AEGIS project LOA 10/048:

**TOWARDS COMPREHENSIVE PEA GERMPLASM MANAGEMENT
FOR FUTURE USE.**

(16.4.2010 – 16.4.2011)

FINAL REPORT

AEGIS project LOA 10/048:

Towards comprehensive pea germplasm management for future use.

Project coordinator: **Dr. Petr Smýkal**

Agritec Plant Research Ltd., Šumperk, Czech Republic

Project partners:

- Dr. **Miroslav Hybl**, Agritec Plant Research Ltd., Šumperk, Czech Rep.
- Dr. **Andrew Flavell**, University of Dundee at SCRI, Inwergovrie, Dundee, UK
- Mr**. Mike Ambrose**, John Innes Centre, Norwich, UK
- Prof. **Jukka Corander**, Abo University, Finland

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 help cope with environmental and demographic changes (Esquinas-Alcazar 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 7 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 one or more of evolutionary, agroecological and molecular data sets (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) (UPOV 1990, 2002). 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 on traits with low heritability. In contrast, molecular markers accurately represent the underlying genetic variation and now dominate the genetic diversity field. For the analysis of pea diversity, Simple Sequence Repeats (SSRs or microsatellites) have become

popular because of their high polymorphism and information content, co-dominance and reproducibility (Baranger et al. 2004, Loridon et al. 2005, Smýkal et al. 2008a,b). Alternately, marker systems based on retrotransposon insertion polymorphism have been extensively used for phylogeny and genetic relationship studies in pea, providing a highly specific, reproducible and easily scorable method (Jing et al. 2007, 2010, Smýkal et al. 2008a, 2011). Using these markers, several of the world's major pea germplasm collections have been analyzed and core collections formed (Smýkal et al. 2009, 2011, Jing et al. 2010). Although SSR and RBIP marker types are widespread now, their potential is at their limit. With advances in model legume sequencing and genomic knowledge, there is a switch to gene-based markers in pea (Jing et al. 2007). Improvements in marker methods have been accompanied by refinements in computational methods to convert original raw data into useful representation of diversity and genetic structure. Initially and still largely used distance-based methods (Reif et al. 2005) have been challenged by model-based Bayesian approaches. The incorporation of probability, measures of support, ability to accommodate complex model and various data types (Beauomont and Rannala 2004, Corander et al. 2007) make them more attractive and powerful. Large volumes of polymorphic data points have been produced for each collection, which were subsequently subjected both to genetic distance analysis and/or model-based Bayesian diversity analysis. However, after data processing, further use of such data is highly limited, especially in the absence of cross-comparison between collections. Furthermore, most of these accessions have been evaluated for morphological, agronomical and phytopathological traits, thus the data has enormous scientific and breeding potential. Thus, a very important issue is the deposition and availability of both original scores for molecular and agronomical, morphological traits data, as so far data held at national level are not broadly accessible and do not include searchable interfaces. Recently an international consortium (PeaGRIC) has been formed to try to take on a coordinating function from the international Pisum community (Furman et al. 2006, Smýkal et al. 2009). Among the objectives of this consortium are the combining of available data sets into a virtual global collection and the development of a dispersed international reference collection (Smýkal et al. 2009). As proposed, such a set would provide a useful and powerful resource for next generation markers and importantly for phenotypic analysis of agronomic traits, both as toolkits for association mapping, as a strategy to gain insight on genes underlying desired traits (Zhu et al. 2008). The core collection concept is well established but assessment of representativeness is often lacking. No standardised method has yet been accepted for core selection, although numerous strategies have been proposed and tested. The most commonly used strategy combines geographical and morphological characteristics but these parameters are unreliable for reflecting genetic diversity accurately.

The aim of this study was to investigate genetic diversity in a selection of pea assessions, of the Czech and Slovak origin, representing the breeding in respective countries over the last *ca.* 50 years. We have combined morphological qualitative and quantitative characters, with molecular markers and tested a variety of clustering approaches to reveal the diversity of the sample set. We then proceed to suggest the composition of a working core collection to faithfully represent this germplasm (e.g. of Czech and Slovak origin).

Material and methods:

Pisum germplasm collection kept in the Agritec Ltd. Šumperk, Czech Republic currently includes 1,307 accessions of field grain peas (*Pisum sativum L. var. sativum*)(79%) and 21% fodder peas (*Pisum sativum L. var. arvense*). The collection is guided according to the general rule of National Programme for Plant Genetic Resources of Czech Republic and passport data are available on http://genbank.vurv.cz/genetic/resources/ (EVIGEZ). We have used the subset (166 accessions, **Figure 1**) which focuses on Czech/Czechoslovak accessions bred over last 70 years, which has been already integrated into EU TEGERM and GLIP projects, together with selected world-wide reference accessions (Smýkal et al. 2008). Cultivars Gotik, Alan, Adept and Bohatyr were included as controls for quantitative traits.The plants were grown in field trials during the 2004 and 2005 seasons, in a randomised complete block design with two replications (**Supplementary table S1**). Each block represented a single plot of 5 m^2. Field trials were established on a site of orthic luvisol soil type at Šumperk town at an altitude of 315 m, with a long-term average annual temperature of 7.45°C and long-term average annual rainfall of 693 mm. The data are being deposited in IS EVIGEZ and GERMINATE databases.

DNA isolation

Young leaves from 10 randomly chosen (but morphologically characterized) plants per accession were bulked together and DNA isolated according to Smýkal et al. (2008a) using commercial kits.

DNA marker analysis

SSR primer pairs (**Table 1**) were selected from Ford *et al.* (2001) and Loridon *et al.* (2005). RBIP analysis of selected retrotransposone locus specific markers (**Table 2**) was performed according to Smýkal et al. (2008a). PCRs and gel analysis were performed as described in Smýkal et al. (2008a,b). In addition to already generated morphological and molecular data (Smýkal et al. 2008a), these samples were analyzed by additional RBIP and SSR loci to meet compatibility with other germplasm data (Smýkal et al. 2011). Genetic similarity coefficients were calculated using the Jaccard index of similarity using NTSys (Rohlf 2006) and PopGene v1.32. (Yeh and Boyle 1997) software. 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 *j*th allele in clone (*i*). Visualisation of genetic data in factorial space was by multidimensional scaling (MDS) based on a similarity matrix of Jaccard coefficients (Kruskal 1964). Cluster analysis was performed on the genetic similarity matrices by the method of Ward using Statistica programme. The silhouette method (Rousseeuw 1987) was applied for the identification of the optimal number of the most homogeneous clusters (Smýkal et al. 2008a). The resulting clusters were expressed as dendrograms. Goodness of fit was assessed by Mantel test (Mantel 1967) using NTSYS-pc version 2.2. The PopGene and FSTAT v2.9.3.2 (Goudet 1995) programes were used to calculate the following parameters: allele frequencies at each locus for complete and subdivided populations; gene diversity H value, expected and observed homozygosities, population genetic distance expressed as Nei unbiased genetic distance (Nei 1973, 1978), F- and Shannon index (**Table 1** and **2**).

Morphological data

The collection was desribed by the Descriptor List of Genus *Pisum* L. (Pavelková et al. 1986), using 45 morphological descriptors. Qualitative and quantitative variables were scored at the flowering stage by using the Czech descriptor list of the genus *Pisum* L. (Pavelková et al. 1986). The **qualitative leaflet characters** (8) scored were: colour, shape at the first flowering node, apex shape, waxy bloom, margin shape at the first flowering node, margin shape on the second true leaf, character of anthocyan spots, and leaf type. The **qualitative seed characteristics** (9) were: testa pattern, seed colour at maturity; flower characteristics were: base vexillum shape, wings shape, vexillum apex, wing colour, vexillum colour and callyx sepala-termination of upper pair. The **quantitative stem characters** (7) were: shape, length, length to the first productive node, length internode uder the first productive node, number of sterile nodes and type of branching, branching at base. Length of stems and length up to first productive node were analysed in laboratory from 10 plants in 2 replications. **Stipules quantitative characters (2)** scored were: size and spot intensity. **Flower quantitative characters (2)** were: number of flowers in raceme, vexillum size. Pod colour at green ripeness was scored consecutively. In the stage of seed ripeness 10 plants were randomly selected from each plot in both of replications. Plants were packed into the bag and analysed in laboratory. **Qualitative pod characters** scored (2) were: parchment coating and apex shape. **Qualitative seeds characters** scored (4) were: colour, testa colour, hilum colour and cotyledon colour. **Quantitative pod characters** scored (4) were: degree of curving, length, width and number per plant. **Quantitative seeds characters** evaluated (4) were: funiculus stability, weight per plant, number per plant and thousand seeds weight. All variables were converted to nominal classes and were numerically coded (Pavelková et al. 1986), arranged in the matrix (45 characters and 166 accessions). The standardised data matrix of the qualitative and quantitative data was used to generate dissimilarity indices based on Euclidean distances. Cluster analysis was performed on the Euclidean distance matrices by the method of Ward using Statistica programme. The silhouette method (Rousseeuw 1987) was applied for the identification of the optimal number of the most homogeneous clusters (Smýkal et al. 2008a). The resulting clusters were expressed as dendrograms. The principal component analysis (PCA) and cluster analysis were calculated by NTSYS-pc package (Rohlf 2006).

Bayesian structure analysis

To investigate the genetic structure of the collection, the Bayesian method available in the BAPS software (Corander *et al.* 2004, 2006, Smýkal et al. 2008a, 2011) was used.

Core set analysis:

Core Hunter software (Thatchuk et al. 2009) was used to select accesions for core collection assembly. To test the representativeness of 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:

Morphological data analysis

The distribution of morphological characteristics across the pea germplasm set was calculated (**Table 3A, B**). Calculation of Euclidean distances derived from these data resulted in the range of 1.5 to 16.5, with two peaks at 6-7 and 12-13 (**Figure 3A**). It indicates that morphological characters distinguish between dry-seed and fodder pea types, based on several highly correlated characters (related to flower colour and stipule and seed testa pigmentation) which subsequently govern distances calculations. Correlations between morphological characters were examined. A number of traits were found to be strongly correlated, for example stipules with anthocyan spots with flower-*vexillum* colour (r =0.91), flower-wing colour (r =0.94), seed-colour at maturity (r =0.65) and seed-testa colour (r =0.64). Due to different effects of environment, the morhological traits were separately evaluated by: qualitative (with low environmetal impact) and quantitative (high environmental impact) characteristics. The principal component analysis (PCA) as tool for evaluation of traits within the largest ammount on the total diversity was used. The 15 qualitative and 8 quantitative traits were selected. PCA was used for rationalization of further evaluation to select characteristics to capture the most variability using the lowest number of descriptors. Nine out of the fifteen qualitative traits were used to estimate phenotypic diversity (**Figure 8**). More than 90% of the total variation of qualitative traits was explained three principle components (PC) (43.85%, 36.64% and 5.86%), based on the nine qualitative eigenvectors (**Table 4A**). The flower characters of anthocyan spotted stipules and colour of flower wings and vexillum 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 (**Table 4B**). 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 (as listed in **Table 4**) were then used for evaluation of morphologic diversity from the same reason as in case of qualitative traits. PCA of quantitative traits included 93% of total variation in four axes. The most important eigenvectors for PC1 were seed and pod number per plant, together with length of stem and length of 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 (**Table 4A**). The morphological characteristics were loaded into dummy variables and clustered using simple matching coefficients and Ward method. The silhouette method revealed 4 clusters as the most homogeneous solution for morphological parameters, with 3, 5 and 6 clusters also providing meaningful solutions (**Figure 4**). Moreover, both quantitative and qualitative characters showed high heritability for most traits, as each accession had at least 4

independent records (2 replicates per year and 2-3 years replicates). The exceptions were disease resistances to fungal and viral diseases, as might be expected.

Molecular data scoring

The 166 selected accessions were genotyped by additional SSR and RBIP markers giving a total of 20 SSR and 25 RBIP loci, using established protocols (Smýkal et al. 2008a, b). These markers were selected from different linkage groups/chromosomes (**Table 1 and Figure 2**). Of these, 17 SSR and 16 RBIP were shown to be polymorphic and informative for the given subset. It has to be noted, that in order to capture possible heterogeneity of accessions, 10 morphologically assessed plants per accession were used to form a sample for DNA analysis. Based on our recent study (Cieslarova et al. 2011), we estimated that about 10 % of the accessions of the entire collection are heterogenous, and in the selected 166 accessions of this study this figure was approaching 15% (26/166 accessions). Importantly, used markers are fully compatible with the entire dataset of Czech National Pea Collection as well as JIC pea collection (Jing et al. 2010), using RBIP and ATFC pea collection (Zong et al. 2009), using SSR markers (Smýkal et al. 2011). The simplicity and unequivocal scoring of essentially binary mode RBIP markers was clearly demonstrated, as multiloci fragment length scoring of SSR markers proved to suffer from an intrinsic technical error. It should be noted that the analysis of microsatellite fragments was performed on PAGE gels, with silver or EtBr staining and molecular weight marker reading. When checked by sequencing analysis of selected samples, reading accuracy was estimated to be of 6-8 bp. In addition, our recent study of microsatellite homoplasy and mutation rate estimates (Smýkal et al. *submitted*) showed that fragmentation analysis using sequencer (eg. fragment length reading with internal standard) cannot detect homoplasy which can only be done by direct sequencing analysis. As the *Pisum* genus is very diverse, this suggests that for microsatellites the risk of homoplasy in wide surveys of pea germplasm is high. Jing et al. (2007) estimated the age of alleles segregating in *Pisum*, and found this to be 1.9 ± 0.7 million years. If we take 10^{-4} as the microsatellite length mutation rate per year, then we expect on average 2% of (correct) microsatellite allele calls to misattribute ancestry. Obviously this will be a less severe problem in narrow (for example cultivated) germplasm and more severe in wide germplasm.

Molecular data analysis

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 and recorded in Excel spreadsheets. In the case of RBIP analysis, a third state, namely complete absence of any PCR product corresponding to primer site mutation (Jing et al. 2007, 2010) was added. We surveyed 166 pea accessions using 20 SSR loci. Of these 17 were polymorphic and yielded a total of 53 alleles with a minimum 3 and maximum 8 alleles per locus (**Table 1**). Twelve rare alleles (22%) with frequencies below 0.05 were found at 6 SSR loci. Calculated Polymorphic Information Content (PIC) values were high, ranging from 0.697 to 0.964, with an average of 0.89. Heterogeneity, associated with the use

of 10 bulked plants per accession, was detected in 60 out of 166 accessions (35%) at 8 loci, averaging 0.069 (7%). Analysis of 20 individual plants per accesions, in the case of 15 selected accessions, indicated heterogeneity within accession rather than individual plant heterozygosity. This agrees well with our recent genetic erosion study (Cieslarová et al. 2011). The same sample set was analysed with 25 retrotransposon RBIP markers. Sixteen of these detected polymorphism in the investigated germplasm set (**Table 2**), identifying a total of 42 alleles. Ten RBIP loci repeatedly produced occasional zero scores (e.g. no PCR product detected, which could be technical failure and/or expression of diversity due to priming site mutation, for detailed explaination see Jing et al. 2005, 2010) (frequencies 0.011-0.35). Fourteen of the informative RBIPs detected residual heterogeneity, varying from 0.006 to 0.335 in 26 accessions (average 16%). 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 166 accessions, apart from 2201Cyc6, 1074Cyc12, 95x2 and MKRBIP4, where the occupied site allele dominated over the empty site (0.84 to 0.91).

Genetic relationships revealed by SSR and RBIP molecular markers

We compared genetic diversity analyzed separately by SSR versus RBIP markers. Although both marker types are derived from repetitive sequences, they clearly sample different proportions of the large pea genome, as can be seen from genetic distances matrix comparison (**Figure 5**). Consequently, all molecular markers results were used in the final analysis. Pairwise genetic distances were calculated from Jaccard similarity coefficients for combined SSR and retrotransposon data. Ward hierarchical ascendant classification was then performed on the distance matrix and finally a dendrogram was built (**Figure 6**). The silhouette method, adopted after the Ward clustering, identified 9 clusters as the most probable estimate (**Figure 6**, **7**). Ward cluster I contains mainly fodder type accessions, cluster II contains 5 fodder and 23 dry-seed types, 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 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 1980´s, 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 32% of fodder pea accessions in the same group as field peas, while combining SSR and RBIP data clustered 67% of fodder pea accessions into 2 of the 9 clusters. In the case of RBIP markers, no specific allele was linked to seed type. Frequency calculations for all SSR and RBIP marker-based distances of the entire data set resulted in a column graph with a normal-like distribution in the range of 0.2 to 1.0, with mean of 0.65 (**Figure 3B**). To reveal another level of structure for the collected sample set, multidimensional scaling (MDS) was performed on the SSR and RBIP data. This identified a broad, continuous variation for the pea sample set with no clear outgroup (**Figure 9**).

Molecular and morphological data integration

Bayesian genetic structure analysis
To investigate the genetic structure of the pea collection, the Bayesian method available in the BAPS software was used. We used this software, as it was shown previously to be superior to commonly used STRUCTURE (Pritchard et al. 2000), including in pea germplasm diversity analysis (Smýkal et al. 2008a, 2011).

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 (Smýkal et al. 2008a). 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. In the case of morphological data, three or six clusters were found by optimal partitioning, with log marginal likelihood values of -14971.4 for 6 clusters and -17237.7 for 3 clusters. This clear and strong partitioning is because of high correlation of numerous traits as revealed also by PCA. One cluster comprised 107 dry-seed plus 2 fodder varieties, a second cluster comprised 47 fodder accessions plus 2 dry-seed varieties and the third cluster comprised 8 dry seed varieties of *afila* type. Therefore, partitioning into 3 clusters was accepted, with a probability of 1.0 (**Figure 10**). Bayesian model-based analysis of combined (SSR plus RBIP) **molecular data** partitioned the sample set into 29 clusters, with a log marginal likelihood value of optimal partition at -7184.9 and a probability of 0.948, showing high structuring of the set. Eleven of these clusters contained nearly exclusively fodder type accessions, 4 others grouped 23 out of 47 dry-seed accessions. The remaining clusters provided no clear assignment of the accessions to either type or breeding period. To reveal the genetic distance between BAPS identified clusters, a Neighbour joining tree was computed as one of the direct outputs of BAPS analysis and shows the relationship of the clusters (**Figure 12**). Conversely, computation of genetic distances within individual clusters show high homogenity (not shown), as can be expected based on Bayesian method setup.

Bayesian model-based analysis of composed **molecular and morphological data** resulted in partitioning into 3 clusters, identical to morphology based data. This is again due to the high correlation of several morphological traits such as 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) (**Table 4A, B**). Thus we have undertaken sequential BAPS clustering using first morphological data resulting in 3 clusters, followed by analysis of molecular data. This approach has structured the analyzed dataset into 17 subclusters of morphology-derived cluster I (field pea), 12 subclusters of cluster II (fodder pea) and 5 subclusters of cluster III (afila type semi-leafless accessions) (**Figure 11**).

Core collection establishment based on BAPS analysis
The final aim of this study was to formulate a core collection of 166 analyzed Czech and Slovak origin accesions, using the combined diversity data. Exploring the Bayesian BAPS analysis of integrated data, a single accession per cluster was selected out of 34 subclusters (morphology – molecular sequential analysis), to form a core collection (Smýkal et al. 2008a) (**Figure 13**). To determine whether this 34 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 nature of the RBIP and SSR data classes (3 possible alleles for the former vs. multiple alleles for the latter), the two marker classes were analysed separately. **Table** 7 shows that both, the average gene diversity value and allelic richness per locus are similar for both molecular marker types between the core collection of 34 accesions and the original 166 accessions. These data indicate that the core collection represents the diversity of the complete collection very well. A similar comparison between the core and total germplasm sets was performed using all 15 qualitative morphological traits. Sixty-three out of 78 trait categories (descriptor states) shown by the entire set are present in the core selection. This decreasement in trait cathegories can be explained by several factors. Especialy quantitative traits are more variable due to environmental conditions during vegetation period. This can be one of the possible reasons of decrease of number of the categories ("trait category" means expression of some descriptor). Another possible explanation of lower number of the "trait category" is the weight of the evaluating aproaches which was done during the molecular and morphological data analysis. In relation to this, indeed more value was given to molecular data, as these offered more detailed clustering. Consequently, the expression of some morphological traits was lost. However, it has to stressed that "core collection" is not definitive "status quo", but it is dynamic unit which can be (and will be) modified according to novel data and accesions. 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 8**). However, since cluster size (number of accessions/cluster) varies from 2 to 14, further accessions were chosen in accordance with cluster size. Thus, 7, 6, 5, 4, 4 and 4 accesions were arbitrarily selected from subclusters 1, 2, 3, 4, 5 and 6 (of cluster I) respectively, resulting in 52 accessions of BAPS-derived core. Although the size of the core collection based on this criteria is similar to 20% of the core set (49 accessions) identified by Core Hunter (see below), the selected accessions differ. However, this approach of selecting further accessions based on the BAPS analysis cluster size contains substantial part of subjectivity. Thus, for further work we have used purpose-build Core Hunter software (Thatchuk et al. 2009) which select accesions without any additional selection criteria from researcher.

Core collection selection using Core Hunter and assessment of representativeness

In the course of the Czech national pea germplasm project (MSM26424601, 2004-2010) we established a collaboration with C. Thatchuk from Dept. Computational Science, University of British Columbia, Canada and J. Crossa from CIMMYT, Mexico. These collaborators recently developed the Core Hunter software (Thatchuk et al. 2009) and demonstrated a better representative core set selection over other software. Computations used in this software use several diversity measurements and correspondingly, respective core collections based on different diversity measures can be formed. As previously tested, the so called „multiobjective" selection, based on combination of Modified Roger´s distance (MR), Cavalli-Sforza (CV) and Shannon-Weaver index (SN), was used preferentially. Moreover, Core Hunter can be set to select core sets of various sizes, eg. with selection intensity of 10-

20-30% (**Figure 14, Table 5**). The minimum size core sets derived from respective molecular markers were also computed (**Figure 15**). Thus 11 accessions from SSR-based, 10 RBIP and 19 RBIP and SSR-based, sufficiently (by 95% of marker diversity) represent the diversity of the original 166 accession dataset . The representativeness of core sets was assessed both for molecular and phenotypic data (**Table 7,8**). In comparison to the BAPS-derived (34 accesions) core set, similarly sized Core Hunter set selected on 20% sampling intensity (e.g. 33 accessions) showed better representation of original diversity measures, especially for molecular data (**Table 7**). This is not surprising since these data were actually used for Core Hunter development and computations. Moreover, this software is designed to maximize genetic diversity, while minimizing accession numbers. On the other hand, since the BAPS-derived core was based on both, morphologic and molecular data (although not integrated in one analysis as discussed earlier), Shannon indexes for traits were slightly better for this core set (**Table 8**). Unfortunately, at this stage, there is no specific algorithm (for example Euclidean distance-based) available for morphological dataset computation which would also allow an integrative approach (eg. composition of molecular and morphological data) (Franco, Crossa and Desphande 2010). However, further improvement of the software, including an user's friendly interface, is underway.

Data deposition into databases

All above obtained data will be made available within national EVIGEZ (http://genbank.vurv.cz/genetic/resources/) database, which in currently in process of being changed to the GRIN-Global system. Steps for entry of data into existing and functional GERMINATE database (Lee et al. 2005)(http://bioinf.scri.ac.uk/germinate_pea/app/) have been assessed and are in progress. Furthemore, relevant data are already available for the entire collection as well as compatible data from JIC pea collection analyzed within the frame of previous TEGERM and GLIP EU projects (Smýkal et al. 2011).

Supportive comments – further ongoing steps:

In addition to the 166 accessions used in this study, we have performed identical analysis over the 1,283 accessions of the entire Czech National Pea Collection (CzNPC), analyzed by RBIP markers. The BAPS analysis of molecular dataset partitioned the CzNPC collection into 9 statistically supported clusters (Smýkal et al. 2011) (**Figure 13**). Subsequent establishment of a representative core collection, using both BAPS and Core Hunter approaches, is currently in progress. This larger set of accessions, i.e. the CzNPC) better demonstrates the potential of the methodology. The results have shown that while upon BAPS clustering still some subjective choice of accessions from individual BAPS-derived clusters is needed, Core Hunter identifies selected accession directly. On the other hand Bayesian clustering helps in visualization and provides supportive measures of diversity distribution as can been seen in Figures 13, 14 and 15. Furthermore, 19 shared RBIP markers enabled us to integrate data of JIC pea collection [3,029 accessions, composed of cultivars (33%), landraces (19%) wild accessions (13%) and genetic stocks (26%)], with 1,283 of CzNPC, composed of mainly

commercial varieties and breeding lines (75%), landraces (24%) and mutants or wild material (1%) and 117 Chinese origin Australian Temperate Field Crop Collection (ATFC) core set accessions. This analysis provided a visual analysis of *Pisum* genus diversity from wild to cultivated material (Smýkal et al. 2011). Currently we are in process of establishment of virtual world-wide pea germplasm collection to be used for further studies such as association mapping.

Recommendations:

The core collection concept is well established but both the methods of accession selection and assessment of representativeness have not been standardised. The most commonly used strategy combines geographical and morphological characteristics but these parameters are unreliable for reflecting genetic diversity accurately. We have used precise molecular (i.e. locus specific) markers as well as standardized morphological descriptors, to assess genetic diversity of selected pea germplasm accessions/collections, together with various approaches of core collection establishment. Based on results of this and previous studies (Smýkal et al. 2008a,b, 2011, Cieslarová et al. 2011) we suggest the use of several plants per accession (at least 10 per one DNA bulk samples) as this, in comparison with single plant sampling, assure adequate representation of the total diversity in an accession, reduces the possibility of mis-scoring and reveals heterogeneity within accessions (Cieslarová et al. 2011). We observed 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 therefore suggest that combining various data types provides better representation of diversity than using just one alone (parallel to more analyzed plants per accession).

Molecular markers display much less of the total variance in the first 2-3 axes of ordination analysis (such as MDS or Paco) than do morphological traits, unless highly distinct accessions are analysed. Consequently, such analysis might not be expected to spread the germplasm into clearly separated outgroups. Furthermore, we see no 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 here derive from multiple dispersed loci in large *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.

While additional computing, such as bootstrapping or the silhouette methods, are needed to provide support for genetic distance-based clustering, all these parameters are directly provided by model-based approaches. There are several types of Bayesian modelling software currently available. Although they perform similarly in relatively small datasets, there are differences, especially when the level of subpopulation differentiation is low, as it is common in germplasm collections. We compared commonly used STRUCTURE versus BAPS, and

found later to clearly and decisively identify clusters, while former needed subclustering approach and supportive measures for cluster number estimation (Smýkal et al. 2011).

Our BAPS analysis has shown that consecutive rather than combined morphological and molecular data computation leads to better interpretable results, owing to the high correlation of numerous morphological traits. No direct computational comparison between distance and model-based population structure has been attempted, as these methods rely upon different principles. Nevertheless, the utility and complementarity of these approaches has been shown here and previously (Smýkal et al. 2008a). In contrast to probabilistic assignment of genotypes into user defined cluster numbers as performed by STRUCTURE, the partitioning based BAPS software use an analytical integration strategy combined with stochastic search methods. As shown in our study (Smýkal et al. 2008a, 2011), BAPS requires much less computational time, is suited to more complex data sets and accommodates spatial models of genetic population and investigate admixture inference. Therefore we recommend it as a useful approach for germplasm management. Furthermore, we have used BAPS analysis to select a core collection of 34 accessions. The core set includes the majority of diversity of the original collection, validating this multi-factorial approach. Purpose build Core Hunter software, performed very well on this test set (as well as much larger entire collections) and can be recommended for efficient core collection establishment. These two methods are not mutually exclusive and in fact BAPS visualization of genetic diversity provides a rational for Core Hunter.

We strongly argue for the establishment of core collections for pea and other crops, using the approaches described here, combining suitably reproducible molecular platforms with morphological parameters to address population structure and to allow better cross-comparison of results. Such collections will be valuable for producing an integrated framework of genetic and phenotypic data generated by different studies.

Based on the present analysis we recommend a set of core accessions be offered up and assigned as AEGIS accessions. These accessions are already listed in the Czech National Inventory as lines available as part of the Multilateral System from the Czech Republic. We recommend these core accession set be flagged as AEGIS accessions in the AEGISSTAT field of EURISCO.

We have undertaken initial steps in composing a world-wide as well as European virtual pea germplasm assembly, using approaches described in this report (Smýkal et al. 2011 and manuscript in preparation). Clearly the analytical approaches presented here should help in the identification of accessions which should be maitained in other European genebanks as part of a decentralised European Core Collection.

Bibliography:

Baranger A. et al. (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.

Cieslarová J, Smýkal P, Dočkalová Z, Hanáček P, Procházka S, Hýbl M, Griga M (2011) Molecular evidence of genetic diversity changes in pea (Pisum sativum L.) germplasm after long-term maintenance. Genet Resour Crop Evol 58: 439-451

Corander J. et al. (2007) Random partition models and exchangeability for Bayesian identification of population structure. Bull Mathem Biol 69: 797-815.

Erskine W and 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

Franco J., Crossa J. and Desphande S (2010) Hierarchical multiple-factor analysis for classifying genotypes based on phenotypic and genetic data. Crop Science 50:105-117.

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

Furman B. et al. (2006) Formation of PeaGRIC: An international consortium to co-ordinate and utilize the genetic diversity and agro ecological distribution of major collections of Pisum. Pisum Genetics 38:32-34.

Goudet J (1995) Fstat version 1.2: a computer program to calculate Fstatistics. Journal of Heredity 86, 485-486

Jing R, Johnson R, Seres A, Kiss G, Ambrose MJ, Knox MR, Ellis TH and Flavell AJ (2007) Gene-based sequence diversity analysis of field pea (Pisum). Genetics 177: 2263-75.

Jing R, Vershinin A, Grzebyta J, Shaw P, Smýkal P, Marshall D, Ambrose MJ, Ellis THN and Flavell AJ(2010)The genetic diversity and evolution of field pea (Pisum) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis. BMC Evolutionary Biology 10: 44.

Jing R.C. et al. (2005) Insertional polymorphism and antiquity of PDR1 retrotransposon insertions in Pisum species. Genetics 171: 741-752.

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

Lee M.J. et al. (2005) GERMINATE. A Generic Database for Integrating Genotypic and Phenotypic Information for Plant Genetic Resource Collections. Plant Physiol. 139: 619–631.

Loridon K. et al. (2005) Microsatellite marker polymorphism and mapping in pea (Pisum sativum L.) Theor Appl Genet 111:1022-1031.

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 Nat 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

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

Reif J.C. et al.. (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci 45: 1-7.

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

Smýkal P. et al. (2009) Effort towards a world pea (Pisum sativum L.) germplasm core collection: The case for common markers and data compatibility. Pisum Genetics 40: 11-15.

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

Smýkal P., Horáček J., Dostálová R., Hýbl M. (2008b): Variety discrimination in pea (Pisum sativum L.) by molecular, biochemical and morphological markers. J Appl Genetic 49: 155-166.

Smýkal P., Hýbl M., Corander J., Jarkovský J., Flavell A.J., Griga M. (2008a) Genetic diversity and population structure of pea (Pisum sativum L.) varieties derived from combined retrotransposon, microsatellite and morphological marker analysis. Theor Appl Genet 117:413–424.

Sneath PHA and Sokal RR (1973) Numerical Taxonomy. Freeman. San Francisco. SPSS for Windows, Rel. 12.0.1. 2003. Chicago: SPSS Inc.

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

Thatchuk C, Crossa J, Franco J, Dreisigacker S, Warburton M and Davenport GF (2009) Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinformatics 10: 243.

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, E.A.V. Morales EAV (eds.) Core collections of plant genetic resources. John Wiley & Sons, Chichester, UK. pp. 95-107

UPOV 1990. Guidelines for the Conduct of Tests for Distinctness, Homogeneity, and Stability. Document UPOV TG/4/7. Geneva, Switzerland.

UPOV-BMT: BMT/36/10. 2002. Progress report of the 36th session of the technical committee, the technical working parties and working group on biochemical and molecular techniques and DNA-profiling in particular. Geneva, Switzerland.

Vavilov NI (1926) Studies on the origin of cultivated plants. Bull Appl Bot 26. Leningrad, USSR

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

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

Zhu C., Gore M., Buckler E.S., Yu J. (2008) Status and prospects of association mapping in plants. The Pl. Genome 1: 5-20.

Zong X, Redden RJ, Liu Q, Wang S, Guan J, Liu J, Xu Y, Liu X , Gu J, Yan L, Ades P and Ford R (2009) Analysis of a diverse global Pisum sp. collection and comparison to a Chinese local collection with microsatellite markers. Theoretical and Applied Genetics 118: 193–204.

Publication dedicated to the project:

Attachments:

Figure 1

Composition of the entire Czech national pea germplasm (1,283 accessions) according to geographical origin (in percentage). Czech and Slovak origin accessions (166) used in this study are indicated.
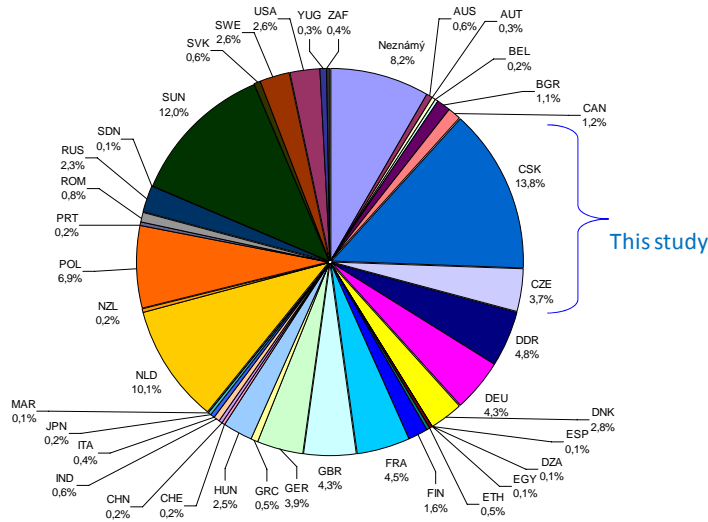

Figure 2

Distribution of SSR loci shown on pea linkage groups (according to Loridon et al. 2005).
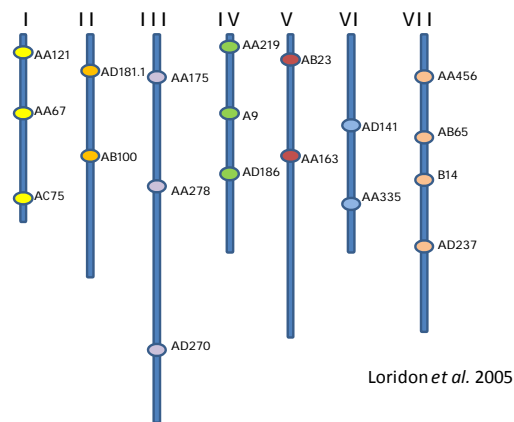
Table 1

Microsatellite markers derived diversity data for 166 analyzed accessions.

SSR locus	Linkage group	Position (cM)	Number of alleles	Observed heterozygozity	Size range (bp)	Polymorphic Information Content (PIC)
AA-121	I	10.2	5	0.033	450-490	0.786
AA-67	I	80.3	6	0.012	330-390	0.882
AC-75	I	154.5	4	0.021	520-550	0.865
AD-186	II	36.2	8	0.147	220-320	0.961
AB-100	II	102.0	4	0.035	520-550	0.722
AD-270	III	254.3	7	0.0	230-290	0.964
A-278	III	154.9	3	0.062	130-170	0.827
AA-175	III	43.9	5	0.031	420-500	0.846
A-9	IV	62.1	3	0.073	330-380	0.886
AA-219	IV	0.8	1	0.0	520	0.0
AA-163	V	100.3	5	0.120	250-320	0.869
AB-23	V	36.8	6	0.057	560-590	0.846
AD-141	VI	70.1	7	0.185	210-330	0.973
AA-335	VI	124.4	3	0.057	540-600	0.756
AA-456	VII	25.VIII	5	0.025	460-510	0.847
B-14	VII	113.9	4	0.017	430-470	0.929
AD-237	VII	152.1	7	0.073	220-360	0.934
AB-65	VII	94.1	3	0.0	140-180	0.697
Mean			**4.5**	**0.048**		**0.763**

Table 2

Retrotransposone-based (RBIP) markers derived diversity data for 166 analyzed accessions.

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**

Figure 3A

Frequency calculation of Euclidean distances of morphological (quantitative and qualitative) characters.
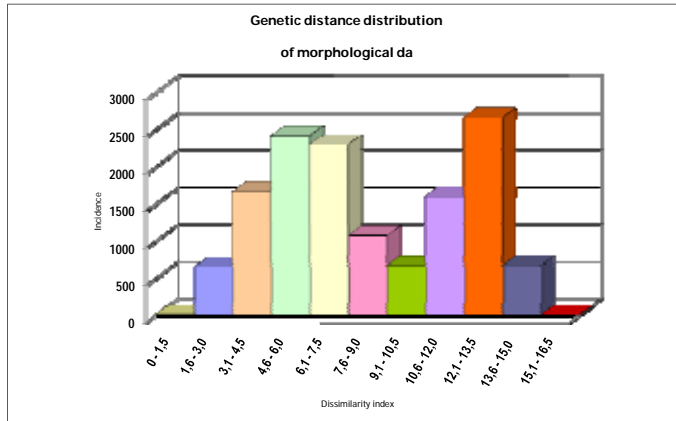

Genetic distance distribution of morphological da

Figure 3B

Frequency calculation of distances of molecular (SSR and RBIP) markers. Values are expressed as ($\sqrt{1}$-Jaccard similarity coefficient) on the x axis, while y axis indicates the number of accessions of each class.
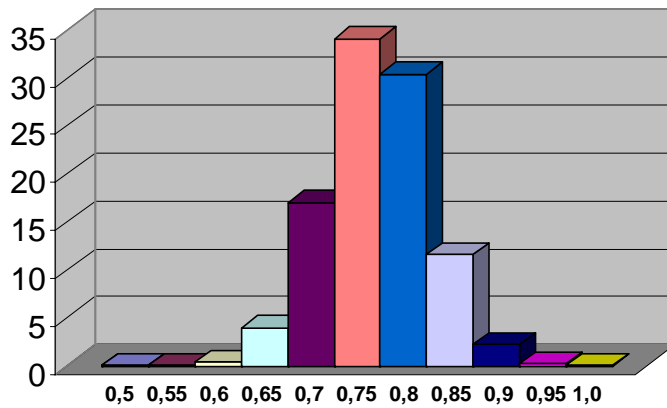
Table 3A

Class frequency distributions of 15 qualitative morphological characters

Descriptor and Classes	Frequency of Class (%)								
	1	2	3	4	5	6	7	8	9
Leaf-type:									
1=leafless; 3=leafless partly; 5=paripinnate leaf;									
7=imparipinnate leaf; 8=non=coupled vinnate	5	-	-	-	94	-	1	-	-
Leaflet-shape:									
1=oblong; 2=oblong ovate; 3=ovate; 4=obovate;									
5=obovate to broadly ovate; 6=broadly ovate;									
7=broadly obovate; 8=rounded	1	26	36	9	11	11	1	-	-
Leaflet-margin shape on the second realleaf:									
1=entire; 2=finely serrate; 3=finely dentate;									
4=serrate; 5=dentate; 6=coarsely serrate;									
7=coarsely dentate	51	27	5	6	5	1	-	-	-
Leaflet-margin shape at the first flowering node:									
1=entire; 2=finely serrate; 3=finely dentate;									
4=serrate; 5=dentate; 6=coarsely serrate;									
7=coarsely dentate	73	13	7	1	1	-	-	-	-
Leaflet-appex shape:									
1=acuminate; 2=acute; 3=obtuse; 4=truncate;									
5=sinucate; 6=obcordate	6	51	30	7	1	-	-	-	-
Leaflet-colour:									
1=basic green absent; 2=yellow green;									
3=light green; 4=green; 5=gray-green;									
6=dark-green; 7=blue-green;									
8=with anthocyan spot; 9=with anthocyan									
on the contour line	-	4	19	40	17	15	-	-	-
Stipules-character of anthocyan spot:									
1=absent; 2=light on the base;									
3=light all over the plant; 4=dark on the base;									
5=dark all over the plant; 6=doubled ont the base;									
7=doubled all over the plant	70	-	1	1	27	1	-	-	-
Flower-vexillum colour:									
1=white; 2=cream; 3=yellow; 4=light pink;									
5=pink; 6=red; 7=light-violet; 8=violet; 9=red-violet	12	58	1	2	2	2	15	8	-
Flower-wings colour:									
1=white; 2=cream; 3=yellow; 4=light pink;									
5=pink; 6=red; 7=light-violet; 8=violet; 9=red-violet	40	31	-	-	2	2	2	5	18
Seed-colour at full ripeness:									
1=light-yellow; 2=yellow-pink;									
3=waxy (two-coloured); 4=yellow-green;									
5=gray-green; 6=dark-green; 7=light brown;									
8=brown; 9=black	24	18	6	1	7	25	15	4	-
Seed-testa colour:									
1=absent; 2=with violet dots; 3=violetly striped;									
4=with brown dots; 5=brown marble like;									
6=brown marbel like with violet dots; 7=brown									
marbel like with violet spots; 8=brown marble									
like with violet stripes; 9=other	76	19	2	1	1	1	-	-	-
Seed-funiculus stability:									
3= no seed-coat coalescence; 7=seed-coat coalescence	-	-	100	-	-	-	-	-	-
Seed-cotyledons colour:									
1=yellow; 2=orange; 3=yellow-green (two coloured);									
4=green; 5=dark green (emerald)	62	10	5	15	8	-	-	-	-
Seed-hilum colour:									
1=light; 2=brown; 3=dark-brown; 4=black	87	6	-	7	-	-	-	-	-
Seed-surface:									
1=smooth; 2=superficially wrinkled; 3=faveolate									
4=irregulary wrinkled; 5=wrinkled	73	22	4	1	1	-	-	-	-

Table 3B

Class frequency distributions of 18 quantitative morphological characters

Descriptor and Classes	Frequency of Class (%)								
	1	2	3	4	5	6	7	8	9
Stem-number of sterile nodes:									
1=less than 6; 2=6-7; 3=8-9; 4=10-11; 5=12-14;									
6=15-16; 7=17-18; 8=19-20; 9=more than 20	-	5	6	15	46	22	4	2	-
Stem-lenght of internode under the first									
productive node:									
1=less than 2.0cm; 2=2.0-3.0cm; 3=3.1-4.0cm;									
4=4.1-5.0cm; 5=5.1-6.0cm; 6=6.1-7.0cm;									
7=7.1-8.0cm; 8=8.1-9.0cm; 9=more than 9.0cm	-	-	5	8	20	15	18	15	19
Stem-lenght to first productive node:									
1=less than 10cm; 2=10-20cm; 3=21-30cm;									
4=31-40cm; 5=41-50cm; 6=51-60cm;									
7=61-70cm; 8=71-90cm; 9=more than 90cm	-	2	4	16	26	15	16	17	4
Plant-seeds number (to standard cv.):									
1= less than 65%; 2=65-75%; 3=76-85%;									
4=86-95%; 5=96-105%; 6=106-115%;									
7=116-125%; 8=126-135; 9=more than 135%	15	21	8	9	7	9	6	11	14
Plant-pods number (to standard cv.):									
1= less than 65%; 2=65-75%; 3=76-85%;									
4=86-95%; 5=96-105%; 6=106-115%;									
7=116-125%; 8=126-135; 9=more than 135%	7	11	14	17	10	7	7	9	18
Plant-seeds weight:									
1= less than 65%; 2=65-75%; 3=76-85%;									
4=86-95%; 5=96-105%; 6=106-115%;									
7=116-125%; 8=126-135; 9=more than 135%	17	24	26	10	10	9	2	2	-
Stem-length:									
1=less than 30cm; 2=30-45cm; 3=46-60cm;									
4=61-80cm; 5=81-100cm; 6=101-120cm;									
7=121-140cm; 8=141-160cm; 9=more than 160cm	-	2	13	35	18	21	10	1	-
Thousand seeds weight:									
1=less than 50g; 2=50-100g; 3=101-150g									
4=151-200g; 5=201-250g; 6=251-300g;									
7=301-350g; 8=351-400g; 9=more than 400g	-	3	10	25	32	22	7	-	1

Table 4A

Matrix of eigenvalues and vectors of principal components for 9 qualitative characters

	Principal components (PC)		
	PC1	PC2	PC3
Eigenvalues			
Variance	4.397	3.299	0,527
% Total contribution	48.85	36.64	5.86
% Accumulated	43.85	85.50	91.36
Eigenvectors			
Stipules-character of anthocyan spot	**0.812***	-0.516	-0.021
Flower-wings colour	**0.819**	-0.522	-0.030
Flower-*vexillum* colour	**0.811**	-0.519	-0.021
Leaflet-colour	0.611	**0.731**	-0.013
Leaflet-shape	0.604	**0.711**	0.055
Leaflet-*apex* shape	0.634	**0.737**	-0.036
Seed-colour at full ripeness	0.679	-0.403	**-0.458**
Leaf-type	0.639	**0.743**	0.025
Seed-testa colour	0.631	-0.439	**0.557**

* Values in the bold are larger than the treshold (average from highest and lowest absolute values of eigenvectors for a column).

Table 4B

Matrix of eigenvalues and vectors of principal components for 8 quantitative characters

	Principal component (PC)			
	PC1	PC2	PC3	PC4
Eigenvalues				
Variance	4.347	1.244	0,979	0,885
% Total contribution	54.35	15.56	12.25	11,06
% Accumulated	54.35	69.91	82.15	93,21
Eigenvectors				
Plant-seeds number	**0.895***	**-0.349**	0.168	-0.010
Plant-pods number	**0.847**	**-0.362**	0.120	-0.028
Stem-lenght to first productive node	**0.828**	**0.442**	-0.205	0.117
Stem-length	**0.851**	**0.399**	-0.090	-0.049
Thousand seeds weight	-0.649	**0.394**	-0.034	**0.612**
Plant-seeds weight	0.598	-0.252	0.315	**0.671**
Stem-lenght of internode under the first productive node	0.486	**0.647**	**0.507**	-0.184
Stem-number of sterile nodes	0.634	0.008	**-0.728**	0.100

* Values in the bold are larger than the treshold (average from highest and lowest absolute values of eigenvectors for a column).

Figure 4

Ward hierarchical ascendant classification of morphological characters calculated by simple matching coefficient. Inset shows Silhouette method determination of most probable 4 cluster numbers.
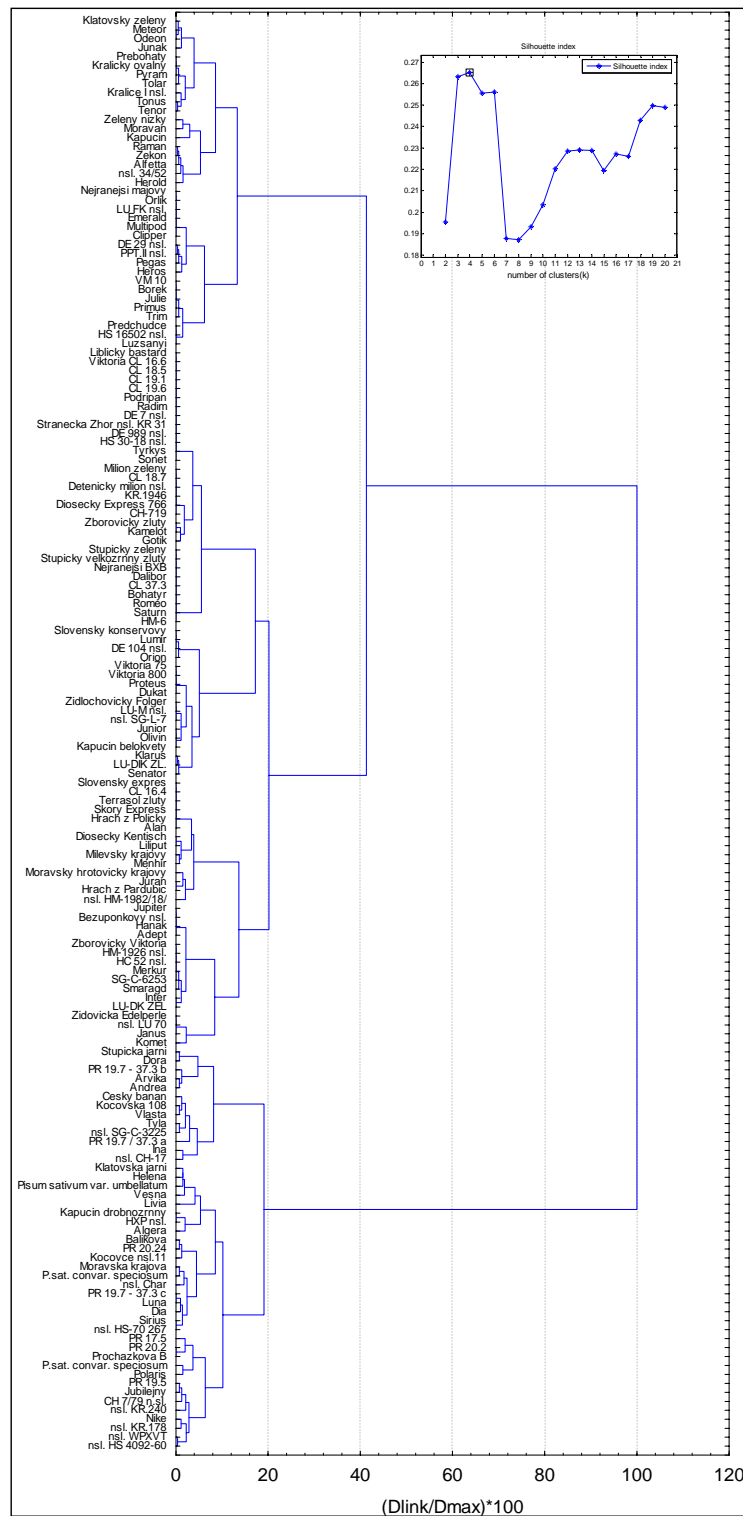
Figure 5

The SSR (x axis) versus RBIP (y axis) derived matrix comparison, derived from Dice similarity coefficient. Matrix correlation: r = 0.02071(= normalized Mantel statistic Z), approximate Mantel t-test: t = 1.286, prob. random Z < obs. Z: p = 0.9009

Figure 6

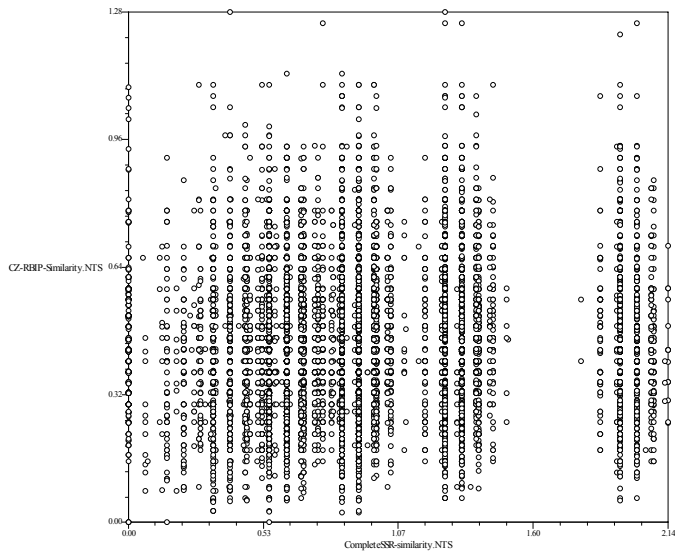Ward hierarchical classification performed on the combined SSR and RBIP molecular distance matrix. The most probable solution based on Silhouette method (inset) for cluster number estimation (9) is indicated by red arrow. Red vertical bar indicates this in dendrogram.
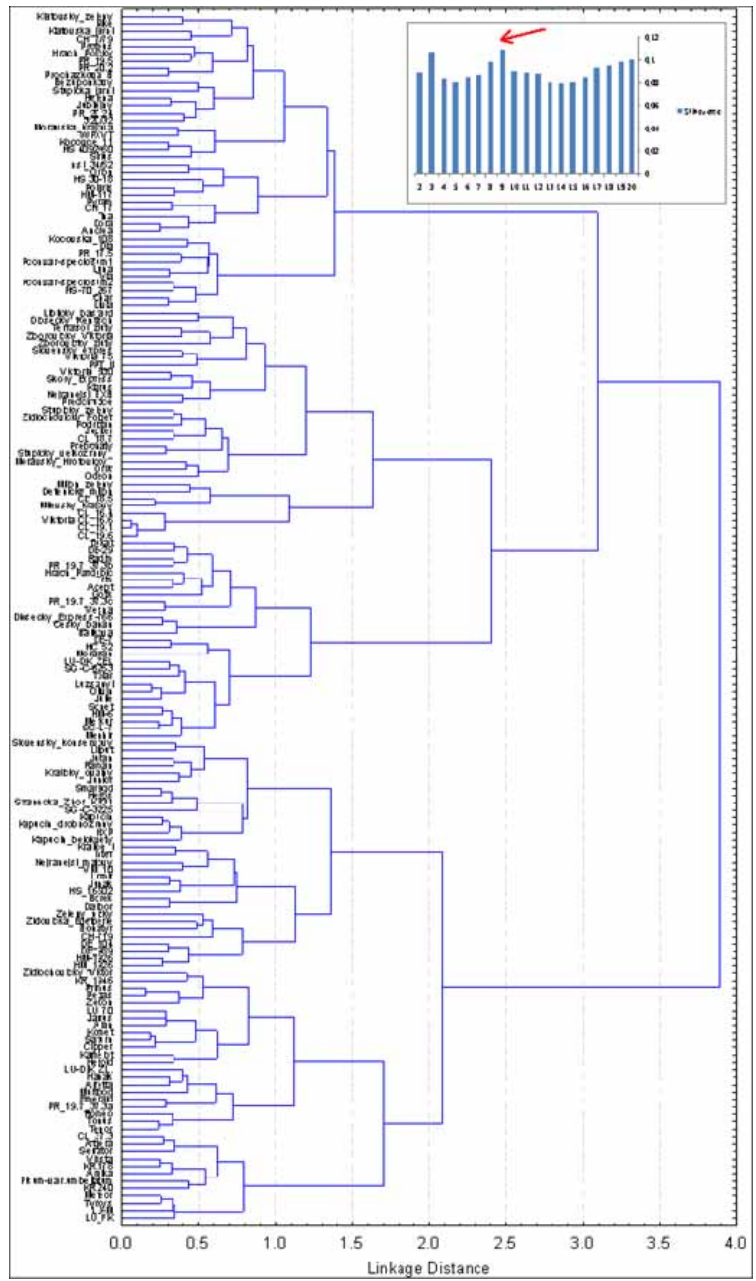
Figure 7

Molecular markers derived Ward clusters (2 to 15) colour visualized and ordered by clustering at level 9 (in red circle).

Figure 8

Principal Component Analysis (PCA) of morphological descriptors. PCA1: 33.2%, PCA2: 18.5%. Three breeding periods (I; up to 1960, II; 1970-80, III; 1980 to the present) are shown in blue circle, red square and green triangle.
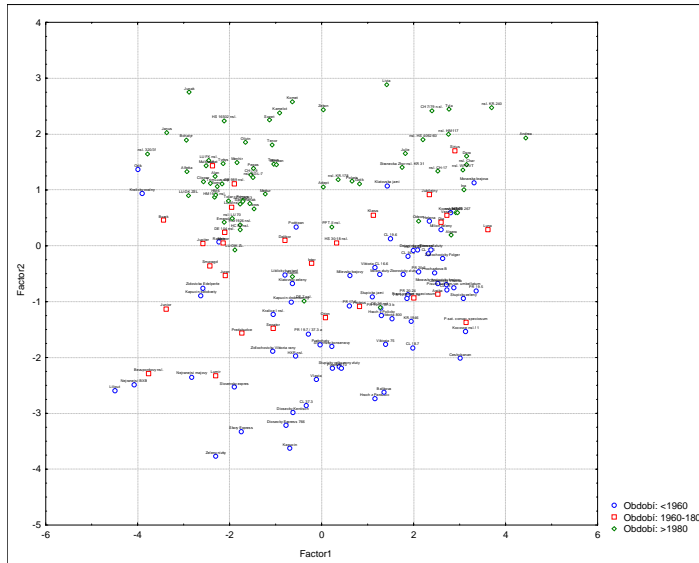
Figure 9

Multi-dimensional scaling (MDS) for combined molecular data. Three breeding periods (I; up to 1960, II; 1970-80, III; 1980 to the present) are shown in blue circle, red square and green triangle.
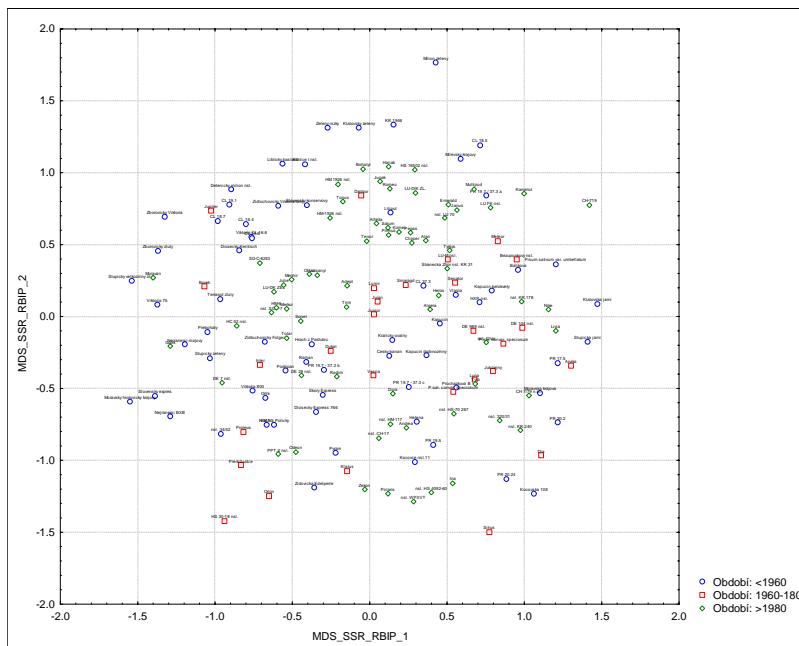
Figure 10

Bayesian model-based analysis of separate and combined morphological and molecular marker data.

Molecular data (RBIP+ SSR) 29 clusters
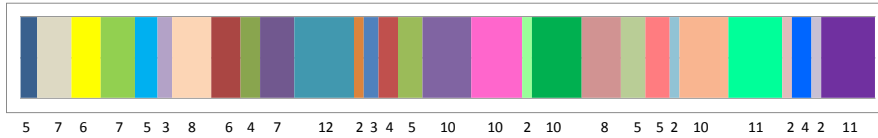
Morphological data 3 clusters (15 qualitative + 8 quantitative)

Combined - morphology + molecular data

Figure 11

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.
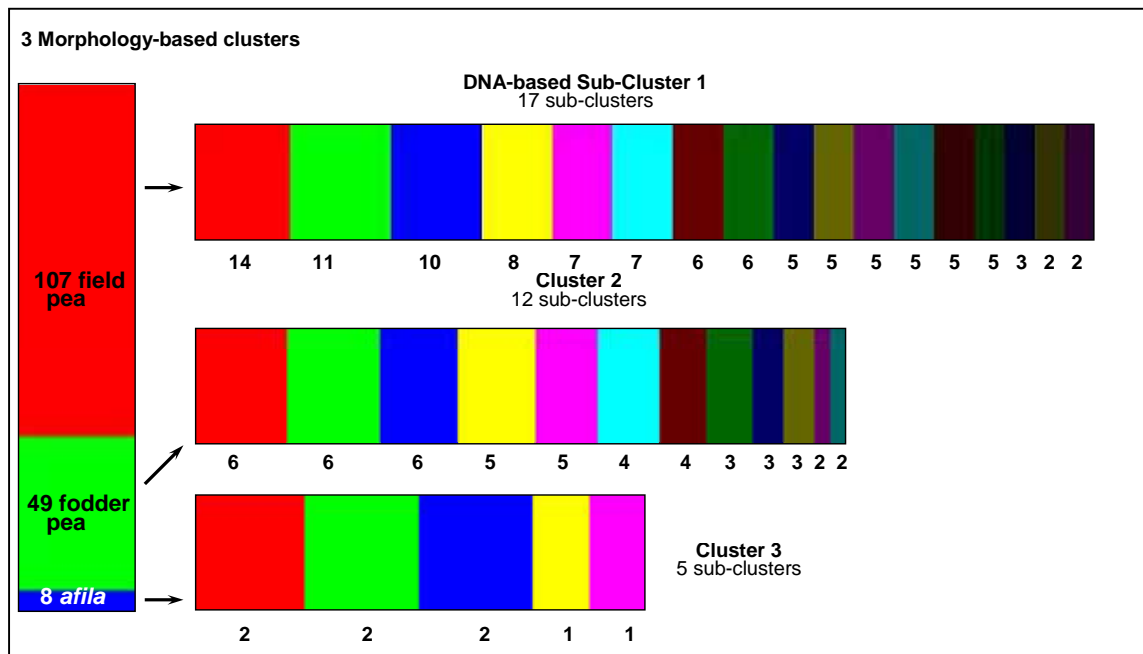
Figure 12

Neighbour joining tree of 29 BAPS clusters derived from composed molecular (RBIP+SSR) data.
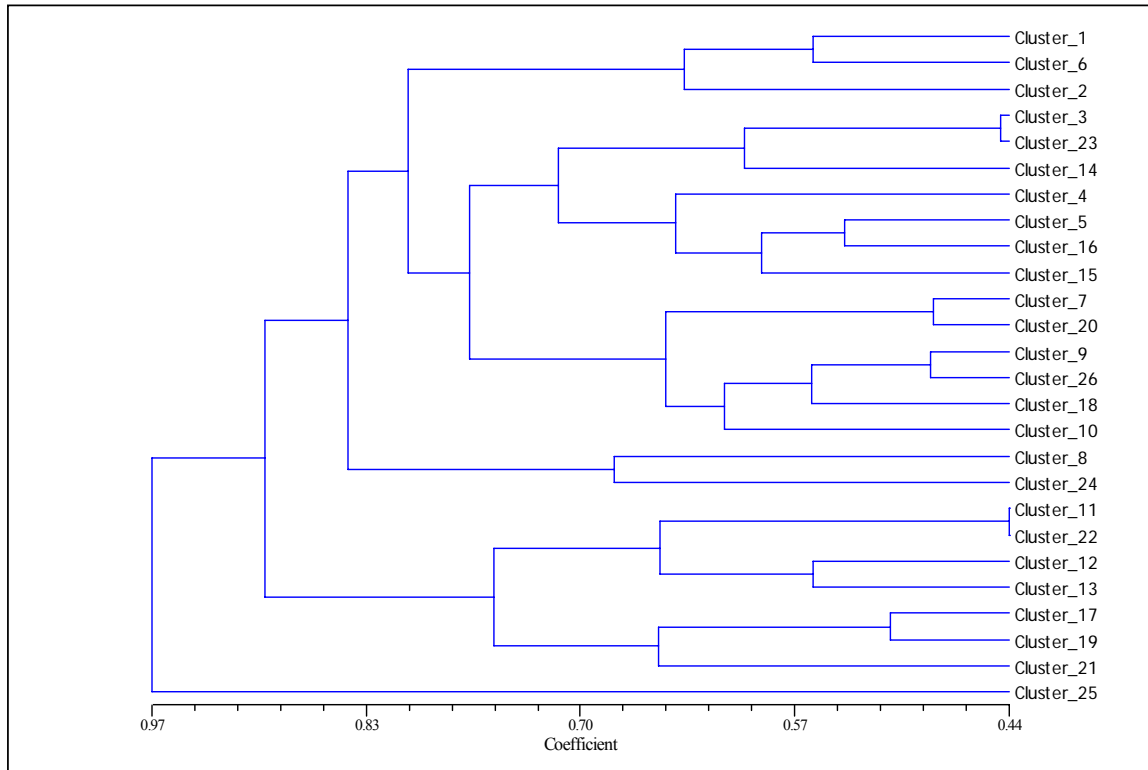
Figure 13

BAPS analysis of entire Czech pea collection (1,283 accessions analyzed by 19 RBIP loci e.g. 56/58 alleles/accession with resulting 9 identified clusters) using Core Hunter software. Each accession is represented by single bar (x axis), with percentage into each identified cluster (from 0 to 100%, y axis).

166 accesions of Czech/Slovak origin (from study of Smýkal et al., 2008a) shown as black bars.
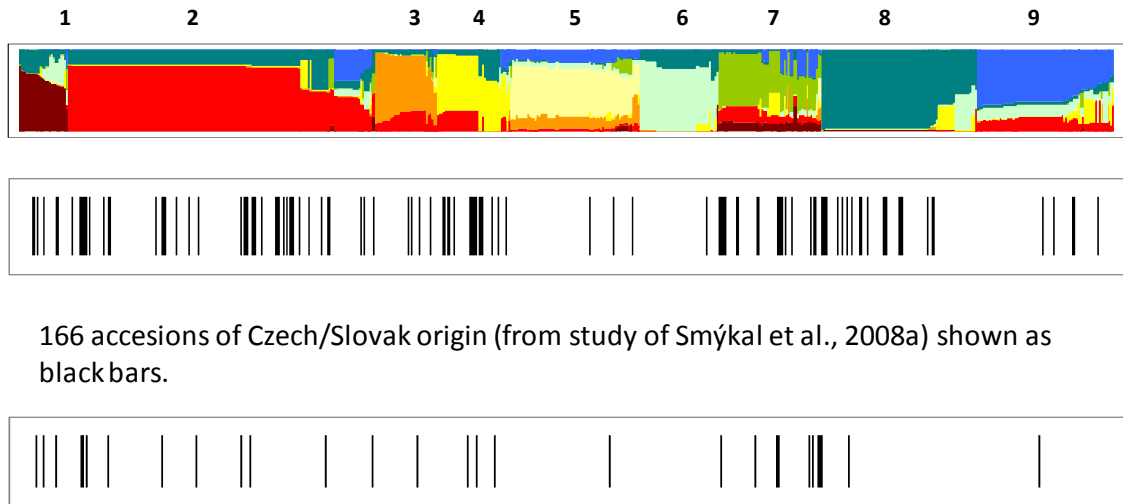
34 core set accesions of Czech/Slovak origin (from study of Smýkal et al., 2008a)

Table 5

Accession numbers of CoreHunter selected multiobjective core collections (10, 20, 30%). Some accessions are arbitrarily coloured for better visualization of their selection by different sampling intensities and markers (L01 versus L02 in pink).

| RBIP+SSR-based | | | SSR-based | RBIP-based | RBIP+SSR-based | RBIP-smallest core | SSR-smallest core | RBIP+SSR-smallest core |
Multi-10%	Multi-20%	Multi-30%	Multi-20%	Multi-20%	Multi-20%	10 acc.	10 acc.	19 acc.
L0100004	L0100002	L0100002	L0100015	L0100031	L0100002	L0100002	L0100015	L0100002
L0100020	L0100004	L0100004	L0100011	L0100002	L0100004	L0100032	L0100020	L0100013
L0100026	L0100013	L0100006	L0100020	L0100003	L0100013	L0100088	L0100026	L0100015
L0100028	L0100015	L0100013	L0100026	L0100004	L0100015	L0100249	L0100028	L0100020
L0100032	L0100020	L0100015	L0100028	L0100006	L0100020	L0100388	L0100345	L0100026
L0100105	L0100026	L0100019	L0100029	L0100017	L0100026	L0100469	L0100530	L0100028
L0100252	L0100028	L0100020	L0100088	L0100019	L0100028	L0100586	L0100535	L0100032
L0100581	L0100031	L0100025	L0100252	L0100028	L0100031	L0200201	L0100603	L0100101
L0100585	L0100032	L0100026	L0100318	L0100032	L0100032	L0200391	L0100927	L0100252
L0100586	L0100083	L0100028	L0100345	L0100082	L0100083	L0200406	L0200202	L0100530
L0100927	L0100088	L0100029	L0100381	L0100088	L0100088		L0200212	L0100585
L0200202	L0100248	L0100031	L0100530	L0100248	L0100248			L0100586
L0200207	L0100252	L0100032	L0100535	L0100251	L0100252			L0100927
L0200209	L0100414	L0100082	L0100549	L0100388	L0100414			L0200202
L0200211	L0100530	L0100083	L0100577	L0100414	L0100530			L0200207
L0200391	L0100581	L0100088	L0100581	L0100469	L0100581			L0200211
	L0100585	L0100101	L0100603	L0100548	L0100585			L0200391
	L0100586	L0100248	L0100687	L0100585	L0100586			L0200397
	L0100687	L0100252	L0100866	L0100586	L0100687			L0200406
	L0100865	L0100345	L0100927	L0100684	L0100865			
	L0100900	L0100381	L0200201	L0100900	L0100900			
	L0100911	L0100388	L0200202	L0100911	L0100911			
	L0100927	L0100414	L0200207	L0200201	L0100927			
	L0200201	L0100469	L0200209	L0200202	L0200201			
	L0200202	L0100530	L0200212	L0200207	L0200202			
	L0200207	L0100548	L0200366	L0200209	L0200207			
	L0200209	L0100581	L0200393	L0200243	L0200209			
	L0200211	L0100585	L0200405	L0200366	L0200211			
	L0200212	L0100586	L0200406	L0200388	L0200212			
	L0200366	L0100684	L0200433	L0200391	L0200366			
	L0200391	L0100687	L0200434	L0200406	L0200391			
	L0200406	L0100865	L0200439	L0200425	L0200406			
		L0100900						
		L0100911						
		L0100927						
		L0200201						
		L0200202						
		L0200207						
		L0200209						
		L0200211						
		L0200212						
		L0200243						
		L0200365						
		L0200366						
		L0200391						
		L0200399						
		L0200406						
		L0200425						
		L0200439						

Figure 14

Visualization of core collections selected by BAPS or Core Hunter (multiobjective 20%, 32 accessions) approaches. **A.)** 3 morphology followed by molecular data BAPS sub clustering (Smýkal et al. 2008), **B.)** SSR_based, **C.)** RBIP_based and **D.)** RBIP+SSR_based Core Hunter. Accessions are shown in order of 29 BAPS clusters derived from molecular data-based.
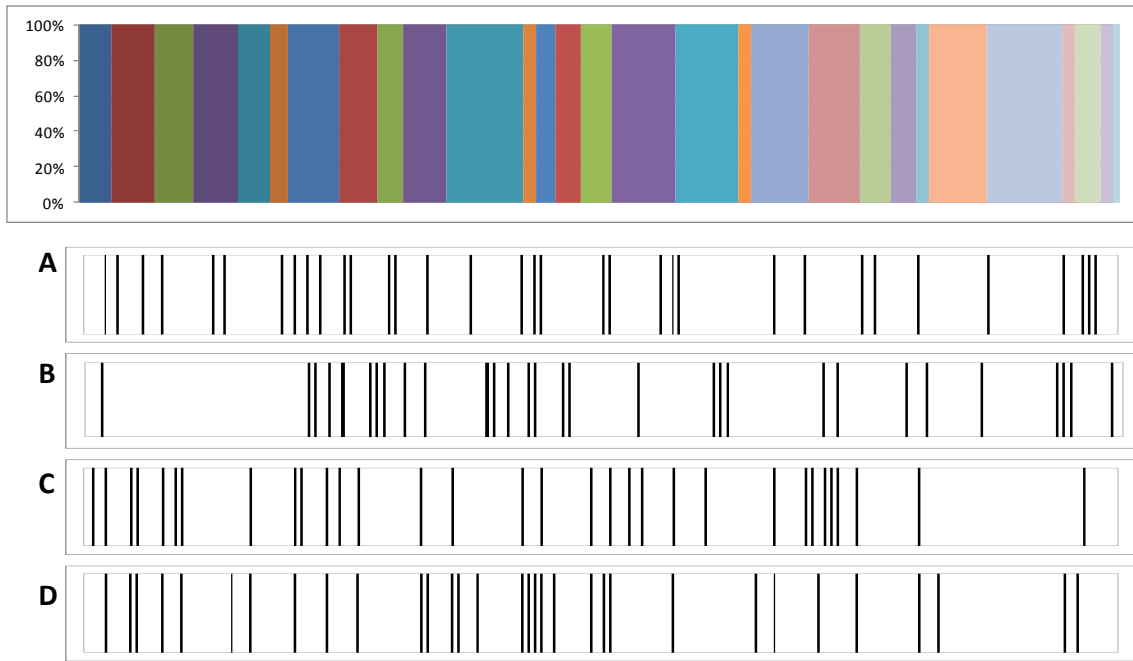

Figure 15

Minimal size core collections selected by Core Hunter (smallest size) approach. Accessions are visualized in order by 29 molecular data-based BAPS clusters.
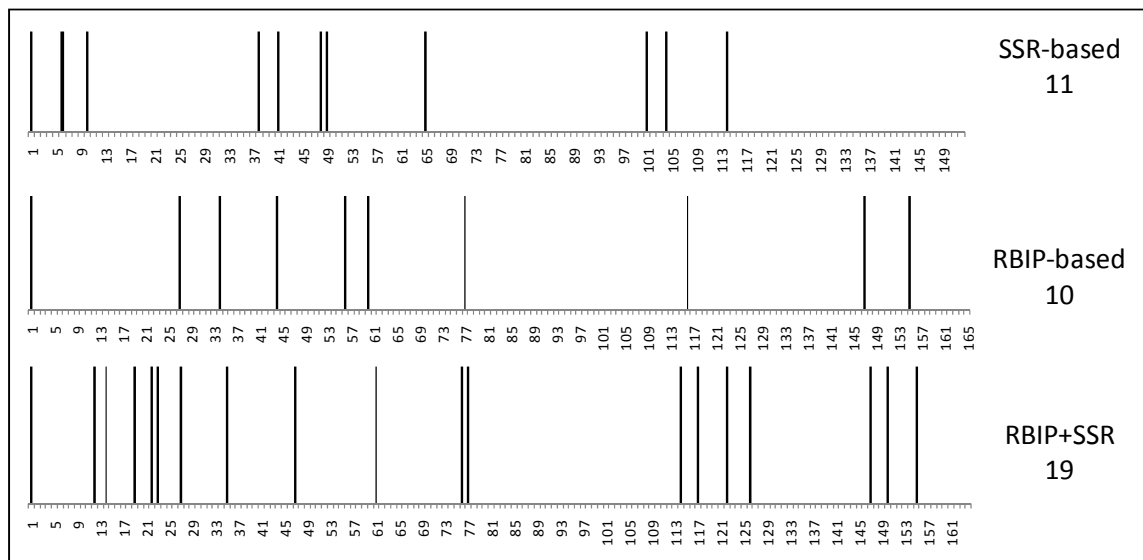

Table 6

Diversity indexes of Core Hunter identified core collections based on SSRs, RBIP and combined datasets.

Dataset	Number of accessions	Modified Roger´s distance	Shannon index	Dataset	Number of accessions	Modified Roger´s distance	Shannon index	Dataset	Number of accessions	Modified Roger´s distance	Shannon index
SSR-based	166 (100%)	0.3311	3.1456	RBIP-based	166 (100%)	0.34	3.711	RBIP+SSR-based	166 (100%)	0.33094	4.12501
MR-derived	33(20%)	0.3702	3.3175	MR-derived	33(20%)	0.374	3.824	MR-derived	33(20%)	0.36565	4.22457
SH-derived	33(20%)	0.3702	3.4074	SH-derived	33(20%)	0.373	3.827	SH-derived	33(20%)	0.36298	4.23544
CV-derived	33(20%)	0.3508	3.2846	CV-derived	33(20%)	0.332	3.706	CV-derived	33(20%)	0.34679	4.17917
Multiobjective	33(20%)	0.3673	3.4074	Multiobjective	33(20%)	0.371	3.839	Multiobjective	33(20%)	0.36237	4.23531
smallest-core	11	0.3714	3.3787	smallest-core	10	0.367	3.801	smallest-core	19	0.35847	4.21511

Table 7

Diversity indexes of Core Hunter (only molecular data, 20% multiobjective) and BAPS (morphology followed by molecular sub clustering) derived core collections.

RBIP Locus	Gene Diversity			Allelic Richness per locus			Shannon index		
	Entire (166)	Core-BAPS (34 acc.)	Core Hunter (33acc.,20%)	Entire (166)	Core-BAPS (34 acc.)	Core Hunter (33acc.,20%)	Entire (166)	Core-BAPS (34 acc.)	Core Hunter (33acc.,20%)
1006-X19	0.468	0.553	0.587	3.680	2.961	3.320	0.73	0.82	0.77
399-14-9	0.499	0.511	0.523	2.000	2.000	2.000	0.72	0.72	0.72
Birte-X5	0.508	0.529	0.556	2.766	2.472	2.530	0.72	0.78	0.82
45-x31	0.667	0.676	0.682	3.999	3.000	3.560	0.98	0.99	0.99
64-X45	0.525	0.430	0.531	2.913	2.961	2.899	0.28	0.35	0.42
281-X40	0.147	0.205	0.243	2.000	2.000	2.000	0.36	0.42	0.51
2055-NR51	0.271	0.328	0.336	3.504	2.472	3.320	1.02	0.99	1.01
95-X2	0.238	0.337	0.389	3.376	2.472	3.180	0.42	0.51	0.62
2201Cycl-6	0.224	0.194	0.284	3.714	2.854	3.579	0.44	0.38	0.42
1074Cycl-22	0.236	0.227	0.235	3.900	2.995	3.210	0.47	0.42	0.47
281-X44	0.419	0.473	0.489	2.766	2.472	2.560	0.72	0.82	0.95
Birte X16	0.443	0.516	0.518	3.998	2.999	3.300	0.75	0.82	0.95
RBIP-3	0.478	0.536	0.542	3.913	3.000	3.320	0.83	0.88	0.89
RBIP-4	0.264	0.368	0.389	2.000	2.000	2.000	0.43	0.56	0.62
RBIP-7	0.367	0.383	0.423	2.767	2.854	2.790	0.60	0.56	0.58
Birte-B1	0.385	0.412	0.419	2.000	2.000	2.000	0.65	0.57	0.64
Mean	**0.384**	**0.417**	**0.447**	**3.081**	**2.595**	**2.848**	**0.63**	**0.66**	**0.71**

SSR Locus	Entire (166)	Core-BAPS (34 acc.)	Core Hunter (33acc.,20%)	Entire (166)	Core-BAPS (34 acc.)	Core Hunter (33acc.,20%)	Entire (166)	Core-BAPS (34 acc.)	Core Hunter (33acc.,20%)
AD-270	0.783	0.809	0.810	6.379	6.184	6.370	1.59	1.46	1.59
A-9	0.67	0.674	0.678	3.000	3.000	3.000	1.09	1.03	1.05
B-14	0.691	0.681	0.695	4.000	4.000	4.000	1.26	1.36	1.26
AD-237	0.764	0.73	0.762	6.720	6.660	6.720	1.60	1.34	1.60
A-278	0.56	0.63	0.664	3.854	3.851	3.854	0.98	0.67	0.98
AD-141	0.757	0.723	0.756	7.491	7.259	7.489	1.68	1.24	1.66
AB-65	0.149	0.121	0.168	2.959	2.956	5.600	0.36	0.35	0.35
AD-186	0.722	0.671	0.756	7.380	6.467	7.368	1.43	1.12	1.42
AA-67	0.528	0.544	0.647	3.616	3.467	3.605	0.89	0.95	0.88
AA-121	0.662	0.598	0.623	5.126	5.000	5.116	1.23	0.98	1.22
AC-75	0.423	0.401	0.423	4.211	3.780	4.210	0.97	0.85	0.95
AB-100	0.398	0.385	0.402	4.220	3.569	4.200	0.89	0.75	0.89
AA-175	0.742	0.741	0.734	5.238	5.166	5.226	0.96	0.86	0.96
AA-163	0.556	0.553	0.543	5.232	4.723	5.226	1.15	1.09	1.15
AB-23	0.467	0.459	0.526	6.563	5.641	6.504	1.23	1.11	1.23
AA-335	0.125	0.116	0.243	2.897	2.658	2.890	0.86	0.56	0.85
AA-456	0.756	0.732	0.788	4.983	4.023	4.978	0.88	0.79	0.88
Mean	**0.574**	**0.563**	**0.601**	**4.933**	**4.612**	**5.080**	**1.12**	**0.97**	**1.11**

Table 8

Shannon index of all 45 scored morphological traits in entire in comparison to Core Hunter (only molecular data, 20% multiobjective) and BAPS (morphology followed by molecular sub clustering) derived core collections.

Shannon index

Trait	Stem shape	Stem length	Length to 1-st prod. node	Length internode uder the 1-st prod.node	No. of sterile nodes	Type of branching	Branching at base	Leafs presence	Leaf type	Leaflet shape	Leaflet margin shape on the second realleaf	Leaflet margin shape at first flowering node	Leaflet apex shape	Leaflet colour	Leaflet waxy bloom	Stipules size	Stipules spoted	Character of anthocyanin spot	Tendrils shape	No. of flowers in raceme	Wexillum size	Base wexillum shape	Wexillum apex
Entire (166)	1.56	1.75	1.78	0.89	0.22	1.20	0.56	1.42	0.22	1.62	1.35	1.23	0.97	1.41	0.89	1.23	2.34	0.64	0.56	0.15	0.84	0.26	0.68
Core-BAPS (34 acc.)	0.98	1.23	1.23	0.87	0.13	0.99	0.48	1.35	0.25	1.53	1.23	0.98	1.12	1.42	0.87	1.08	1.97	0.68	0.65	0.12	1.23	0.56	0.56
Core Hunter (33acc., 20%)	0.87	0.98	1.56	0.76	0.16	0.87	0.56	1.40	0.16	0.87	0.89	1.02	0.74	0.93	0.46	1.02	2.04	0.56	0.47	0.11	0.81	0.13	0.42

Shannon index

Trait	Wexillum colour	Wings shape	Wings colour	Degree of carina keel development	Callyx sepala-termination of upper pair	Pod-parchment coating	Pod-degree of curving	Pod appex shape	Pod length	Pod width	Pod colour in green ripeness	Ripe seed surface	Seed colour at green ripeness	Seed colour at full ripeness	Seed testa colour	Seed hilum colour	Seed funiculus stability	Seed cotyledons colour	Seeds weight/ plant	Pods no./ plant	Seeds no./ plant	Thousand seed weight (TSW)	Mean
Entire (166)	1.29	0.54	1.43	0.79	0.56	0.17	2.46	0.47	2.56	2.13	1.89	0.82	0.11	1.67	0.68	0.46	0.00	1.03	2.35	1.89	2.68	2.47	**1.16**
Core-BAPS (34 acc.)	1.36	0.43	1.42	0.46	0.59	0.12	2.17	0.42	2.43	1.89	1.54	0.77	0.09	1.80	0.71	0.47	0.00	1.06	2.01	1.45	1.45	2.13	**1.05**
Core Hunter (33acc., 20%)	1.13	0.46	1.27	0.77	0.51	0.12	2.13	0.33	2.24	2.06	1.76	0.77	0.11	1.49	0.44	0.33	0.00	1.01	2.14	1.48	2.32	2.17	**0.95**