© NIAB 2010
ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2011) **9**(1); 4–18
doi:10.1017/S147926211000033X

Phylogeny, phylogeography and genetic diversity of the *Pisum* genus

Petr Smýkal[1]*, Gregory Kenicer[2], Andrew J. Flavell[3], Jukka Corander[4], Oleg Kosterin[5], Robert J. Redden[6], Rebecca Ford[7], Clarice J. Coyne[8], Nigel Maxted[9], Mike J. Ambrose[10] and Noel T. H. Ellis[10]

[1]*Agritec Plant Research Limited, Department of Biotechnology, Zemedelská 2520/16, CZ-787 01 Šumperk, Czech Republic,* [2]*Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, UK,* [3]*Division of Plant Sciences, University of Dundee at SCRI, Invergowrie, Dundee DD2 5DA, UK,* [4]*Department of Mathematics, Abo Akademi University, Biskopsgatan 8, FIN-20500 Åbo, Finland,* [5]*Institute of Cytology and Genetics, Siberian Department of Russian Academy of Sciences, 630090 Novosibirsk, Russia,* [6]*Australian Temperate Field Crops Collection, Horsham VIC 3401, Australia,* [7]*Melbourne School of Land and Environment, The University of Melbourne, Victoria 3010, Australia,* [8]*USDA – Agricultural Research Service, WSU, Pullman WA99164, USA,* [9]*School of Biosciences, University of Birmingham, Birmingham B15 2TT, UK and* [10]*John Innes Centre, Colney, Norwich NR4 7UH, UK*

Abstract

The tribe *Fabeae* (formerly *Vicieae*) contains some of humanity's most important grain legume crops, namely *Lathyrus* (grass pea/sweet pea/chickling vetches; about 160 species); *Lens* (lentils; 4 species); *Pisum* (peas; 3 species); *Vicia* (vetches; about 140 species); and the monotypic genus *Vavilovia*. Reconstructing the phylogenetic relationships within this group is essential for understanding the origin and diversification of these crops. Our study, based on molecular data, has positioned *Pisum* genetically between *Vicia* and *Lathyrus* and shows it to be closely allied to *Vavilovia*. A study of phylogeography, using a combination of plastid and nuclear markers, suggested that wild pea spread from its centre of origin, the Middle East, eastwards to the Caucasus, Iran and Afghanistan, and westwards to the Mediterranean. To allow for direct data comparison, we utilized model-based Bayesian Analysis of Population structure (BAPS) software on 4429 *Pisum* accessions from three large world germplasm collections that include both wild and domesticated pea analyzed by retrotransposon-based markers. An analysis of genetic diversity identified separate clusters containing wild material, distinguishing *Pisum fulvum*, *P. elatius* and *P. abyssinicum*, supporting the view of separate species or subspecies. Moreover, accessions of domesticated peas of Afghan, Ethiopian and Chinese origin were distinguished. In addition to revealing the genetic relationships, these results also provided insight into geographical and phylogenetic partitioning of genetic diversity. This study provides the framework for defining global *Pisum* germplasm diversity as well as suggesting a model for the domestication of the cultivated species. These findings, together with gene-based sequence analysis, show that although introgression from wild species has been common throughout pea domestication, much of the diversity still resides in wild material and could be used further in breeding. Moreover, although existing collections contain over 10,000 pea accessions, effort should be directed towards collecting more wild material in order to preserve the genetic diversity of the species.

Keywords: Bayesian inference; core collection; domestication; genetic diversity; germplasm; microsatellite; pea; phylogeny; *Pisum*; retrotransposon

* Corresponding author. E-mail: smykal@agritec.cz

Introduction – domestication of pea

Pea (*Pisum sativum* L.) is one of the world's oldest domesticated crops. It is the third most widely grown legume, as its seeds serve as a protein-rich food for humans and livestock alike. Domesticated about 10,000 years ago (Ambrose, 1995; Zohary and Hopf, 2000), pea is currently cultivated in temperate zones worldwide. Centuries of selection and breeding have resulted in thousands of pea varieties many of which are maintained in numerous germplasm collections worldwide (Smýkal *et al.*, 2008b). Pea (*P. sativum* L.) was used in the earliest of genetic studies, most famously by Mendel (1866) and previously by Knight (1799). However, owing to its large genome size (4000 Mb) and the high occurrence of repetitive sequences (Macas *et al.*, 2007), much of the recent progress in molecular genetics and genomics has not been conducted on pea.

Pisum within tribe *Fabeae*

Reconstructing the phylogenetic relationship of the *Leguminosae* is essential to understanding the origin and diversification of this economically and ecologically important family. The monophyly of the family (*Leguminosae/Fabaceae*) as a natural group has never been in doubt, but it was not until the phylogenetic analyses of groups such as Kass and Wink (1996, 1997) and Doyle *et al.* (1997) that the group's monophyly was demonstrated through molecular DNA sequence data. Since then, molecular phylogenetic research has provided a solid understanding of relationships at all levels in the family (Lewis *et al.*, 2005). Tribe *Fabeae* (syn. *Vicieae*) is considered one of the most advanced groups in the legumes (Kupicha, 1981; Steele and Wojciechowski, 2003; Wojciechowski *et al.*, 2004; Lock and Maxted, 2005), and one of the most recently evolved. Estimates based on rates of evolution in the *maturase* K (*matK*) chloroplast gene place the age of the crown clade at 17.5 Mya in the mid-Miocene (Lavin *et al.*, 2005). The centre of diversity and posited area of origin is the Eastern Mediterranean (Kupicha, 1981; Kenicer, 2007). The tribe contains five genera, including *Vicia* with most of the ancient Old-World grain legume crops: *Lathyrus* (grass pea/sweet pea; about 160 species); *Lens* (lentils; 4 species); *Pisum* (peas; 3 species); *Vicia* (vetches; about 140 species) (Steele and Wojciechowski, 2003; Lock and Maxted, 2005; Endo *et al.*, 2008; Kenicer *et al.*, 2008; Smýkal *et al.*, 2009a) (Fig. 1).

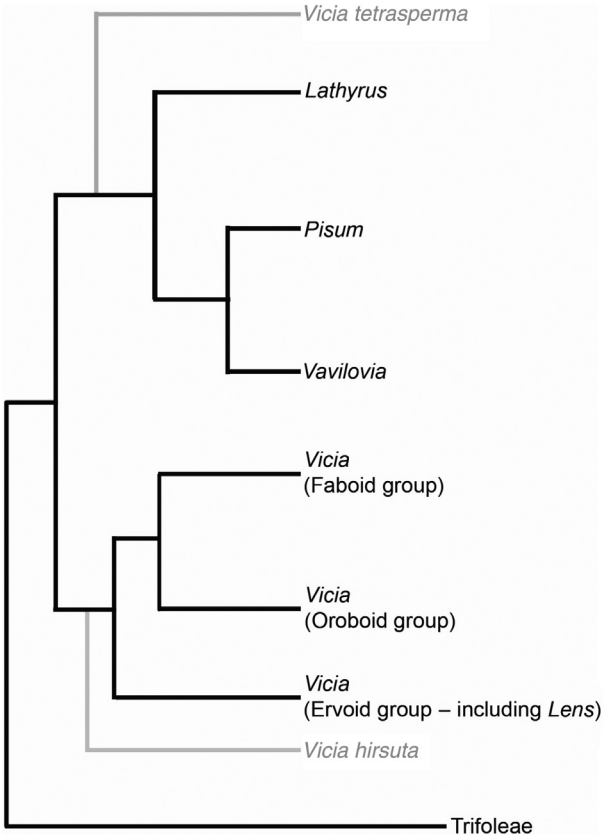

Fig. 1. Phylogeny of *Fabeae* tribe, based on chloroplast and ITS DNA sequence data.

Morphology-based classifications of *Pisum*

The classification of *Pisum* L. based on morphology and karyology has changed over time from a genus with five species (Govorov, 1937) to a monotypic genus (Lamprecht, 1966; Marx, 1977). While Davis (1970) recognized two species, *P. fulvum* Sibth. & Sm. and *P. sativum* L., both native to Turkey, he did not consider the third putative species *P. abyssinicum* A. Br., which is endemic to Yemen and Ethiopia. Subsequently, the nomenclature of the group is complex, and numerous names have been proposed for wild representatives of *P. sativum*. However, only three have been used to denote taxa of subspecies or species rank: *P. elatius* Bieb. (Bieberstein, 1808), *P. humile* Boiss. & Noe and *P. syriacum* Boiss. & Noe. (Makasheva, 1979). *P. elatius* was classified as a subspecies first by Schmalhausen (1895), although many authors ascribe this to Ascherson and Graebner (1910). *P. humile* was described by Boissier and Noe (1856) and given a name used earlier by Miller (1768) for a form of cultivated pea. Berger (1928) downgraded the rank to subspecies and gave it a new name: *P. sativum* subsp. *syriacum* (Boissier and Noe) Berger. Its status was again raised to species by Lehmann (1954), though this remained unsupported. Crossing experiments undertaken by Ben-Ze'ev and Zohary (1973) partially clarified relationships among four species recognized by Boissier (1856) – *sativum, elatius, humile* and *fulvum* – while

wider hybridization experiments between *Lathyrus* and *Pisum* species have shown cross-incompatibility (Ochatt *et al.*, 2004). The domestication of cultivated pea from northern populations of '*humile*' was proposed by Ben-Ze'ev and Zohary (1973), but the source could just as likely be the 'northern elatius' (Kosterin *et al.*, 2010). A thorough description of the genus was performed by Makasheva (1979) based on morphological, ecological and some biochemical data. This placed *Pisum* together with *Vicia* and *Lathyrus.* The ancestor of *Vavilovia formosa* was placed as the last common ancestor for all three genera, from which an extinct perennial and later an annual *Pisum* ancestor evolved (Fig. 2). The more recent and most used classification of Maxted and Ambrose (2000) adopted three species:

- *P. sativum* L.
    - Subsp. *sativum* (includes var. *sativum* and var. *arvense*)
    - Subsp. *elatius* (Bieb.) Aschers. & Graebn (includes var. *elatius,* var. *brevipedunculatum* and var. *pumilio*)
- *P. fulvum* Sibth. & Sm.
- *P. abyssinicum* A. Br.

This classification is accepted in this paper.

The taxonomic position of *P. abyssinicum* is often discussed, namely whether this lineage has diverged sufficiently from other taxa to be considered a separate species or whether it should be placed within *P. sativum* as a subgroup (Maxted and Ambrose, 2000). Based on morphological characteristics, Govorov (1937) labelled it a separate cultivated species, while Makasheva (1979) regarded it as a subspecies. A serious karyologic barrier for crossing to *P. sativum* (Ben-Ze'ev and Zohary, 1973) and clear-cut phenotypic differences support the view of its species status (Lamprecht, 1963). Although its origin is not fully understood, it has been proposed that it was domesticated independently 4000–5000 years ago in Early or Middle-Kingdom Egypt (Vershinin *et al.*, 2003; Jing *et al.*, 2010).

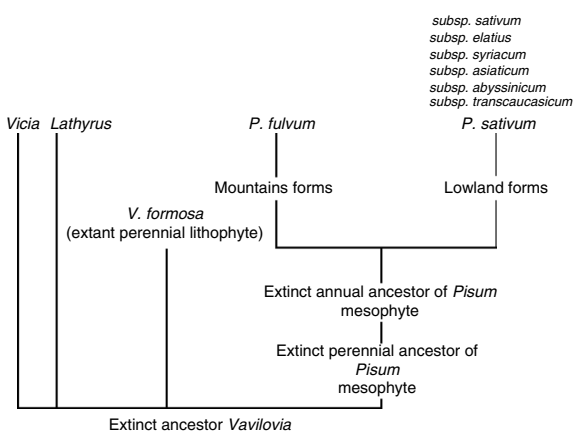

Fig. 2. Hypothetical origin of *Pisum,* according to Makasheva (1979).

Pisum classification based on molecular data

Early data from electrophoretic patterns of albumin and globulin (Waines, 1975) and chloroplast DNA polymorphism (Palmer *et al.*, 1985) have separated *P. fulvum* as a distinct species and *P. sativum* as an aggregate of *'humile', P. elatius* and *P. sativum*. Recent phylogenetic studies based on retrotransposon insertion markers support the model of *P. elatius* as a paraphyletic group, within which all *P. sativum* is nested (Vershinin *et al.*, 2003; Jing *et al.*, 2005, 2010). The study by Hoey *et al.* (1996) using morphological, allozyme and random amplification of polymorphic DNA (RAPD) characteristics on a set of Ben-Ze'ev and Zohary (1973) accessions resulted in separation of *P. fulvum* and 'southern' *humile*, while cultivated peas were grouped among *P. elatius* accessions. The position of 'northern' *humile* varied between a sister group to cultivated peas and *P. elatius.* More recently, studies of internal transcribed spacer (ITS) sequence variation have supported this (Saar and Polans, 2000; Polans and Saar, 2002). Extensive phylogenetic relationships between pea forms were reconstructed by Ellis *et al.* (1998), Pearce *et al.* (2000) and Vershinin *et al.* (2003) using both amplified fragment length polymorphism (AFLP) and its derived retrotransposon insertion-based marker method, sequence-specific amplification polymorphisms (SSAP). Using these approaches, *P. fulvum* and *P. abyssinicum* formed neighbouring but separate branches, a subset of *P. elatius* was positioned between *P. fulvum and P. abyssinicum,* and further branches were found within cultivated pea. The most recent studies of *P. abyssinicum* place it between *P. fulvum* and a subset of *P. elatius* (Vershinin *et al.*, 2003; Jing *et al.*, 2010) and showed very low diversity in molecular analyses, which could be explained by passage through a bottleneck. A general feature of molecular phylogenetic analysis of *Pisum* has been the impact of introgression on pea diversity and evolution (Jing *et al.*, 2007). Moreover, good conservation between SSAP (Vershinin *et al.*, 2003), retrotransposon insertions (Jing *et al.*, 2005) and gene-based derived (Jing *et al.*, 2007) trees was observed, in spite of the fact that they derive from different genomic components. The gene-based study showed that *Pisum* is a diverse genus with one polymorphic site every 15 bp on average. Linkage disequilibrium (LD) analysis has suggested that, owing to recombination,

different genetic loci display very different pictures of genetic diversity. Another study on relationships among wild *Pisum* forms used a combination of mitochondrial, chloroplast and nuclear genome markers (Kosterin and Bogdanova, 2008; Kosterin *et al*., 2010), separating *P. fulvum* and *P. abyssinicum* accessions and about half of those of wild *P. sativum* from the rest of the wild and all cultivated *P. sativum*. The distinction within *P. sativum* coincided with the cytogenetic classes of Zohary and Ben-Ze'ev (1973). However, a comparison of results between different phylogenetic analyses is limited and difficult due to differences in studied accessions as well as markers. Moreover, incomplete information on taxonomic attribution and the origin of wild accessions hinder such studies.

Phylogeography of the *Pisum* genus

The geographical range of wild representatives of *P. sativum* extends from Iran and Turkmenistan through Anterior Asia, northern Africa and southern Europe (Makasheva, 1979; Maxted and Ambrose, 2001; Maxted *et al*., 2010). However, due to their early cultivation, it is often difficult to identify the precise location of the centre of diversity, especially considering that large parts of the Mediterranean region and Middle East have been substantially modified by human activities and changing climatic conditions. Moreover, reliable and thorough passport data are often missing or incomplete, especially for valuable older acquisitions gained through expeditions. Thus, some so-called 'wild accessions' may simply have escaped cultivation. Furthermore, as in other crops, wild species are often found in secondary habitats as weeds and in direct contact with domesticated pea (sympatric), resulting in spontaneous hybridizations between cultivars and wild forms (Ben Ze'ev and Zohary, 1973). As stated earlier, it is widely accepted that the genus *Pisum* contains the clear-cut and rather homogenous wild species *P. fulvum* Sibth. et Smith. found in Jordan, Syria, Lebanon and Israel. It also contains cultivated subspecies *P. sativum* subsp. *abyssinicum* A. Br. from Yemen and Ethiopia (Westphal, 1974), which was domesticated independently *of P. sativum* (Jing *et al*., 2010). Lastly, the *Pisum* genus contains a large and loose aggregate of both wild (*P. sativum* subsp. *elatius*) and cultivated forms that comprise the wild species *P. sativum* L. in a broad sense. Both *P. fulvum* and *P. abyssinicum* differ from *P. sativum* by several chromosomal rearrangements, which make them nearly incompatible with *P. sativum*. Hybridization between them is also hampered by nuclear–cytoplasmic conflict (Bogdanova, 2007; Bogdanova *et al*., 2009). Analysis using three dimorphic nuclear, plastid or mitochondrial markers was performed, and four contrasting combinations of alleles (lineages A to D) were introduced (Kosterin and Bogdanova, 2008; Kosterin *et al*., 2010). These authors proposed a scenario for the evolution of wild *P. sativum* and its domestication in which the ancestral state of the genus (combination A) originated in the eastern Mediterranean, based on the present area of this lineage in Israel, Lebanon, Syria and southern Turkey. Here, *P. sativum* grows often sympatrically with *P. fulvum*, which also has combination A. *P. abyssinicum*, another taxon with exclusively combination A, occurs in Yemen and Ethiopia. It was proposed that the westward spread of lineage A occurred during the Pleistocene, when the sea occupied less area. The accessions with combination A found on Sardinia and Menorca are thought to represent island refugia of early spread. During this westward dispersal, lineage C appeared and spread over the central and western Mediterranean areas and northeastern Africa. The representatives of lineage D were found in Egypt (cultivated), Sicily and Turkey (wild), while the lineage B was located near the Black Sea (Kosterin *et al*., 2010) (Fig. 3). Thus, Asia Minor was an area affected by two opposite spreading waves of peas: that of lineage A from the south and that of lineage B from the north. It was suggested that it was in the West and/or Central Mediterranean where the transition between lineages A and B took place, and that this transition left intermediate descendants (Kosterin *et al*., 2010). Jing *et al*. (2010), using retrotransposon markers,

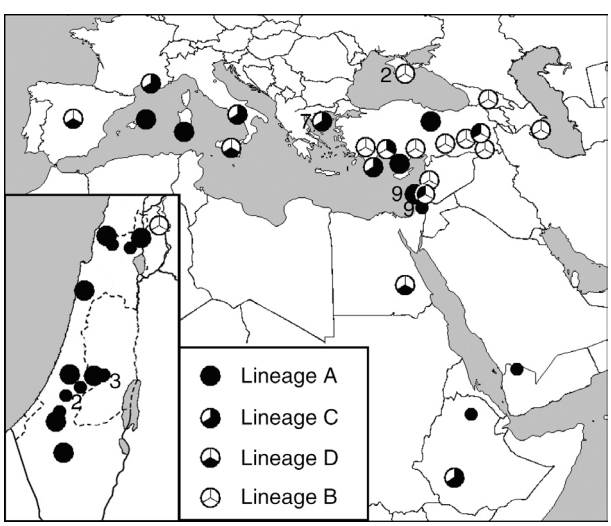

Fig. 3. Phylogeography of *P. fulvum, P. abyssinicum* (small circles) and wild *P. sativum* subsp. *elatius*) accessions with indication of alleles of the three markers studied (taken from Kosterin *et al*., 2010). Lineage A, Anterior Asia, islands (*cox1* +, *rbcL* + and *SCA*^{f}); lineage B, Tauro-Caucasian area, Turkey (*cox1-, rbcL-* and *SCA*^{s}); lineage C, Mediterranean (France and Greece), Ethiopia (*cox1-, rbcL* + and *SCA*^{f}); lineage D, Egypt, Sicily, Spain (*cox1-, rbcL-* and *SCA*^{f}).

proposed a related model whereby a subset of *P. elatius* was selected by early farmers in the Fertile Crescent and grown extensively, thus broadening its distribution across Southern Eurasia and additionally differentiating in two opposite directions. The first of these was an expansion eastwards into the Indian subcontinent and the Himalayan regions, subsequently giving rise to the diverged Afghan *P. sativum* ecotypes found today. A second proposed diversification of another strand of *P. elatius*-derived primitive *P. sativum* was the main domestication route that gave rise to the mainstream of modern cultivated *Pisum*. Jing *et al.* (2010) further concluded that *P. abyssinicum* derived from a cross between *P. fulvum* and a third subset of the diverse *P. elatius* species in the western half of the Fertile Crescent. A small sample was then transferred by humans to northeastern Africa (introducing the bottleneck mentioned above), where it was developed into the modern *P. abyssinicum*.

Genetic diversity of *Pisum* germplasm collections

Accessions of pea have been collected and maintained within several major collections worldwide (Smýkal *et al.*, 2008b). These include the John Innes Centre (JIC), UK (3557 accessions); the Nordic GeneBank, Sweden (2724 accessions); the United States Department of Agriculture (USDA), USA (5404 accessions); the International Center for Agricultural Research in the Dry Areas (ICARDA), Syria (6105 accessions); Instituto del Germoplasma, Bari, Italy (4297 accessions); Leibnitz Institute of Plant Genetics and Crop Plant Research, Germany (5336 accessions); the Australian Temperate Field Crops Collection (ATFC), Australia (6567 accessions); the Vavilov Institute of Plant Breeding, Russia (6790 accessions); and the National Genebank of China, China (3837 accessions). Simple sequence repeats (SSRs or microsatellites) have been popular for assessing *Pisum* diversity because of their high polymorphism and information content, co-dominance and reproducibility (Burstin *et al.*, 2001; Ford *et al.*, 2002; Baranger *et al.*, 2004; Loridon *et al.*, 2005; Ta'ran *et al.*, 2005; Smýkal *et al.*, 2008a; Zong *et al.*, 2009; Nasiri *et al.*, 2009). On the other hand, microsatellites have high mutation rates (Vigouroux *et al.*, 2002; Raquin *et al.*, 2008) and suffer from homoplasy (Bhargava and Fuentes, 2010), e.g. a state in which alleles are identical, but not identical by descent. Also, SSR primers are often difficult to transfer for assessing the relationships among related taxa, as previously shown between *P. sativum* and *P. fulvum* (Ford *et al.*, 2002). As the *Pisum* genus is very diverse, this suggests that the risk of homoplasy in wide surveys of pea germplasm using microsatellites is high. Other marker types used for diversity studies include retrotransposon-based methods, such as SSAP (Ellis *et al.*, 1998) and inter-retrotransposon amplified polymorphism (Kalendar and Schulman, 2006; Smýkal, 2006). Both suffer from a dominant nature and show band intensity variation, leading to reproducibility problems. Alternatively, insertion site polymorphism of the *PDR1 Ty1-copia* group retrotransposon (Lee *et al.*, 1990) has been investigated by SSAP linkage and diversity analysis (Ellis *et al.*, 1998; Pearce *et al.*, 2000). Ellis *et al.* (1998) found that AFLP and SSAP methods were in strong agreement, but AFLP overestimated variation. An alternative marker system based on scoring presence and absence of individual retrotransposon insertions (RBIP) gives greater power for phylogeny and genetic relationship studies in pea and is suitable for in-depth phylogeny and germplasm diversity studies (Jing *et al.*, 2005, 2010). Jing *et al.* (2007) showed good correlation among SSAP (Vershinin *et al.*, 2003) and RBIP (Jing *et al.*, 2005) studies by assessing single nucleotide polymorphisms (SNPs) in 49 genes. All of the mentioned markers and trait loci were used to develop and integrate genetic maps (Ellis and Poyser, 2002; Loridon *et al.*, 2005). Recently, opportunities have arisen through advances in the sequencing of model legumes *Medicago truncatula and Lotus japonica*. The synteny between these genomes and that of *Pisum* has been demonstrated by functional mapping (Aubert *et al.*, 2006). Many molecular studies have indicated that *Pisum* is very diverse and that the structured diversity reflects taxonomic identifiers, ecogeography and breeding gene pools. These studies show the pattern of diversity within which *Pisum* is consistent with the taxonomic scheme of Maxted and Ambrose (2001), with the exception of '*elatius*' ranked as a subspecies of *P. sativum,* rather than of equal rank. *P. elatius* in either sense includes a greater diversity than *P. sativum* subsp. *sativum,* likely due to the fact that *P. sativum* subsp. *sativum* is the cultigen, domesticated from a wild ancestor, probably a type (or types) of *P. elatius*. It would seem more reasonable to position *P. sativum* subsp. *sativum* subordinate to *P. elatius*. Moreover, *Pisum* is capable of genetic exchange (Maxted and Ambrose, 2001), as supported by studies of Jing *et al.* (2005, 2010) and Vershinin *et al.* (2003), which showed that allelic introgression between very diverse material occurs, suggesting the view of *Pisum* as one species.

Using the molecular methods, several major world pea germplasm collections have been analyzed and core collections were formed. In summary, over 2000 accessions of the Chinese collection have been analyzed by 21 SSR loci (Zong *et al.*, 2009); 310 USDA pea accessions have been assessed by 37 RAPD and 15 SSR markers (Coyne *et al.*, 2005 and unpublished). Similarly, The French National Institute for Agricultural Research used 121 protein and SSR markers to genotype 148 accessions

(Baranger *et al.*, 2004; Loridon *et al.*, 2005), and the Crop Development Centre Canada pea collection (~100 accessions) was studied by RAPD, Inter simple sequence repeats and SSR (Ta'ran *et al.*, 2005). Almost the entire JIC pea germplasm (3029 accessions), consisting of a broad balance of cultivars (33%), landraces (19%), wild accessions (13%) and genetic stocks (26%), was analyzed using 45 RBIP markers (Jing *et al.*, 2010); and 1283 pea accessions, representing much of the cultivated pea diversity, held at the Czech National Pea Germplasm collection (CzNPC), were genotyped using a combination of 25 RBIPs and 10 SSRs (Smýkal *et al.*, 2008a and in preparation). The latter study has shown that both SSRs and RBIPs have similarly high information content and offer comparable diversity measurements. This is an important finding, as SSRs are more difficult to transfer between laboratories and suffer from homoplasy.

Data processing – analysis of genetic diversity structure

Altogether a large number of polymorphic data points have been produced and analyzed; however, the extended use of such data is limited, especially in the absence of cross-comparison between collections. Thus, an international initiative was formed to coordinate the international *Pisum* research community (Furman *et al.*, 2006; Smýkal *et al.*, 2008b) in order to allow combining available datasets into a virtual global pea collection and the development of a dispersed international reference pea collection. Such a collection would provide a useful and powerful resource for generation of next generation markers, such as SNPs, or even whole genome sequencing and, more importantly, phenotypic analysis. These would act as toolkits for association mapping and offer a strategy to gain insight into genes and genomic regions underlying desired traits.

Other than conventional linkage mapping based on time-consuming mapping population development, LD mapping, which uses the non-random associations of loci in haplotypes, is a powerful, high-resolution tool for elucidating complex quantitative traits. In contrast to biparental crosses, the higher resolution and the possibility of historical trait data exploitation indicate and provide enormous potential for the LD method in crop breeding and genetics.

Improvements in marker methods have been accompanied by refinements in computational methods to convert raw data into useful representations of diversity and genetic structure. Still, largely used distance-based methods have been challenged by model-based approaches. In particular, Bayesian inference of phylogeny has become popular in the field of population genetics (Pritchard *et al.*, 2000; Rosenberg, 2002; Falush *et al.*, 2003; Corander *et al.*, 2003, 2004, 2005, 2006). This has revolutionized phylogeny estimation by incorporating probability, the provision for measure of support and, especially, complex model and data character processing (Huelsenbeck *et al.*, 2001; Holder and Lewis, 2003; Beaumont and Rannala, 2004; Corander *et al.*, 2004). The high rate of genetic exchange within *Pisum* means that tree-like descriptions of variation patterns can be misleading because different markers produce different tree structures among the same genotype sets. Moreover, the composition of various data types, such as morphology and DNA-based data (Smýkal *et al.*, 2008a), supports the use of alternative approaches, such as principal coordinate or component analyses, multidimensional scaling and, particularly, modelling methods (Pritchard *et al.*, 2000; Corander *et al.*, 2003, 2004, 2006). Although applied largely in population genetics, their usefulness has also been demonstrated in germplasm genetic structure assessment, including in pea (Smýkal *et al.*, 2008a; Jing *et al.*, 2010). Model-based analysis of population structure provides information that cannot be gained from distance-based analysis, which can introduce distortions and simplify relationships between members in complex clusters. Furthermore, these methods were introduced to overcome the constraint of accession partitioning between two distinct clusters, which is common in modern varieties with distant parent crosses. No direct computational comparison between distance- and model-based population structures is possible, since these methods rely upon different principles. Nevertheless, the utility and complementarity of these approaches have been shown (Rosenberg, 2002; Corander *et al.*, 2003, 2004; Smýkal *et al.*, 2008a).

Several types of Bayesian modelling software are currently available. Although they perform similarly in relatively small datasets, there are differences, especially when the level of subpopulation differentiation (F_{ST}) is below 0.1 (Latch *et al.*, 2006), as is common in germplasm collections. To date, published plant germplasm studies have primarily relied upon STRUCTURE software (Pritchard *et al.*, 2000; Falush *et al.*, 2003), which assigns genotypes probabilistically to a user-defined number of clusters or gene pools. Partition-based alternatives provided by BAPS software use analytical integration strategy combined with stochastic search methods and are also appropriate when the number of genetically diverged sources contributing to observed data is unknown (Corander *et al.*, 2003, 2004, 2006, 2007). Additionally, BAPS provides the following advantages over STRUCTURE: (1) analytical integration for the fitting of the models provides a more reliable estimation for complex datasets; (2) spatial models of genetic

Table 1. Description of three pea germplasm collections used in this study: CzNPC, JIC *Pisum* collection and ATFC

Collection	Number of accessions	Composition
JIC	3029	Cultivars (33%), landraces (19%), wild accessions (13%) and genetic stocks (26%)
ATFC	2120	1243 Chinese origin, 774 globally diverse *P. sativum,* 103 of wild *Pisum* sp.
CzNPC	1283	Commercial varieties and breeding lines (75%), landraces (24%) and mutants or wild material (1%)

population structure can be accommodated in BAPS; (3) admixture inference in BAPS enables the investigation of the statistical significance of estimated admixture coefficients; and (4) BAPS requires much lower computation time than STRUCTURE (Latch *et al.*, 2006). We therefore propose BAPS analysis as a suitable approach for future germplasm management.

Towards the world pea core collection

In keeping with the above-mentioned methods, and in order to have a compatible dataset needed for composite pea germplasm analyses, we have chosen easily scorable (essentially binary) retrotransposon insertion (RBIPs) markers to conduct an analysis of three large collections (Table 1). We have used the entire JIC dataset from Jing *et al.* (2010), which consists of 3029 accessions comprised largely of expedition acquisitions and mutant stocks; the 1283 accessions from the CzNPC, consisting of cultivated varieties, landraces and breeding lines (Smýkal *et al.*, 2008a); and a selected core set of 117 accessions of Chinese origin from the ATFC collection (Zhong *et al.*, 2009a) (Table 2). The latter were of particular interest, as an analysis of SSR loci showed the Chinese samples to be genetically distinct from the global gene pool sourced outside of China (Zhong *et al.*, 2009a, b).

As an initial step, we have conducted a Bayesian BAPS analysis of the original datasets for all three of the above-mentioned collections. The ATFC germplasm, comprised of the 1243 accessions of Chinese origin, 774 globally diverse *P. sativum* genotypes and 103 wild pea accessions, was analyzed using 21 SSR loci (42 data points/accession) partitioned into $K = 2–10$ clusters, with optimal clustering being $K = 6$ and 8 (Fig. 4(a)). Cluster 1 contained 97 accessions from the Shanxi, Yunnan, Henan and Inner Mongolia parts of China; cluster 2 contained 420 accessions from the Yunnan, Tibet, Sichuan, Inner Mongolia, Hubei, Qinghai and Shanxi provinces; cluster 3 contained 286 accessions of worldwide distributed varieties and breeding lines; cluster 4 included 282 accessions of mostly wild pea material, such as *P. fulvum* (10), *P. sativum* subsp. *elatius* (6), *P. abyssinicum* (12) and cultivated accessions from Australia, Germany, Nepal and Pakistan, with an additional 62 accessions from China (Sichuan, Tibet, Inner Mongolia, Qinghai). Cluster 5 contained 250 tightly clustered accessions from the Anhui, Gansu, Guizhou, Henan, Inner Mongolia, Shanxi, Sichuan and Qinghai provinces. The 181 accessions comprising cluster 6 originated mainly from the Shanxi province, and the large (400 accessions) cluster 7 possessed has a notable number of Afghan (20), Ethiopian (37) and Australian breeding lines (108). Finally, cluster 8 (204 accessions) contained 170 samples from the Inner Mongolia and Shanxi regions (18). Thus, the BAPS analysis also indicated a range of gene pools unique to China (clusters 1, 2, 5, 6 and 8), enlarging on the diversity revealed by SSR analysis (Zong *et al.*, 2009). BAPS analysis also revealed the positioning of 117 accessions of Chinese origin within the ATFC core set (Fig. 4(b)), which was originally assembled based on a distance-generated dendrogram (Zong *et al.*, 2009). The 117 accessions selected for inclusion in a core germplasm set were placed within all eight clusters identified by BAPS, although their distribution was not even and could be further improved to capture original set diversity using the BAPS data. In contrast to SSR analysis, the RBIP marker data did not identify specific, private alleles; thus, it is allelic frequency that makes retrotransposon insertion data informative. Interestingly, although 115 alleles in total were detected across the 21 microsatellite loci, this did not separate wild pea (especially *P. fulvum* and *P. elatius*) from cultivated germplasm, as found previously using

Table 2. List of material used for composed dataset in this study with indicated levels of missing data (zero scores owing to primer annealing versus accessions; see Jing *et al.,* 2010 for details) and heterogeneity (bulk of 10 or 20 plants per sample used in CzNPC and ATFC datasets) used in the composed dataset study

Germplasm	Number of accessions	Missing data (%)	Heterogeneity (%)
JIC	3029	20	Not assessed
CzNPC	1283	7	10
ATFC Chinese core	117	5	8
Composed set	4429	16	3

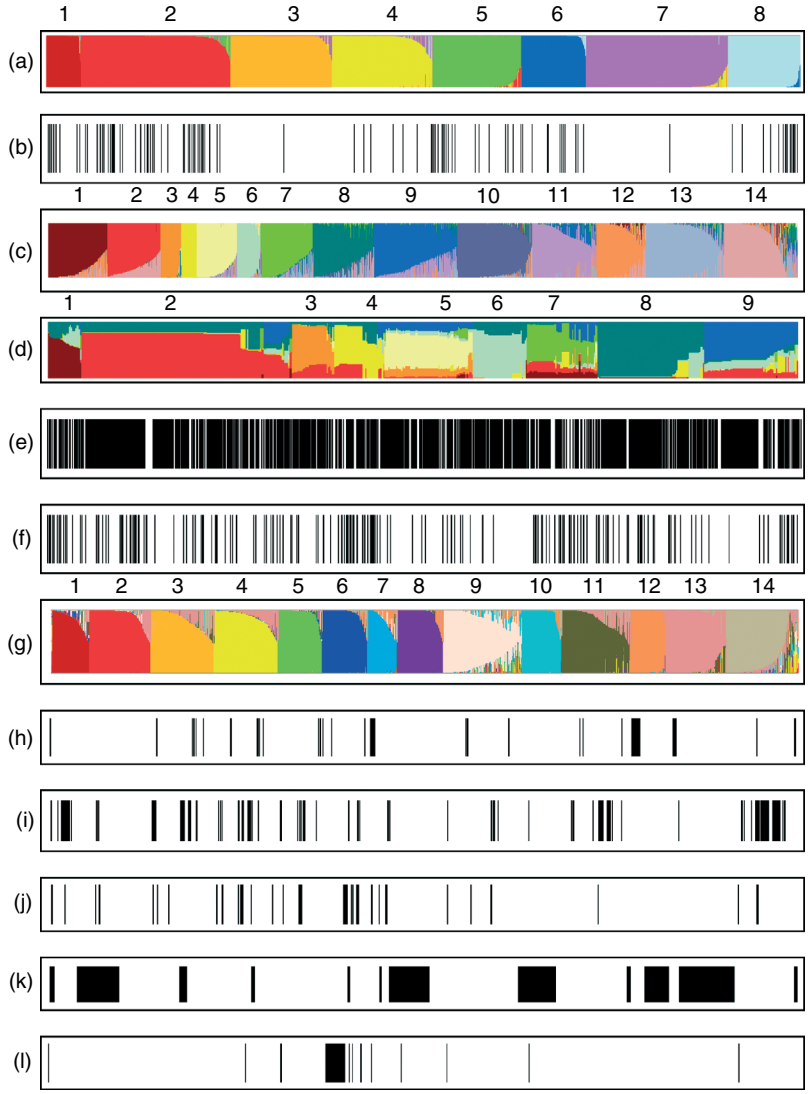

Fig. 4. BAPS analysis partitioning. (a) BAPS at $K = 8$ of 2120 accessions of ATFC collection (Zong *et al.*, 2009) genotyped by 21 SSR loci. (b) Black bars indicate distribution of 117 core set accessions of Chinese origin (according to Zong *et al.*, 2009) used for composed dataset analysis. (c) BAPS at $K = 14$ of 3029 accessions of JIC collection (Analysis and exploitation of germplasm resources using transposable element molecular markers dataset, Jing *et al.*, 2010) genotyped by 45 RBIP loci. (d) BAPS at $K = 9$ of 1283 accessions of CzNPC analyzed by combination of 25 RBIP and 10 SSR loci. (e) Dry-seed pea (*P. sativum* subsp. *sativum* var. *sativum*) accessions are indicated as black bars, while fodder pea (var. *arvense*) accessions are shown in white. (f) 203 accessions of Czech/Slovak origin (from Smýkal *et al.*, 2008a) are shown as black bars. (g) 4429 accessions of the combined set analyzed by 17 selected RBIP loci. (h) 140 accessions of Chinese origin (*P. sativum* cultigen). (i) 349 accessions of Ethiopian origin (*P. sativum* cultigen). (j) 100 accessions of Afghan origin (*P. sativum* cultigen). (k) 1283 accessions from the Czech collection (*P. sativum* cultigen). (l) 140 accessions of wild forms (*P. fulvum, P. sativum* subsp. *elatius* and *P. abyssinicum*).

retrotransposon-based RBIP assay of the JIC germplasm (Jing *et al.*, 2010). Also in the study by Nasiri *et al.* (2009), 20 SSR loci clearly discriminated wild *Pisum* sp. accessions, including, *P. sativum* subsp. *elatius* and *P. sativum* subsp. *abyssinicum*. Moreover, in contrast to the distance-based analysis and principal component analysis applied by Zong *et al.* (2009), which identified 214 clusters, the model-based BAPS analysis clearly showed clustering in eight well-supported clusters. In addition, Zong *et al.* (2009) showed that a microsatellite-based 146 core germplasm set captured better allele diversity within the original collection than a core constructed solely based on geographic origin.

The 3029 accessions of the JIC collection (http://www.jic.ac.uk/germplas/pisum/) analyzed by 45 RBIP loci (45 data points/accession) were assigned into $K = 2$ to 14 clusters (Fig. 4(c)). This is in contrast to

STRUCTURE analysis, which partitioned the collection into maximally $K = 7$ clusters (Jing *et al.*, 2010). In a direct comparison of the BAPS and STRUCTURE methods, a good level of agreement was found when $K = 3$ (Fig. 5(a)); however, the former method was favoured since the resultant substructuring revealed biological meaningful diversity. It has to be noted that this comparison also takes into account the order of each accession. As in the case where the BAPS posterior cluster assignment often equals 1.0, as indicated by colour bars, one can see some block-like cluster correspondence, rather than diagonal, which would be the case of a complete match. This is well preserved in cluster 2, containing the majority of the wild material using both BAPS and STRUCTURE. On the other hand, comparison of cluster assignments for higher K values, such as $K = 7$, did not shown any significant correspondence (Fig. 5(b)). In the analysis of Jing *et al.* (2010), each of the three $K = 3$ sets was separately subjected to further STRUCTURE analysis. Groups 1 and 3 were further subclustered into $K = 6$, while group 2 was subclustered into $K = 2$. Group 1 was dominated by *P. sativum* landraces and cultivars, largely round- and large-seed phenotypes; group 2 contained *P. sativum* cultivars with primarily wrinkled-seed types. In contrast, group 3 showed a considerable amount of substructuring with regard to both taxonomy and phenotypic traits (Jing *et al.*, 2010). Subgroups separated almost all *P. abyssinicum, P. elatius* and *P. fulvum*, along with accessions of Afghan origin. In contrast, the BAPS analysis at $K = 14$ directly identified separate clusters containing wild material, distinguishing *P. fulvum* (cluster 3 in Fig. 4(b)) from *P. elatius* and *P. abyssinicum* (cluster 4 in Fig. 4(b)), supporting the view of separate species or subspecies. This wild material was clearly already separated at a $K = 5$ value, while at $K = 11$, *P. fulvum* was clustered from *P. elatius* and *P. abyssinicum*. In addition to these, accessions of domesticated (*P. sativum* subsp. *sativum*) peas of Afghan and Ethiopian origin (clusters 6 and 7, 8, respectively, Fig. 5(i,j)) were readily separated by BAPS. The remaining nine clusters at $K = 14$ contained well-structured, cultivated material lacking much geographical or user type stratification. However, *P. sativum* of Ethiopian origin constituted a large part of the JIC germplasm, and these accessions proved well resolved starting from $K = 11$, and along with the Afghan accessions, separated at $K = 14$ (cluster 7). There was a significant (306) proportion of modern varieties partitioned in cluster 10. Outgrouping of Afghan types was supported by previous studies, which classified them as resistant to European *Rhizobium* strains (Young and Matthews, 1982). Moreover, positioning of *P. abyssinicum* together with *P. elatius* and *P. fulvum* is in agreement with phylogenetic analysis using chloroplastDNA and ITS markers, supporting the view that *P. abyssinicum* is an ancient hybrid of the two species.

Finally, 1283 accessions from the Czech National Pea Collection (CzNPC, http://genbank.vurv.cz/genetic/resources) analyzed by a combination of 25 RBIP and 10 SSR loci (57 data points/accession) were assigned into $K = 9$ (with the highest posterior probability of 0.9997; Fig. 4(d)). This germplasm contained largely commercial varieties and breeding lines (75%), while the remainder were landraces (24%) and mutants or wild material (1%) (Smýkal *et al.*, 2008a). The prevalence of highly bred material likely explains consistently lower posterior assignment values (≤ 1) in most of the accessions, in comparison to those in the ATFC and JIC collections. Furthermore, part of this maybe attributed to the use of a combination of RBIP and SSR data, as shown by separate marker analysis (data not shown). CzNPC was divided by morphological descriptors into dry-seed pea *P. sativum* subsp. *sativum* var. *sativum* (L01 accessions, 1006) (Fig. 4(e)) and fodder pea *P. sativum* subsp. *sativum* var. *arvense* (L02 accessions,

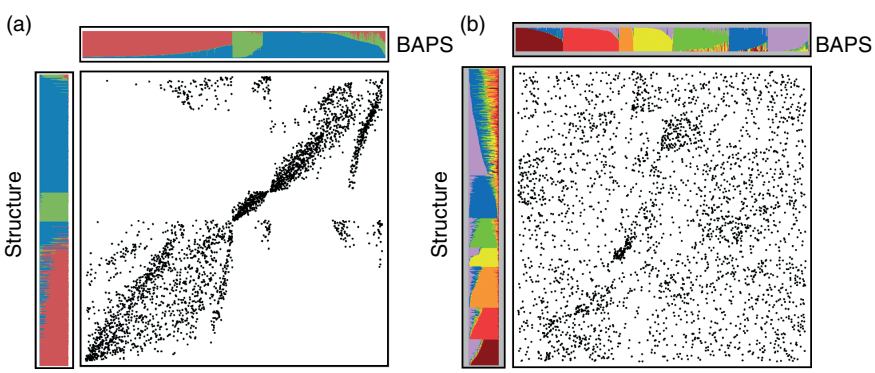

Fig. 5. Comparison of 3029 accessions from the JIC dataset genotyped by 45 RBIP loci assigned by BAPS or STRUCTURE software at $K = 3$ (a) or K = 7 values (b), respectively. Respective clusters are colour-coded and ordered according to accessions assignment.

277). As shown previously (Smýkal *et al.*, 2008a), both RBIP and SSR markers do not discriminate between these classes.

Overall, large amounts of diversity were captured among cultivated material, which could be clearly visualized by model-based methods. Although no clear geographical assignments were found in most of the older varieties, likely owing to a large degree of interbreeding, landraces were found in clusters distinct from the modern varieties (Smýkal *et al.*, 2008a and in preparation).

In the combined set of the three germplasm collections, a total of 4429 accessions were analyzed by 17 selected RBIP loci, providing a total of 75,293 data points with zero scores treated as missing data for 16% of the dataset (see Jing *et al.*, 2005, 2010 for explanation of zero scores) (Table 1). Subsequent BAPS analysis provided posterior assignments for $K = 2$ to 14, with optimal partitioning into $K = 11$ (Fig. 4(g)). Although 17 RBIP loci might be considered a low number to sample the diversity of the large pea genome, clear genetic structure could be observed. Notably, all wild pea (*P. fulvum*, 53; *P. sativum* subsp. *elatius*, 28; and *P. abyssinicum,* 26) were placed in cluster 6 at $K = 14$ (Fig. 4(l)), together with the accessions of Afghan origin (27) (Fig. 4(j)). Furthermore, cluster 14 contained a large proportion of *P. sativum* subsp. *sativum* (140 accessions of Ethiopian origin.). Also, 117 accessions from the ATFC plus 23 JIC core of Chinese origin were distributed into clusters 8, 11, 12 and 14 (Fig. 4(h)). The remaining clusters contained all cultivated material (Fig. 4(k)) plus the JIC set of mutant lines. It was proposed that the distinct differentiation of the Chinese *P. sativum* genotypes may in part reflect the early isolation of agriculture in eastern Asia from that in southern Asia, Europe and northern Africa (Zong *et al.*, 2009) and the restricted initial gene pool and opportunities for recombination outside this relatively closed gene pool.

Furthermore, multivariate analysis (see Smýkal *et al.*, 2008a and Jing *et al.*, 2010 for methods) revealed relatively closer genetic distance within cultivated material, especially of modern varieties and breeding lines, while wild material provides much of the *Pisum* genus diversity (Fig. 6). The greater genetic distance of wild forms and some of the material of Chinese origin suggests usefulness of this material for further breeding.

Although the above-mentioned marker types are now widespread, their potential is limited due to the small amount of the genome that is actually assessed. With advances in model legume sequencing, increased genomic knowledge and rapidly progressing next generation sequencing technologies, there is a progression towards gene-based markers such as high-throughput SNP generation and detection assays. Recently, the first highly multiplexed SNP genotyping assay was published for pea (Deulvot *et al.*, 2010).

Data deposition and core collections

One very important, if not critical, issue is the deposition and availability of data. So far, data held at the national level have not been broadly accessible. Although the European EURISCO Web catalogue (http://eurisco.ecpgr.org), maintained by Bioversity International and the USDA National Plant Germplasm System (GRIN), provides information on around two million accessions, this information is largely passport-based and is thus limited. From GRIN, pea descriptor data (153,812 observations) and digital images (10,643) are downloadable at http://www.ars-grin.gov/cgi-bin/npgs/html/crop.pl?177. Fortunately, the recent EU-funded PGR Secure project, on *Avena*, *Brassica*, *Beta* and *Medicago* case studies, should lead to a database system that will bring together passport, morphological and genotypic data (Lee *et al.*, 2005) that will both improve germplasm management and enable data exploration across a wide range of data types.

Defining a pea core together with a set of markers provides a basis for the comparison of phenotypic and molecular analyses and would form a useful additional case study for the PGR Secure project. No standardized method for core collection (Hodgkin *et al.*, 1995) assembly has been established, although numerous strategies have been proposed and tested (Van Hintum, 1999; Hu *et al.*, 2000; Wang *et al.*, 2007; Thachuk *et al.*, 2009). Further methods continue to be developed as new approaches and algorithms become available. The most commonly used grouping strategy relies on geographical (e.g. passport) data, followed by morphological characteristics (Brown and Spillane, 1999). Since most traits are quantitative and influenced by many genes, they are affected by environmental and experimental conditions. Consequently, stratification based on phenotypic traits would not accurately reflect genetic relationships. Pairwise genetic distance calculation followed by the subtraction of the most commonly related accessions is a widely adopted method. However, as shown earlier, genetic distance does not properly reflect population structure as Bayesian inference. The application of a model-based method for pea core collection establishment was successfully tested on a subset of the Czech National Pea Collection (Smýkal *et al.*, 2008a) and is currently being further developed. Such collections will be valuable for producing an integrated framework of genetic and phenotypic data generated by different studies.

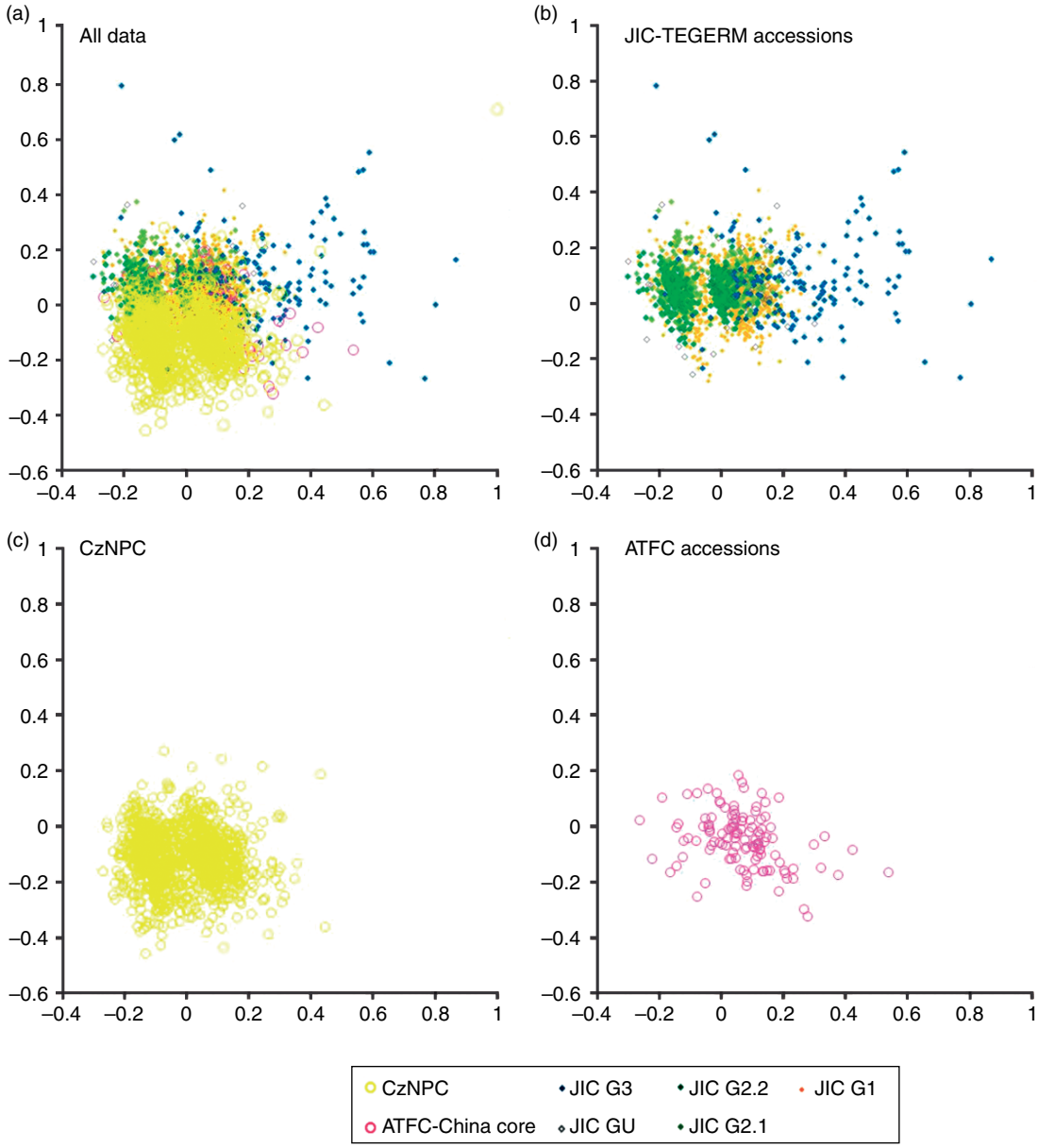

Fig. 6. Multivariate analysis of a composed dataset. For the entire dataset, the fraction of shared alleles for all pairwise combinations of samples was analyzed by multidimensional scaling. The output for the first two dimensions is shown. All points are plotted, and each sample is colour-coded according to germplasm assignment: (a) composed dataset, (b) JIC-TEGERM dataset (as in Jing *et al.*, 2010), (c) CzNP collection, L accessions, (d) ATFC core collection of Chinese origin.

Germplasm collections are dynamic

The maintenance of germplasm genetic integrity is essential for long-term *ex situ* conservation. Periodic regeneration, performed on limited plots with a small number of individuals, increases the risk of genetic drift, which in turn leads to a decrease or even loss of genetic diversity (Breese, 1989). Modern techniques of seed storage can maintain seed viability for over 100 years. In such conditions, base collections are stored. This category of collection is generally used exclusively for the regeneration and maintenance of the stocks in active collections, where the emphasis is on characterization, evaluation and distribution. Only a few published studies were devoted to germplasm integrity evaluations. We have shown that over a 20–40 year period, with about four to ten regeneration cycles performed to maintain seed viability, the genetic diversity contained within pea germplasm accessions was reduced or even lost (Cieslarová *et al.*, 2010). These findings imply that regeneration procedures should be improved to accommodate more numerous samples and that the

composition of the collection should be continuously monitored to prevent the risk of genetic diversity loss.

General conclusions

This study, based on molecular data, has positioned *Pisum* between *Vicia* and *Lathyrus* and shown it to be closely allied to *Vavilovia*. Study of phylogeography supports the spread of wild pea from centre of origin (Middle East) eastwards (to the Caucasus, Iran and Afghanistan) and westwards to the Mediterranean region. Analysis of wide pea germplasm has demonstrated that *Pisum* is a diverse genus. Bayesian analysis of a combined dataset of 4429 pea accessions, using locus-specific retrotransposon insertion markers, has separated wild species and subspecies (*P. fulvum*, *P. sativum* subsp. *elatius* and *P. abyssinicum)* from cultivated material. Within cultivated pea (*P. sativum* subsp. *sativum*), accessions from Afghanistan, Ethiopia and China were distinguished. These results showed that comparably large diversity is captured among cultivated material, which could be clearly visualized by model-based methods. We have demonstrated the superiority of BAPS over STRUCTURE software and propose BAPS analysis as a suitable approach for germplasm exploration and management. Despite multiple introgression between cultivated and wild *Pisum*, significant genetic variation is present in wild *Pisum*. However, plant breeders are reluctant to use wild germplasm because hybrids with wild material have a high likelihood of having impaired rather than improved performance. This underlines the necessity for increased pre-breeding efforts, whereby the traits of interest, such as biotic and abiotic resistance, are made available in backgrounds more acceptable to breeders.

Acknowledgements

P. S. acknowledges financial support from the Ministry of Education of Czech Republic, MSM 2678424601, LA08011 and the Bioversity International AEGIS LOA 10/048 projects. O. K. acknowledges financial support from Russian Fund for Fundamental Research, grant 10-04-00 230-a. M. J. A. acknowledges financial support from Defra GC0142 project for the maintenance of the JIC Pisum collection.

References

Ambrose MJ (1995) From Near East centre of origin the prized pea migrates throughout world. *Diversity* 11: 118–119.

Ascheron P and Graebner P (1910) Synopsis der mitteleuropaischen Flora Bd 6, Abt 2, IV Leipzig 1093 S.

Aubert G, Morin J, Jacquin F, Loridon K, Quillet MC, Petit A, Rameau C, Lejeune-He'naut I, Huguet T and Burstin J (2009) Functional mapping in pea, as an aid to the candidate gene selection and for investigating synteny with the model legume. *Medicago truncatula*. *Theoretical and Applied Genetics* 112: 1024–1041.

Baranger AG, Aubert G, Arnau G, Lainé AL, Deniot G, Potier J, Weinachter C, Lejeune-Hénaut J, Lallemand J and Burstin J (2004) Genetic diversity within *Pisum sativum* using protein- and PCR-based markers. *Theoretical and Applied Genetics* 108: 1309–1321.

Beaumont MA and Rannala B (2004) The Bayesian revolution in genetics. *Nature Reviews in Genetics* 5: 251–261.

Ben-Ze'ev N and Zohary D (1973) Species relationship in the genus *Pisum* L. *Israel Journal of Botany* 22: 73–91.

Berger A (1928) Systematic botany of peas and their allies. In: Hedrick U (ed.) *The Vegetables of New York*. 1. Albany: State of New York, Education Department, pp. 10–18.

Bhargava A and Fuentes FF (2010) Mutational dynamics of microsatellites. *Molecular Biotechnology* 44: 250–266.

Bieberstein M (1808) Flora taurico-caucasica exhibens stirpes phaenomagas, in Chersoneso Taurica et regionibus caucasicis sponte crescentes. Bd 2 Charkouiae, Typ Akad 477 S.

Bogdanova VS (2007) Inheritance of organelle DNA markers in a pea cross associated with nuclear-cytoplasmic incompatibility. *Theoretical and Applied Genetics* 114: 333–339.

Bogdanova VS, Galieva ER and Kosterin OE (2009) Genetic analysis of nuclear-cytoplasmic incompatibility in pea associated with cytoplasm of an accession of wild subspecies *Pisum sativum* subsp. elatius (Bieb.) Schmahl. *Theoretical and Applied Genetics* 118: 801–809.

Boissier E (1856) Diagnoses plantarum orientalum novarum. *Lipsie* 3: 125.

Breese EL (1989) *Regeneration and Multiplication of Germplasm Resources in Seed Genebanks: The Scientific Background*. Rome: International Board for Plant Genetic Resources.

Brown AHD and Spillane C (1999) Implementing core collections-principles, procedures, progress, problems and promise. In: Johnson RC and Hodgkin T (eds) *Core Collections for Today and Tomorrow*. Rome, Italy: International Plant Genetic Resources Institute, pp. 1–9.

Burstin J, Deniot G, Potier J, Weinachter C, Aubert G and Baranger A (2001) Microsatellite polymorphism in *Pisum sativum*. *Plant Breeding* 120: 311–317.

Cieslarová J, Smýkal P, Dočkalová Z, Hanáček P, Procházka S, Hýbl M and Griga M (2010) Molecular evidence of genetic diversity changes in pea (*Pisum sativum* L.) germplasm after long-term maintenance. *Genetic Resources and Crop Evolution*. doi 10.1007/s10722-010-9591-3.

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

Corander J, Waldmann P and Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics* 164: 367–374.

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

Corander J, Gyllenberg M and Koski T (2007) Random partition models and exchangeability for Bayesian identification of

population structure. *The Bulletin of Mathematical Biology* 69: 797–815.

Coyne CJ, Brown A, Timmerman-Vaughan GM, McPhee KE and Grusak MA (2005) Refined USDA-ARS pea core collection based on 26 quantitative traits. *Pisum Genetics* 37: 3–6.

Davis PH (1970) *Pisum* L. In: Davis PH (ed.) *Flora of Turkey and East Aegean Islands*. vol. 3. Edinburg: Edinburg University Press, pp. 370–373.

Deulvot C, Charrel H, Marty A, Jacquin F, Donnadieu C, Lejeune-Hénaut I, Burstin J and Aubert G (2010) Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. *BMC Genomics* 11: 468.

Doyle JJ, Doyle JL, Ballenger JA, Dickson EE, Kajita T and Ohashi (1997) A phylogeny of the chloroplast gene *rbcL* in the Leguminosae: taxonomic correlations and insights into the evolution of nodulation. *American Journal of Botany* 84: 541–554.

Ellis THN and Poyser SJ (2002) An integrated and comparative view of pea genetic and cytogenetic maps. *New Phytologist* 153: 17–25.

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

Endo Y, Choi BH, Ohashi H and Delgado-Salinas A (2008) Phylogenetic relationships of New World Vicia (Leguminosae) inferred from nrDNA internal transcribed spacer sequences and floral characters. *Systematic Botany* 33: 356–363.

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

Ford R, Le Roux K, Itman C, Brouwer JB and Taylor PWJ (2002) Genome-specific sequence tagged microsatellite site (STMS) markers for diversity analysis and genotyping in *Pisum* species. *Euphytica* 124: 397–405.

Furman BJ, Ambrose M, Coyne CJ and Redden B (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.

Govorov LI (1937) Goroch (Peas). *Kulturnaja Flora SSR* (in Russian). Moscow: State Printing Office, pp. 229–336.

Hodgkin T, Brown AHD and van Hintum TJL and Morales EAV (eds) (1995) *Core Collections of Plant Genetic Resources*. Chichester: John Wiley & Sons, pp. 95–107.

Hoey BK, Crowe KR, Jones VM and Polans NO (1996) A phylogenetic analysis of *Pisum* based on morphological characters, allozyme and RAPD markers. *Theoretical and Applied Genetics* 92: 92–100.

Hu J, Zhu J and Xu HM (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic valued of crops. *Theoretical and Applied Genetics* 101: 264–268.

Huelsenbeck JP, Ronquist F, Nielsen R and Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294: 2310–2314.

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

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: 2263–2275.

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.

Kalendar R and Schulman AH (2006) IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. *Nature Protocols* 1: 2478–2484.

Kass E and Wink M (1996) Molecular evolution of the Leguminosae: phylogeny of the three subfamilies based on *rbcL*-sequences. *Biochemical Systematics and Ecology* 24: 365–378.

Kenicer GJ (2007) Systematics and biogeography of *Lathyrus* L. (*Leguminosae, papilionoideae*). PhD Thesis, Royal Botanical Garden Edinburgh.

Kenicer G, Smýkal P and Mikič A (2008) Phylogenetic study of mysterious *Vavilovia formosa* (Stev.) Fed., a *Pisum* relative. IV. International Conference on Legume Genetics and Genomics, 2008 Puerto Vallarta, Mexico, pp. 83.

Kenicer GJ, Kajita T, Pennington RT and Murata J (2005) Systematics and biogeography of *Lathyrus* (*Leguminosae*) based on internal transcribed spacer and cpDNA sequence data. *American Journal of Botany* 92: 1199–1209.

Knight TA (1799) Experiments on the Fecundation of Vegetables. *Philosophical Transactions of the Royal Society* 89: 504–509.

Kosterin OE and Bogdanova VS (2008) Relationship of wild and cultivated forms of *Pisum* L. as inferred from an analysis of three markers, of the plastid, mitochondrial and nuclear genomes. *Genetic Resources and Crop Evolution* 55: 735–755.

Kosterin OE, Zaytseva OO, Bogdanova VS and Ambrose M (2010) New data on three molecular markers from different cellular genomes in Mediterranean accessions reveal new insights into phylogeography of *Pisum sativum* L. subsp *elatius* (Bieb.) Schmalh. *Genetic Resources and Crop Evolution* 57: 733–739.

Kupicha FK (1981) *Vicieae* (Adans.) DC. (1825) nom conserv prop. In: Polhill RM and Raven PH (eds) *Advances in Legume Systematics 1*. Kew: Royal Botanical Gardens, pp. 377–381.

Lamprecht H (1963) Zur Kenntnis von *Pisum arvense* L. *oect. abyssinicum* Braun, mit genetischen und zytologischen Ergebnissen. *Agric Hort Genetics* 21: 35–55.

Lamprecht H (1966) *Die Enstehung der Arten und hoheren Kaategorien*. Wien: Springer Verlag.

Latch EK, Dharmarajan G, Glaubitz JC and Rhodes OE (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics* 7: 295–302.

Lavin M, Herendeen PS and Wojciechowski M (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Systematic Biology* 54: 575–594.

Lee D, Ellis THN, Turner L, Hellens RP and Cleary WG (1990) A copia-like element in *Pisum* demonstrates the uses of dispersed repeated sequences in genetic-analysis. *Plant Molecular Biology* 15: 707–722.

Lee MJ, Davenport GF, Marshall D, Ellis THN, Ambrose MJ, Dicks J, van Hintum TJL and Flavell AJ (2005) GERMINATE.

A generic database for integrating genotypic and phenotypic information for plant genetic resource collections. *Plant Physiology* 139: 619–631.

Lehmann C (1954) Das morphologische system der Saaterbsen. *Der Zuchter* 24: 316–337.

Lewis G, Schrirer B, Mackinder B and Lock M (eds) (2005) *Legumes of the World*. Kew: Royal Botanical Gardens.

Lock M and Maxted N (2005) Tribe *Fabeae*. In: Lewis G, Schrire B, Mackinder B and Lock M (eds) *Legumes of the World*. Richmond: Royal Botanic Gardens, Kew.

Macas J, Neumann P and Navrátilová A (2007) Repetitive DNA in the pea *(Pisum sativum* L. genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics* 8: 427.

Makasheva RK (1979) Gorokh (pea). In: Korovina ON (ed.) Leningrad: Kulturnaya Flora SSR, Kolos, pp. 1–324 (in Russian).

Marx GA (1977) Classification, genetics and breeding. In: Sutcliffe JF and Pate JS (eds) *Physiology of the Garden Pea*. New York: Academic Press, pp. 21–43.

Maxted N and Ambrose N (2000) Peas (*Pisum* L.) Chapter 10. In: Maxted N and Bennett SJ (eds) *Plant Genetic Resources of Legumes in the Mediterranean*. Dordrecht: Kluwer Academic Publishers, pp. 181–190.

Maxted N, Hargreaves S, Kell SP, Amri A, Street K, Shehadeh A, Piggin J and Konopka J (2010) Temperate forage and pulse legume genetic gap analysis. Paper given at XIII OPTIMA Meeting in Antalya, Turkey, 22–26 March 2010.

Mendel JG (1866) Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereins in Brünn* 4: 3–47. (see also http://www.mendelweb.org/)

Miller P (1768) *The Gardener's Dictionary; Containing the Methods of Cultivating and Improving the Kitchen, Fruit and Flower Garden*. *etc*. printed by J. and J. Rivington, Reprint 1969, Verlag von J. Cramer, Germany, 8th ed. London.

Nasiri J, Haghnazari A and Saba J (2009) Genetic diversity among varieties and wild species accessions of pea (*Pisum sativum* L.) based on SSR markers. *African Journal of Biotechnology* 15: 3405–3417.

Ochatt SJ, Benabdelmouna A, Marget P, Aubert G, Moussy F, Pontecaille C and Jacas L (2004) Overcoming hybridization barriers between pea and some of its wild relatives. *Euphytica* 137: 353–359.

Palmer JD, Jorgensen RA and Thompson WF (1985) Chloroplast DNA variation and evolution in *Pisum*: patterns of change and phylogenetic analysis. *Genetics* 109: 195–213.

Pearce SR, Knox M, Ellis TH, Flavell AJ and Kumar A (2000) Pea Ty1-copia group retrotransposons: transpositional activity and use as markers to study genetic diversity in *Pisum*. *Molecular and General Genetics* 263: 898–907.

Polans NO and Saar DE (2002) ITS sequence variation in wild species and cultivars of pea. *Pisum Genetics* 34: 9–13.

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

Raquin AL, Depaulis F, Lambert A, Galic N, Brabant P and Goldringer I (2008) Experimental estimation of mutation rates in a wheat population with a gene genealogy approach. *Genetics* 179: 2195–2211.

Rosenberg NA (2002) Genetic structure of human populations. *Science* 298: 2381–2002.

Saar DE and Polans NO (2000) ITS sequence variation in selected taxa of *Pisum*. *Pisum Genetics* 109: 195–213.

Sáenz de Miera LE, Ramos J and Pérez de la Vega M (2008) A comparative study of convicilin storage protein gene sequences in species of the tribe *Vicieae*. *Genome* 7: 511–523.

Schmalhausen I (1895) Flora Srednei y Yuzhnoj Rossii, Kryma i Severnogo Kavkaza. 1: 468 (in Russian).

Smýkal P (2006) Development of an efficient retrotransposon-based fingerprinting method for rapid pea variety identification. *Journal of Applied Genetics* 47: 221–230.

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

Smýkal P, Coyne CJ, Ford R, Redden R, Flavell AJ, Hýbl M, Warkentin T, Burstin J, Duc G, Ambrose M and Ellis THN (2008*b*) Effort towards a world pea (*Pisum sativum* L.) germplasm core collection: the case for common markers and data compatibility. *Pisum Genetics* 40: 11–14.

Smýkal P, Kenicer G and Mikič A (2009*a*) Beautiful Vavilovia (*Vavilovia formosa*) and molecular taxonomy of tribe *Fabeae*. *Book of Abstracts IV Congress of the Serbian Genet Society*, p. 166.

Smýkal P, Kalendar R, Ford R, Macas J and Griga M (2009*b*) Evolutionary conserved lineage of *Angela*-like retrotransposons as a genome-wide microsatellite repeat dispersal agent. *Heredity* 103: 157–167.

Steele KP and Wojciechowski MF (2003) Phylogenetic analyses of tribes *Trifolieae* and *Vicieae*, based on sequences of the plastid gene *matK* (*Papilionoideae: Leguminosae*). In: Klitgaard BB and Bruneau (eds) *Advances in Legume Systematics*. Kew: Royal Botanical Garden, pp. 355–370.

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

Thachuk 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.

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

Vershinin AV, Allnutt TR, Knox MR, Ambrose MJ and Ellis NTH (2003) Transposable elements reveal the impact of introgression, rather than transposition, in *Pisum* diversity, evolution, and domestication. *Molecular Biology and Evolution* 20: 2067–2075.

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JSC and Doebley J (2002) Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution* 19: 1251–1260.

Wang JC, Hu J, Xu HM and Zhang S (2007) A strategy on constructing core collections by least distance stepwise sampling. *Theoretical and Applied Genetics* 115: 1–8.

Westphal E (1974) Pulses in Ethiopia, their taxonomy and agricultural significance. *Versl Landbouwkundl Onderzoek*. The Netherlands: Wageningen.

Wojciechowski MF, Lavin M and Sanderson MJ (2004) A phylogeny of legumes (*Leguminosae*) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany* 91: 1846–1862.

Young JPW and Matthews P (1982) A distinct class of peas (*Pisum sativum* L.) from Afghanistan that show strain specificity for symbiotic Rhizobium. *Heredity* 48: 203–210.

Zohary D and Hopf M (2000) *Domestication of Plants in the Old World*. Oxford: Oxford University Press.

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.