is usually lacking. No standardised method has yet been accepted for core selection, although numerous strategies have been tested (Van Hintum 1999; Hu et al. 2000; Wang et al. 2007). The most commonly used strategy combines geographical and morphological characteristics (Brown and Spillane 1999) but these parameters are unreliable for reflecting genetic diversity accurately (Tanksley and McCouch 1997). We strongly argue for the establishment of core collections for pea and other crops, using approaches described here, combining suitably reproducible molecular platforms with robust morphological parameters to address population structure and to allow better cross-comparison of results.

Acknowledgments This work was financially supported by Ministry of Education of Czech Republic research project MSMT 2678424601 and Czech Ministry of Agriculture project no. 33083/03-3000. Technical support of Mrs. L.Vítámvásová, E. Fialová, M. Vachatová and E. Kamlerová are greatly acknowledged.

References

Baranger A, Aubrey G, Aran G, Laine AL, Deniot G, Potier J, Burstin J (2004) Genetic diversity within *Pisum sativum* using protein and PCR-based markers. Theor Appl Genet 108:1309–1321

Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. Nature Rev 5:251–261

Beckmann JS, Soller M (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. Biotechnology 10:930–932

Bredemeijer GMM, Cooke RJ, Ganal MW, Peeter R, Isakk P, Noordijk Y, Rendell S, Jackson J, Roder MS, Wendehake K, Dijcks M, Amelaine M, Wickaert V, Bertrand L, Vosman B (2002) Construction and testing of a microsatellite database containing more than 500 tomato varieties. Theor Appl Genet 105:1019–1026

Brown AHD, Spillane C (1999) Implementing core collections—principles, procedures, progress, problems and promise. In: Johnson RC, Hodgkin T (eds) Core collections for today and tommorow. IPGRI, Rome

Brown AHD, Weir BS (1983) Measuring genetic variability in plant populations. In: Tanksley SD, Orton TJ (eds) Isozymes in plant genetics and breeding, part A. Elsevier, Amsterdam

Burstin J, Deniot G, Potier J, Weinachter C, Aubert G, Baranger A (2001) Microsatellite polymorphism in *Pisum sativum*. Plant Breed 120:311–217

Corander J, Martiinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. Mol Ecology 15:2833–2843

Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20:2363–2369

Corander J, Gyllenberg M, Koski T (2007) Random partition models and exchangeability for Bayesian identification of population structure. Bull Math Biol 69:797–815

Donini P, Law JR, Koebner RMD, Reeves JC, Cooke RJ (2000) Temporal trends in the diversity of UK wheat. Theor Appl Genet 100:912–917

Ellis THN, Poyser SJ, Knox MR, Vershinin AV, Ambrose MJ (1998) Polymorphism of insertion sites of Ty1-*copia* class retrotransposons and its use for linkage and diversity analysis in pea. Mol Gen Genet 260:9–19

Erskine W, Muehlbauer FJ (1991) Allozyme and morpfological variability, outcrossing rate and core collection formation in lentil germplasm. Theor Appl Genet 83:119–125

Esquinas-Alacazar J (2005) Protecting crop genetic diversity for food security: political, ethical and technical challenges. Nature Rev Genet 6:946–953

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Flavell AJ, Knox MR, Pearce SR, Ellis THN (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J 16:643–650

Ford R, LeRoux K, Itman C, Brouwer JB, Tayler PWJ (2002) Diversity analysis and genotyping in *Pisum* with sequence tagged microsatellite (STSM) primers. Euphytica 124:397–405

Flavell AJ, Bolshakov VN, Booth A, Jing R, Russell J, Ellis THN, Isaac P (2003) A microarray-based high throughput molecular marker genotyping method: the tagged microarray marker (TAM) approach. Nucleic Acids Res e31:e115

Frankel OH, Brown AHD (1984) Current plant genetic resources—a critical appraisal. In: Genetics: new Frontiers, vol IV. Oxford and IBH Publ. Co., New Delhi

Goudet J (1995) FSTAT (ver. 1.2): a computer program to calculate F-statistics. J Heredity 86:485–486

Hu J, Zhu J, Xu HM (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. Theor Appl Genet 101:264–268

Jing RC, Knox MR, Lee JM, Vershinin AV, Ambrose M, Ellis THN, Flavell AJ (2005) Insertional polymorphism and antiquity of *PDR1* retrotransposon insertions in *Pisum* species. Genetics 171:741–752

Jing R, Bolshakov VI, Flavell AJ (2007) The Tagged Microarray Marker (TAM) method for high throughput detection of single nucleotide and indel polymorphisms. Nature Protocols 2:168–177

Kalendar R, Grob T, Regina M, Suoniemi A, Schulman A (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theor Appl Genet 98:704–711

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27

Le Clerc V, Bazante F, Baril C, Guiard J, Zhang D (2005) Assessing temporal changes in genetic diversity of maize varieties using microsatellite markers. Theor Appl Genet 110:294–302

Le Clerc V, Cadot V, Canadas M, Lallemand J, Guerin D, Boulineau F (2006) Indicators to assess temporal genetic diversity in the French Catalogue: no losses for maize and peas. Theor Appl Genet 113:1197–1209

Lewontin RC (1972) The apportionment of human diversity. Evol Biol 6:381–398

Loridon K, McPhee K, Morin J, Dubreuil P, Pilet-Nayel ML, Aubert G, Rameau C, Baranger A, Coyne C, Lejeune-Henaut I, Burstin J (2005) Microsatellite marker polymorphism and mapping in pea (*Pisum sativum* L.). Theor Appl Genet 111:1022–1031

Maccaferri M, Sanguineti MC, Noli E, Tuberosa R (2005) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Mol Breed 15:271–289

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Nei M (1973) Analysis of gene diversity in subdivaded populations. Proc Natl Acad Sci USA 70:3321–3323

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590

Nersting LG, Andersen SB, von Bothmer R, Gullord M, Jorgensen RB (2006) Morphological and molecular diversity of Nordic oat through one hundred years of breeding. Euphytica 150:327–337

Pavelková A, Moravec J, Hájek D, Bareš I, Sehnalová J (1986) Descriptor list genus *Pisum* L. RICP Prague. Genové zdroje 32:46

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

Reif JC, Melchinger AE, Frish M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci 45:1–7

Reynolds J, Weir BS, Cockerham C (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–776

Rohlf F (2006) NTSYSpc: Numerical Taxonomy System (ver. 2.2). Exeter Publishing, Ltd., Setauket

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Roussel V, Leisova L, Exbrayat F, Stehno Z, Balfourier F (2005) SSR allelic diversity changes in 480 European bread wheat varieties released from 1840 to 2000. Theor Appl Genet 111:162–170

Roussel V, Koenig J, Beckert M, Balfourier F (2006) Molecular diversity in French bread wheat accessions related to temporal trends and breeding programmes. Theor Appl Genet 108:920–930

Shannon CE, Weaver W (1962) The mathematical theory of communication. Univer. of Illinois Press, Urbana

Simioniuc D, Uptmoor R, Friedt W, Ordon F (2002) Genetic diversity and relationships among pea cultivars revealed by RAPDs and AFLPs. Plant Breed 121:429–435

Smýkal P (2006) Development of an efficient retrotransposon-based fingerprinting method for rapid pea variety identification. J Appl Genet 47:221–230

Smýkal P, Valledor L, Rodríguez R, Griga M (2007) Assessment of genetic and epigenetic stability in long-term in vitro shoot culture of pea (*Pisum sativum* L.). Plant Cell Rep 26:1985–1998

SPSS for Windows, Rel. 12.0.1 (2003) SPSS Inc., Chicago

StatSoft Inc (2006) STATISTICA (data analysis software system), version 7.1. http://www.statsoft.com

Tanksley SD, McCouch SR (1997) Seed bank and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066

Tar'an B, Zhang C, Wankertin T, Tullu A, Vandenberg A (2005) Genetic diversity among varietis and wild species accessions of pea (*Pisum sativum* L.) based on molecular markers, and morphological and physiological characters. Genome 48:257–272

Tohme J, Jones P, Beebe S, Iwanaga M (1995) The combined use of agroecological and characterization data to establish the CIAT *Phaseolus vulgaris* core collection. In: Hodgkin T, Brown AHD, van Hintum TJL, Morales EAV (eds) Core collections of plant genetic resources. Wiley, Chichester, pp 95–107

Van Hintum TJL (1999) The general methodology for creating a core collection. In: Johnson RC, Hodgkin T (eds) Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome, pp 10–17

Vavilov NI (1926) Studies on the origin of cultivated plants, Bulletin of Applied Botany, vol 26. Leningrad, USSR

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Wang JC, Hu J, Xu HM, Zhang S (2007) A strategy on constructing core collections by least distance stepwise sampling. Theor Appl Genet 115:1–8

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236

Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BT, Powell W (1997) Genetic distribution of BARE-1 retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol Gen Genet 253:687–694

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphism amplified by arbitrary primers are usefull as genetic markers. Nucleic Acids Res 18:6531–6535

Wright S (1965) The interpretation of population structure by F-statistics with special regards to systems of mating. Evolution 19:395–420

Yeh FC, Boyle TJB (1997) Population genetic analysis of co-dominant and dominant markers and quantitative traits. Belg J Bot 129:157

Zietkiewicz E, Rafalski A, Labuda D (1994) Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction. Genomics 20:176–183