Theor Appl Genet (2011) 122:1385–1397
DOI 10.1007/s00122-011-1539-2

ORIGINAL PAPER

Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotransposon-based markers

P. Smýkal · N. Bačová-Kerteszová · R. Kalendar · J. Corander · A. H. Schulman · M. Pavelek

Received: 16 March 2010 / Accepted: 14 January 2011 / Published online: 4 February 2011

Abstract Retrotransposon segments were characterized and inter-retrotransposon amplified polymorphism (IRAP) markers developed for cultivated flax (*Linum usitatissimum* L.) and the *Linum* genus. Over 75 distinct long terminal repeat retrotransposon segments were cloned, the first set for *Linum*, and specific primers designed for them. IRAP was then used to evaluate genetic diversity among 708 accessions of cultivated flax comprising 143 landraces, 387 varieties, and 178 breeding lines. These included both traditional and modern, oil (86), fiber (351), and combined-use (271) accessions, originating from 36 countries, and 10 wild *Linum* species. The set of 10 most polymorphic primers yielded 141 reproducible informative data points per accession, with 52% polymorphism and a 0.34 Shannon diversity index. The maximal genetic diversity was detected among wild *Linum* species (100% IRAP polymorphism and 0.57 Jaccard similarity), while diversity within cultivated germplasm decreased from landraces (58%, 0.63) to breeding lines (48%, 0.85) and cultivars (50%, 0.81). Application of Bayesian methods for clustering resulted in the robust identification of 20 clusters of accessions, which were unstratified according to origin or user type. This indicates an overlap in genetic diversity despite disruptive selection for fiber versus oil types. Nevertheless, eight clusters contained high proportions (70–100%) of commercial cultivars, whereas two clusters were rich (60%) in landraces. These findings provide a basis for better flax germplasm management, core collection establishment, and exploration of diversity in breeding, as well as for exploration of the role of retrotransposons in flax genome dynamics.

Communicated by A. Graner.

Gene Bank accessions DQ767972, GU735096–GU735098, GU929874–GU929878, GU980587–GU980590

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1539-2) contains supplementary material, which is available to authorized users.

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

N. Bačová-Kerteszová
Slovak University of Agriculture, Tr. A. Hlinku 2, 949 76 Nitra, Slovak Republic

R. Kalendar · A. H. Schulman
MTT/BI Plant Genomics Lab, Institute of Biotechnology, University of Helsinki, Viikinkaari 1, Biocenter 3, Helsinki 00801, Finland

J. Corander
Department of Mathematics, Abo Akademi University, Biskopsgatan 8, 20500 Åbo, Finland

J. Corander
Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, 0014 Helsinki, Finland

A. H. Schulman
Biotechnology and Food Research, MTT Agrifood Research Finland, Jokioinen, Finland

M. Pavelek
Department of Grain Legumes and Technical Crops, Agritec Plant Research Ltd, Zemědělská 2520/16, 787 01 Šumperk, Czech Republic

Introduction

Flax (*Linum usitatissimum* L.) belongs among the oldest domesticated plants; it was already cultivated in ancient Egypt and Samaria 10,000 years ago (Zohary and Hopf 2000) to provide both fiber and oil. Recently, 30,000-year-old processed and colored flax fiber was found, indicating that early humans made fabric or threads from the flax (Kvavadze et al. 2009). The fibers are used in textile as well as automobile and construction industries, while flax seeds are used in animal feed and human food, as source of omega-3-fatty acids and biologically active lignans. Linseed oil is used in the production of paint, soap, putty, and polymers. The genus *Linum* is the type genus for the flax family, *Linaceae* DC. (Dumort) comprising 22 genera and about 300 species distributed worldwide (Hickey 1988; McDill et al. 2009). *Linum* is the largest genus (about 200 species) within the family and is found both in the Mediterranean region and the Americas. It includes both horticultural plants with various flower colors and one field crop (*Linum usitatissimum* L.). The center of origin of cultivated flax is believed to be the Middle East, although secondary diversity centers were identified in the Mediterranean basin, Ethiopia, Central Asia, and India (Vavilov 1926; Zohary and Hopf 2000). Molecular analysis of the steroyl-ACP-desaturase II (*sad2*) locus, involved in unsaturated fatty acid synthesis, supports a single domestication origin for all extant cultivated flax (Allaby et al. 2005), in spite of its wide geographical range. Moreover, this analysis supported the antiquity of oil flax over fiber use. On the other hand, from ancient time till twentieth century, flax was the important source of fiber for textile industry in temperate regions, rather than cultivated for oil use. Flax is not found as a wild plant; *L. angustifolium* (*L. bienne*), with which is shares the same chromosome number ($2n = 30$) and is interfertile, is considered its progenitor (Diederichsen and Hammer 1995; Fu et al. 2002a; Gill and Yermanos 1967; Uysal et al. 2010).

As its species name, given by Linné, indicates, flax has long been considered a very useful plant, not only for fiber and oil production but also in traditional medicine. The two major user types are connected to morphotypes, broadly designed as oil (linseed; convar. *mediterraneum*), fiber (flax; convar. *elongatum*), and intermediate (convar. *usitatissimum*) varieties, although this infra-specific grouping is not unified (Diederichsen and Fu 2006). Wider fiber use may be reflected in the historic uses of landraces (Kvavadze et al. 2009), with both fiber and dual uses relatively more important than at present. Flax cultivation has been steadily declining in favor of cotton, with linseed currently occupying about 2.4 million ha and fiber flax is about 0.4 million ha (FAOSTAT data 2010). Nevertheless, the plant is sparking new interest for its sustainability and for novel pharmaceutical and nutritional uses (Haggans et al. 1999; Hilakivi-Clarke et al. 1999). World genebanks maintain about 48,000 accessions of cultivated flax, of which some 10,000 might be unique (Diederichsen 2007).

Both to identify potentially novel genotypes among the many flax accessions, and to assess genetic diversity for both germplasm management and core collection (Frankel and Brown 1984) assembly, molecular markers are highly useful. A variety of marker systems, including random amplified polymorphic DNA (RAPD), inter-simple sequence repeat (ISSR), amplified fragment length polymorphism (AFLP), and simple sequence repeat (SSR), been used to analyze flax germplasm (Cloutier et al. 2009; Diederichsen and Fu 2006; Everaert et al. 2001; Fu 2002, 2005; Fu et al. 2002a, b, 2003; van Treuren et al. 2001; Wiesnerova and Wiesner 2004). Taken together, these studies show that cultivated flax has low genetic diversity compare to wild relatives or some other crops (Smýkal et al. 2008a), possibly resulting from a domestication bottleneck.

In apparent contradiction to the lack of diversity indicated by marker studies, the flax genome (686 Mbp) shows environmentally induced yet heritable genomic changes, a phenomenon of interest for many years (Evans et al. 1966). The activation of transposable elements has been proposed and identified as the mechanism behind this genotypic plasticity (Chen et al. 2005, 2009). Transposable elements, particularly the retrotransposons, comprise much or most of plant genomes; their replication generates genomic diversity and makes them an excellent source of molecular markers (Schulman et al. 2004). The retrotransposon-based marker methods rely on PCR amplification between a conserved retrotransposon feature, most often the long terminal repeat (LTR), and another dispersed and conserved feature in the genome (Schulman et al. 2004). The inter-retrotransposon amplified polymorphism (IRAP) method displays insertional polymorphisms by amplifying the segments of DNA between two retrotransposons. It has been used in numerous studies of genetic diversity (e.g., Antonius-Klemola et al. 2006; Kalendar and Schulman 2006; Smýkal et al. 2008b; Vukich et al. 2009). Here, we have isolated the first large set of *Linum* retrotransposon segments, developed a novel method for fingerprinting flax germplasm based on IRAP and applied it to over 700 accessions in order to assess genetic diversity and germplasm structure.

Materials and methods

Plant materials

The accessions used in this study originated from the Czech National Flax Germplasm collection (http://genbank.vurv.cz/genetic/resources; Agritec Ltd, Šumperk; Table S1, Online Resource S1). For this study, 708 accessions of the

total 2,081 were analyzed. Because the collection is divided by passport data into three sections, we followed this scheme: 143 landraces (L, numbers X12), 387 cultivars (C, numbers X13, X14) and 178 breeding lines (Br, numbers X11). The end-use type was determined according to phenotypic differentiation (Kulpa and Danert 1962; Diederichsen and Fu 2006), with distinctions for fiber (convar. *elongatum*, taller than 70 cm, with only the upper quarter of the stem having side branches), linseed (convar. *mediterraneum*, with 1000-seed weight over 9 g and a lower branching stem), and the intermediate type for both uses (convar. *usitatissimum*). These represent both traditional landraces as well as modern varieties for oil (272) or fiber (351) production or for both uses (86 accessions) and originate from 36 countries spanning Africa, Asia, Australia, Europe, and both Americas. Sections of registered varieties and breeding lines were combined in the diversity analysis, as the latter represent material essentially not registered varieties. Both sections span the period of over last 70 years. Isolated genomic DNA from the following five geographically diverse accessions were used for initial cloning steps: 05X12-0004, -0046, -0052, -0097, -0194). In addition, 10 wild *Linum* species (*Linum angustifolium*, *L. perenne*, *L. tenuifolium*, *L. stelleroides*, *L. suffruticosum*, *L. austriacum*, *L. alpinum*, *L. flavum*, *L. thracicum*, and *L. grandiflorum*) were obtained from Institute of Natural Fibers and Medicinal Plants, Poznan, Poland. For assessment of intra-accession variation, eight individual plants from two accessions per section (e.g., 05X1100001 and 05X1100451 from X11, 05X120365 and 05X120284 from X12, 05X130008 and 05X131043 from X13) were analyzed. To investigate genetic stability over the time, seeds were hand-harvested from 10 individual plants of three accessions (e.g., 05X1100451 from X11, 05X120284 from X12, 05X130008 from X13) and separately cultivated over three subsequent years (2006–2008). All investigated accessions were sown in field trials located at Šumperk on standard 5 m^2 plots and/or in glasshouse conditions.

Genomic DNA isolation—bulk sample approach

Shoots were harvested from 10 randomly chosen young plants per accession and bulked for DNA isolation. This approach was taken in order to capture heterogeneity of accession (Fu 2005; Smýkal et al. 2008a, 2011; Kwon et al. 2010). Leaf material was stored at −80°C until DNA isolation. Genomic DNA was manually isolated using the Invisorb Plant Genomic DNA Isolation Kit (INVITEK, Germany). DNA obtained from approximately 100 mg fresh weight material was eluted in 300 µl of the kit's elution buffer at a concentration of about 50–100 ng/µl and stored at −20°C till use. The DNA quality was checked electrophoretically and spectrophotometrically.

Isolation of retrotransposon fragments

The iPBS method (Kalendar et al. 2010) was used to obtain retrotransposon fragments. The method is based on PCR amplification between the primer binding sites (PBS) for minus-strand cDNA synthesis, which are highly conserved in virtually all retrotransposons and located adjacent to the 5′ LTR (Antonius-Klemola et al. 2006; Vukich et al. 2009). The PCR amplification was performed with one or two primers complementary to the PBS domain (Table 1). All amplified PCR fragments were cloned into the pGEM-5Zf (Promega, USA) plasmid and sequenced using an ABI3700 (Applied Biosystems, USA) capillary sequencer. Retrotransposon segments within the clones were identified by a combination of approaches: by comparison to LTRs of known retrotransposons, by alignment of all sequenced PCR fragments to identify conserved repetitive segments, and by the universal structural features shared by retrotransposon LTRs, in particular the inverted terminal repeats at their flanks and the terminal nucleotides 5′ TG…CA 3′. The 3′ terminus of the 5′ LTR is generally located 1–5 nucleotides away from the PBS. To define full-length LTRs, multiple alignments were carried out as described below using isolated segments amplified by iPCR.

Sequence analyses

BLAST (Altschul et al. 1997) analysis was performed on the NCBI server using default settings. The DNA sequences were viewed, edited, and assembled using Sequence Scanner version 2.0 (Applied Biosystems). Alignments were made with the aid of CLUSTAL W software (Thompson et al. 1994). Final alignments were displayed with Boxshade 3.21 (ftp.ebi.ac.uk) software. Sequence assembly, primer design, and restriction analyses were performed with FastPCR software version 6.0 (Kalendar et al. 2009; http://primerdigital.com/fastpcr.html; PrimerDigital Ltd, Finland).

Table 1 Primer binding site (PBS) primers used to amplify flax LTR retrotransposon fragments

Number	Primer (5′-3′)
792	TAGGTCGGAACAGGCTCTGATACCA
793	GCAGGGAACTTGGATGCTGATACCA
795	ACCGAGCAACTTGAGCTCTGATACCA
796	CTAGGGTCAAGGGGGCTCTGATACCA
797	GGGAAATGGTCCGCTCTGATACCA
800	AGGGAACTTGGATGCCGATACCA
802	TAATGCTGAACTTGCTCCGATGCCA
803	CATGAACCTAGCTCACGATGCCA

IRAP amplification

PCR amplification was performed in 20 µl reactions in 0.2 ml tubes (Trefflab, Degersheim, Switzerland) or on 96-well PCR plates (BioPlastic, Landgraaf, The Netherlands) using an Eppendorf Gradient Master Cycler (Eppendorf, Hamburg, Germany) according to Smýkal (2006) and Smýkal et al. (2008a, b). The cycling program consisted of: 5 min initial denaturation at 94°C; 35 cycles of 94°C denaturation (30 s), 55°C primer annealing (30 s), and 72°C extension (3 min); a final extension for 10 min at 72°C. The PCR products were resolved either by electrophoresis on 1.7% agarose (Serva, Heidelberg, Germany) in 1× TBE buffer or, for better resolution, on 8% PAGE gels followed by EtBr staining and UV visualization.

Data analysis

Each IRAP band was treated as a single locus. The presence or absence of a fragment of a given length was recorded in binary code. Data were evaluated using POPGENE software version 1.32 (Yeh and Boyle 1997) for the Shannon index (*H'j*, Shannon and Weaver 1949) as defined for multilocus markers as $H'j = -\sum p_i \log p_i$, where p_i is the frequency of the *i*th fragment in the sample. Polymorphic information content (PIC) was calculated for each marker using the following formula: $\mathrm{PIC}_i = 1 - \sum P2_{ij}$, where P_{ij} is the frequency of the *j*th allele in clone (*i*) according to Smith et al. (1997). Wild *Linum* species DNA marker data were processed by NTSYS-pc version 2.2 (Rohlf 2006) using the SIMQUAL module with the Jaccard genetic similarity coefficient (GSj). Dendrograms were constructed from cophenic values. In addition, genetic distances were computed from GSj according to Reif et al. (2005). Analysis of molecular variance (AMOVA, Excoffier et al. 1992) was used to partition the total genetic variation to among- and within-population variance components and to calculate the genetic variation in each population, using the ARLEQUIN software ver. 3 (Excoffier et al. 2005). The phi-statistic (Φst), was used to describe the interpopulation distance and its level of significance was also calculated by the AMOVA component in ARLEQUIN ver. 3. The sub-samples (populations) were defined as germplasm sections, e.g., breeding lines, landraces and varieties, and as fiber and linseed (oil use). Partitioning of total allelic diversity (H_T) within and among germplasm sections was calculated using Nei's (1973, 1978) gene diversity statistics. Total allelic diversity (H_T) was partitioned into the within- (H_S) and among-population (D_{ST}) components, such that $H_T = H_S + D_{ST}$.

To investigate the genetic structure of the flax collection, the Bayesian method available in the Bayesian analysis of population structure (BAPS) software (Corander and Marttinen 2006; Corander et al. 2008) was used as described in Smýkal et al. (2008a). We used 25 estimation runs of the clustering algorithm for unlinked loci, such that the upper-bound values (*K*) for the number of clusters in each run ranged from 5 to 50.

Results

LTR isolation

Retroviruses as well as plant pararetroviruses and retrotransposons initiate minus-strand reverse transcription, an essential step of replication, with specific tRNAs of the host cell. The PBS site is complementary to the tRNA primer, and the iPBS method exploits this property as a means of amplifying segments of two LTR retrotransposons in close proximity to one another by use of primers matching common PBS motifs (Kalendar et al. 2010). To screen the flax genome for retrotransposons, a set of eight PBS primers (Table 1) were used in iPBS with a template comprising a mixture of five genotypes. Most primers produced multiple bands, expected for abundant and clustered retrotransposons. A total of 91 iPBS reaction products, which were 400 to 2,000 bp long, were cloned, sequenced, and aligned, identifying clusters of highly similar fragments. A total of 76 clusters were identified, which have features typical for LTRs (Table S2, Online Resource S2). Primers were designed to match these (Tables 2, S3, Online Resource S2), positioned close to the predicted end of the LTR end and oriented outwards.

Development of IRAP for flax germplasm genotyping

In order to test the suitability of the LTR primers for IRAP genotyping, a set of eight geographically diverse, cultivated flax lines (Table S1, Online Resource S1; Fig. 1) was used. On this set, each primer was first tested alone; selected ones were tested in combination with a second LTR primer. The scoring criteria were the number, sharpness, and evenness in intensity of the PCR products following electrophoresis, as well as the degree of polymorphism among the genotypes. About one quarter of primers yielded either poor amplifications or few products; these primers were discarded. Other primers (e.g., 1829, 1830, 1839, 1860, 1883, 1895; Table S3, Online Resource S2) amplified well, but produced primarily monomorphic products. A set of 10 primers with numerous, evenly distributed fragments (Fig. 1), having a degree of polymorphism higher than 30% and PIC values between 0.30 and 0.52 (Table 2), were retained for further work. Magnesium and DNA template concentration variations in the PCR reaction had no significant influence (data not shown).

Table 2 IRAP primers with sequence, polymorphism level, and polymorphic information content (PIC)

Primer number	Sequence (5′-3′)	GenBank accession number	Position and orientation	Amplified bands per sample (mean)	Percentage polymorphism	PIC value (across entire collection)
1826	ACCCCTTGAGCTAACTTTTGGGGTAAG	GU735096	1282 → 1308	28	42	0.30
1833	CTTGCTGGAAAGTGTGTGAGAGG	GU929874	8 ← 30	17	35	0.32
1838	TGTTAATCGCGCTCGGGTGGGAGCA	DQ767972	138 → 162	16	20	0.29
1845	AGCCTGAAAGTGTTGGGTTGTCG	GU980589	1111 ← 1133	17	47	0.51
1846	CTGGCATTTCCATTGTCGTCGATGC	GU980588	971 ← 995	19	41	0.42
1854	GCATCAGCCTGGACCAGTCCTCGTCC	GU929878	586 ← 611	17	47	0.52
1868	CACTTCAAATTTTGGCAGCAGCGGATC	GU735096	460 → 486	14	31	0.45
1881	TCGAGGTACACCTCGACTCAGG	GU929877	538 ← 559	25	36	0.48
1886	ATTCTCGTCCGCTGCGCCCCTACA	GU980590	831 → 854	16	31	0.49
1899	TGAGTTGCAGGTCCAGGCATCA	GU980587	31 ← 52	18	37	0.42

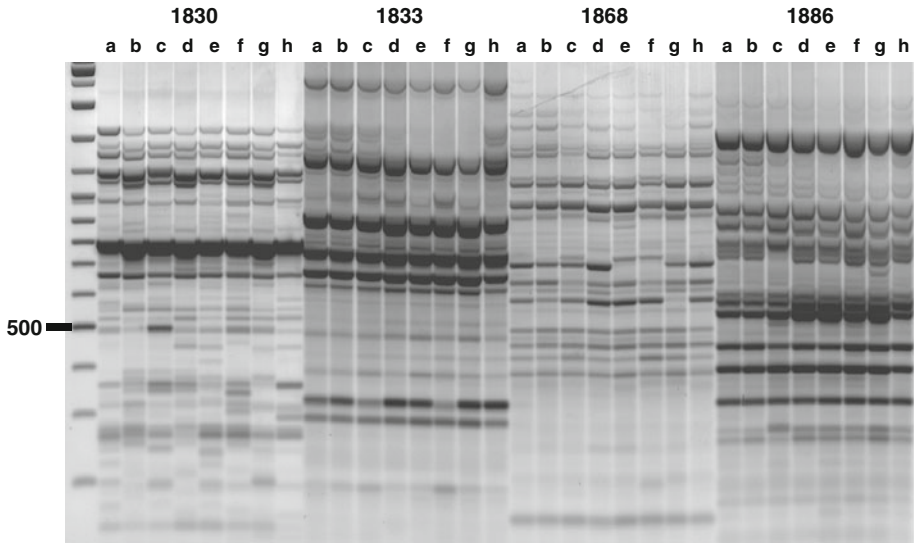

Fig. 1 IRAP analysis on a set of eight flax accessions (*a* 05X12-0004, *b* −0046, *c* −0052, *d* −0097, *e* −0194, *f* −0238, *g* −0269, *h* −0382). The codes for the primers (Table 2) used in the reaction are shown above the each set of lanes. A 100 bp sizing ladder (500 bp labeled) is shown on the left

Intra-accession variation

To assess the level of intra-accession variation, 10 individual plants of two accessions per section (e.g., 05X1100001 and 05X1100451 of X11, 05X120365 and 05X120284 of X12, 05X130008 and 05X131043 of X13) were analyzed with the three most polymorphic IRAP primers (1826, 1833 and 1845; Fig. 2). The data were evaluated by pair-wise resemblance for individual plants (Jaccard coefficients). Intra-accession variation was higher for landraces, with Jaccard similarity coefficients (JSC) of 0.86 and 0.78, while for breeding lines and varieties, JSC ranged from 0.92 to 0.97, respectively. Technical replicates of the same samples showed very low variation (0.99 JIC); (data are not shown).

Because certain flax genotypes have reported to be prone to genetic instability (Cullis 2005; Chen et al. 2005, 2009; Schneeberger and Cullis 1991), we analyzed three accessions and their progenies in three successive years.

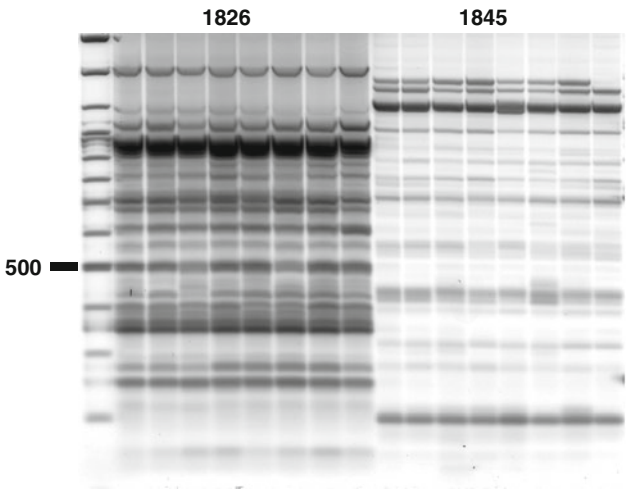

Fig. 2 Intra-accession stability as detected by IRAP. Eight individual plants of accessions 05X1100001 and 05X130008 were analyzed by IRAP with primers 1826 and 1845. A 100 bp sizing ladder (500 bp labeled) is shown on the left

For the three primers used (1826, 1833 and 1845), the IRAP fingerprint remained stable (data not shown).

Genetic diversity within cultivated flax germplasm

A total of 708 flax accessions were scored by IRAP with the 10 most informative primers, yielding 141 PCR fragments that could be scored from gel to gel, or a total of 99,828 scored products. Of these, 74 (52%) were polymorphic across the investigated set. The number of scorable bands per primer ranged from 14 (1868; 41% polymorphic) to 28 (1826; 42% polymorphic). Each score represents the summary fingerprint of 10 bulked plants; the dominant IRAP marker system hides intra-accession differences between the plants among the bulks. However, there were no accessions giving identical profiles in the analyzed set. As shown in Table 3, the percentage of polymorphic loci and Shannon's index, measures of genetic diversity, decreased from respectively 58% and 0.34 for the diverse landraces, through 50% and 0.33 in cultivars, to 48% and 0.3 for breeding lines. Conversely, the average Jaccard similarity index increased from 0.63 for landraces to 0.81 and 0.85 for cultivars and breeding lines, respectively. Analysis of gene diversity in subdivided populations (Nei 1973, 1978) gave a total gene diversity (H_T) value of 0.227, compared with an H_S of 0.215 within populations and a G_{ST} of 0.053 between populations.

Analyses of molecular variance (AMOVA) showed that although most of the genetic variation (76.55%) could be assigned to differences within the populations (i.e., to sections L, C and Br), whereas 23.45% could be ascribed to variation among them (Table S5, Online Resource S3). The inter-population or section phi-statistic (Φst) estimates indicated that the largest distances are between breeding lines and landraces (0.296, significant at $P < 0.001$), and landraces and cultivars (0.234, significant at $P < 0.001$), while the lowest is between cultivars and breeding lines (0.103, non-significant).

A Bayesian analysis of the marker data using the BAPS algorithm partitioned all 708 accessions into 20 clusters, associated with a posterior probability of 0.99933, and showed a high structuring of the set (Fig. 3). The stability

Table 3 Summary statistics for IRAP marker data from 708 *Linum* accessions

	Wild *Linum* species	Cultivated all	Breeding lines (X11)	Landraces (X12)	Cultivars (X13)
Number of accessions	11	708	178	143	387
Number of bands	1,672	99,828	25,098	20,163	54,567
% Polymorphic loci	100	52	48	58	50
Shannon index	0.68	0.34	0.30	0.34	0.33
Jaccard index (mean)	0.57	0.77	0.85	0.63	0.81
Nei gene diversity	0.46	0.22	0.21	0.26	0.22

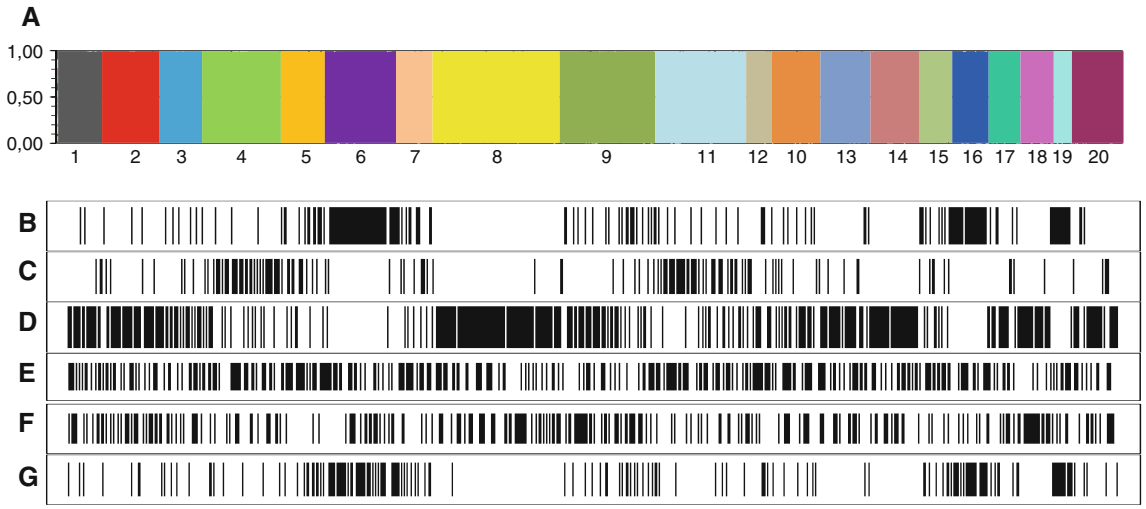

Fig. 3 Bayesian analysis of flax germplasm structure. **a** Altogether 708 accessions of cultivated flax were partitioned into 20 clusters with a posterior probability of 0.99933, using BAPS software. Each *bar* represents the accession and the scale on the left indicates the probability value to cluster assignment. Partitioning of 178 breeding lines (**b**), 143 landraces (**c**), 387 cultivars (**d**), 351 fiber use (**e**), 272 oil use (**f**) and 169 accessions of Czech (Czechoslovak) origin (**g**)

of this partitioning was verified by numerous (25) estimation runs, all of which converged into exactly 20 clusters when the prior upper bound K was set equal to at least 25. All solutions with less than 20 clusters (these were obtained when the upper bound for an estimation run was set <20) had extremely small posterior probabilities compared to the optimal solution. The size of individual clusters ranged from 12 (cluster 19) to 85 (cluster 8) with a mean of 30 accessions. There were five clusters: 4, 6, 8, 9, and 10 (respectively with 52, 50, 85, 63, and 63 accessions) comprising 313 altogether accessions (40% of the total). When the clusters were analyzed for the relative contribution of the germplasm sections to each, partitioning was observed for the breeding lines (Fig. 3b) and landraces (Fig. 3c). Although the breeding lines were dispersed among the clusters, they predominated in some (6, 16, 19) while landraces dominated in others (4, 10) that have been of fiber use (Fig. 3c, Table S4). As seen in Fig. 3d and Table S4, cultivars were the most dispersed across all 20 clusters, being the most dominant in clusters 1, 2, 3, 8, 9, 13, 14, 18, and 20 (Table S4, Online Resource S3). Nevertheless, 91 of the breeding lines (52%) were partitioned into three clusters (6, 9, and 16) and 72 of the landraces (50% of all) into just two clusters (4 and 10).

Retrotransposon markers do not differentiate between linseed and fiber flax

Visualization of user type in BAPS clusters showed even partitioning of both oil and fiber types (Fig. 3e, f). When the marker data from the BAPS analysis are displayed in a neighbor-joining (NJ) tree, two large groups are revealed (Fig. S1, Online Resource S3). To confirm this, we computed also multidimensional scaling (MDS) of the clusters based on average pair-wise Hamming distances between the accessions they contain. A three-dimensional plot (Fig. S2, Online Resource S3) of the MDS distances shows that the topology of the NJ tree is well founded. Five groups of clusters (4, 5, 6, 7, 8, 10, 18; 1, 2, 3, 9, 13, 14, 15, 16, 19, 20; 11; 12; 17) are particularly closely related, but correspond neither to type of use nor to geographical origin (Table S4, Online Resource S3 and data not shown). Closer examination of NJ tree (Fig. S1) of cluster relationships did not reveal any particular partitioning according to origin or user type, except for a higher proportion (107 vs. 36) of landraces in group I and, vice versa, a higher proportion (266 vs. 88) of breeding lines in group II. In group I, there are 88 breeding lines, 107 landraces, 176 cultivars of which 200 accessions are fiber type, 39 combined and 129 of oil type, while in group II, there are 266 breeding lines, 36 landraces, 211 cultivars of which 151 accessions are fiber type, 47 combined and 143 of oil type.

Although the germplasm is biased towards the fiber types (351 of 708), the assignment of user type showed that fiber (351) and combined-use (272) accessions distributed rather evenly except for clusters 18 and 19, with partitioning of over 128 fiber type accessions into four clusters (4, 6, 8, 10). The oil-type accessions (86 in total), however, were predominantly found in clusters 1, 2, 3, 4, 6, 8, 9, 10, and 16 (Table S4, Online Resource S3). A phi-statistic (Φst) analysis showed only low, statistically non-significant distances separated groups distinguished by oil versus fiber use (0.116), as was the separation between these groups and the accessions for combined use (0.058 and 0.063, respectively). Analyses of molecular variance (AMOVA) showed (Table S5, Online Resource S3) that there is a lower variance components number (12.67) between fiber accessions than for oil (21.30) or combined-use (25.20) accessions; consequently, among user types, only 8% of the variation was found.

Geographical distribution

There was no strict partitioning of diversity according to country of origin. Moreover the number of accessions for each country varied greatly (Table S1). The 708 accessions originated from 36 countries, with the largest group (169, e.g., 24%) being breeding lines and cultivars of Czech origin, reflecting the origin of germplasm collection (Table S1, Online Resource S1; Table S4, Online Resource S3). However, certain clustering pattern could be seen. Czech origin accessions were partitioned among nearly all clusters, except 14 and 18 (in which accessions of Bulgarian or Argentinean origin dominate). In clusters 5, 6, 9, 16, and 19, there were more than 10 accessions of Czech origin (Fig. 3g). Cluster 6 was particularly rich in Czech accessions (38 out of 49). This corresponds to breeding lines of fiber use, which was predominant breeding target in the past. As shown in the Fig. S1 dendrogram, there is wide separation between clusters 6 and 9, 16, 19, reflecting past (fiber use) and more recent (oil or combined use) breeding program focus using separate gene pools. Although the accessions of the next most abundant source, Argentina (63), partitioned into 13 clusters, significant proportion was found in two clusters 9 and 18 (10, 11 out of 63 and 22, respectively). These belong to cultivars and either fiber or combined use. Dutch accessions (56) were partitioned fairly evenly, with more than six accessions in each of clusters 8, 9 and 14, which are also fiber use cultivars. The American and Canadian accessions of fiber use cultivars (58 and 18 each) were predominantly in clusters 8 and 9. Those from Russia and the former USSR (34) were in clusters 8 and 10, whereas 10 accessions of German origin (out of 27) were in cluster 4. The two last largest groups of Brazilian and Uruguayan accessions (35 together) and

Bulgarian accessions (35) were in clusters 1, 2, and 8, and 14, respectively (Table S4, Online Resource S3). The remaining 227 accessions from 27 countries were more dispersed, with 1–17 accessions per country and maximally five per cluster (Table S1, Online Resource S1, and data not shown).

Wild flax genetic diversity

To determine the maximal discrimination power of IRAP and to examine genetic diversity within the *Linum* genus, we included 10 wild species. These were analyzed with 10 LTR-derived primers (1830, 1833, 1838, 1845, 1846, 1853, 1868, 1880, 1886, 1899). A dendrogram was built (Fig. 4) from the IRAP marker data and associated JSC. It shows *L. grandiflorum* (scarlet or crimson flax, Northern Africa) as the out-group, separated at a Jaccard coefficient of 0.23. Cultivated flax (*L. usitatissimum*) grouped together with *Linum angustifolium syn. L. bienne* (pale flax, Southern Europe), its presumed ancestor, at 0.42. Another cluster (at 0.76) was formed by *L. perenne*, *L. tenuifolium* (white flax, Central and Eastern Europe), *L. stelleroides* (Eastern Asia), *L. suffruticosum* (Southwestern Europe–Spain), *L. austriacum*, and *L. alpinum* (Central and Eastern Europe). The remaining cluster contained the yellow-flowering flax species *L. flavum* and *L. thracicum* (yellow flax, Southern and Eastern Europe) with a coefficient of 0.62. This separation of groups was maintained, although not so pronounced, even when only data from any two primers were analyzed (not shown), demonstrating the robustness of the IRAP-based separation. As shown in Table 3, analyses of these 11 accessions by 10 IRAP primers yielded 1,672 bands with 100% polymorphism, resulting in a high Shannon index (0.68) and a lower average Jaccard similarity index (0.57).

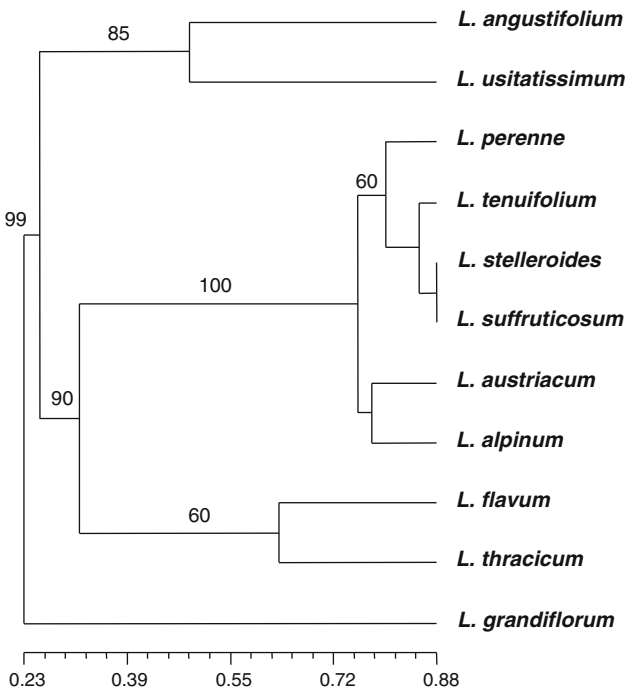

Fig. 4 UPGMA cluster analysis based on Jaccard similarity coefficients, showing the genetic relationships among 11 *Linum* species based on IRAP analysis. Bootstrap percentages above 50 are indicated

Discussion

Knowledge about the distribution of genetic diversity conserved in germplasm collections is crucially important for unlocking that diversity for further use (Tanksley and McCouch 1997). For ex situ collections, identification of redundant accessions (Fu et al. 2006) and establishment of core collections (Brown and Spillane 1999) are key tasks. Earlier molecular marker analyses by RAPD, ISSR; AFLP, and SSR showed a low diversity in flax germplasm (Cloutier et al. 2009; Fu 2002, 2005; Fu et al. 2002a, b, 2003; van Treuren et al. 2001; Wiesnerova and Wiesner 2004; Uysal et al. 2010). In this work, we have taken advantage of the ubiquity and abundance of LTR retrotransposons in plant genomes and their role in genomic diversification to develop and apply retrotransposon markers based on the IRAP method for the first time to flax (Kalendar and Schulman 2006).

Flax genotyping by retrotransposon markers

Using a recently introduced protocol called iPBS (Kalendar et al. 2010), we have cloned and sequenced about 90 DNA segments, in total length over 80 kb, containing retrotransposon segments. This represents the first large-scale identification of *Linum* transposable elements. Earlier, only two transposable elements had been isolated from *Linum*: a non-autonomous *Cassandra*-type TRIM retrotransposon by us (accession DQ767972; Kalendar et al. 2008) and a non-autonomous DNA transposon, *dLute* of the *hAT* Superfamily (AF036935; Luck et al. 1998). A single-copy insertion sequence of uncertain origin, *LIS*-1, has also been found (Chen et al. 2005). Both these elements are not suited for IRAP fingerprinting, due to their low copy number.

Of the 90 retrotransposon segments identified, 76 (85%) contained features typical for the LTRs of retrotransposons, which allowed the design of LTR-specific primers. Fifty primers yielded the multi-locus fingerprints typical for the IRAP method and displayed sufficiently high levels of polymorphism to be generally applicable. The 10 most informative primers were applied to 708 flax accessions. The IRAP primers were highly (100%) polymorphic across

10 wild *Linum* species, but less so (52%) within the cultivated germplasm. Although we have not aimed to estimate phylogenetic relationships, the IRAP analysis of wild *Linum* species showed good correspondence with the phylogeny inferred from chloroplast and nuclear markers (McDill et al. 2009; Fu and Allaby 2010). The resulting IRAP-derived dendrogram placed together cultivated flax (*L. asitatissimum*) and its putative ancestor *Linum bienne* syn. *L. angustifolium*. The placement of *L. grandiflorum* as the out-group by IRAP marker analysis is not in accord earlier results (Fu et al. 2002b) and may be attributable to the sampling of different accessions here, to taxonomic mis-classification or to sampling of different genome regions, or to shared characters (RAPD vs. IRAP bands of shared mobility) actually being of different origins. We can exclude taxonomic mis-classification since all wild material was grown and taxonomically determined; moreover, *L. grandiflorum* is an easily distinguished species. The risk of the shared characters (DNA amplified fragments) error increases with phylogenetic distance. Because wild species and cultivated germplasm did not fully overlap in the IRAP dataset we have not analyzed these by BAPS and consequently we have not compared the diversity between cultivated and wild germplasm.

The high diversity shown across the *Linum* genus is notable because the retrotransposon markers were developed from cultivated accessions. Normally, retrotransposon markers show higher polymorphism in their native species. Hence, although the full-length retrotransposons corresponding to the IRAP primers were not identified (except the individual *Cassandra*-type TRIM retroelement), they appear to be both conserved and active in multiple *Linum* species. Furthermore, the large proportion of interspecies variability expressed as polymorphic bands suggest that only few, pre-speciation element insertions remained fixed, as found in the *Helianthus* genus by Vukich et al. (2009).

Retrotransposons and flax genome stability

The *L. usitassimum* genome is estimated at about 686 Mbp, about the same as the 730 Mbp sorghum genome, which is 55% retrotransposons (Paterson et al. 2009). The difference in genome size between sorghum and rice (Paterson et al. 2009), and indeed between large and small plant genomes in general, can be attributed to the accumulation of retrotransposon insertions over time. Moreover, small genomes remain small in the face of retrotransposon replication through loss of the inserted copies over time (International Brachypodium Initiative 2010). Hence, the polymorphic patterns of inter-retrotransposon amplifications displayed by IRAP are generated by gain and loss of retrotransposons over time, as well as insertions or deletions in the intervening DNA and mutations that destroy PCR priming sites. In practice, recombinational loss has less effect on the retrotransposon display pattern than do new insertions because the LTR–LTR recombination, an important mechanism of loss (International Brachypodium Initiative 2010), leaves behind an LTR and thus does not affect the PCR priming site.

While retrotransposon markers are sufficiently stable to allow their use in mapping projects (e.g., Huo et al. 2009; Tanhuanpää et al. 2008), they are nevertheless sensitive enough to detect rapid genome changes (Belyayev et al. 2010; Petit et al. 2010). Genomic instability was reported (Durrant 1962) in certain flax genotypes, particularly var. Stormont Cirrus (Chen et al. 2005, 2009), whereby a strong influence of the environment on the genome leads to putative adaptive mutations (genotrophs). Morever, there are several reports that demonstrate that transposable element insertional dynamics could promote or intensify morphological and karyotypical changes, some of which may be potentially important for the process of microevolution, thereby allowing species with plastic genomes to survive as new forms or even as new species in times of rapid climatic change (Belyayev et al. 2010). We therefore tested the stability of IRAP fingerprints to see if we might detect retrotransposon mobilization associated with genotrophs. None of the analyzed plants in course of 3 years' trials displayed fingerprint changes, including var. Stormont Cirrus (05X130008) in which the genotrophs were generated (Cullis 2005). However, the IRAP experiments did not display all the retrotransposons in the genome and would not detect nested insertions located within the retrotransposon "behind" (5′ to) the PCR priming sites, which are located near the LTR termini. Hence, we cannot exclude that stress might activate retrotransposons in *Linum*.

IRAP does not detect disruptive selection for oil and fiber morphotypes

Flax was used by early humans even 30,000 years ago as the source of fiber (Kvavadze et al. 2009); it was among the first crops domesticated in the Near East some 10,000 years ago (Zohary and Hopf 2000) and is thought to have undergone disruptive selection for oil and fiber morphotypes (Diederichsen and Fu 2006; Diederichsen 2007). However, the differences between morphotypes are very subtle. The oil varieties are short and have large seeds containing 40% oil, whereas fiber varieties have tall, more sparsely branched plants with fewer smaller seeds (Zohary and Hopf 2000; Diederichsen and Fu 2006). Although these are quantitative traits, more influenced by environment, they have high heritability.

In spite of some variation, there is phenotypic antagonism between linseed, oil (convar. *elongatum*), and fiber (convar. *mediterraneum*) varieties (Diederichsen and Fu 2006). This division is supported to some extent also by molecular data, such as allelic diversity at the *sad2* locus, which encodes steroyl-ACP-desaturase II acting in unsaturated fatty acid biosynthesis (Allaby et al. 2005), as well as by several chloroplast markers (Fu and Allaby 2010). Interestingly, both these studies showed that selection for seed lipids was associated with the domestication of flax, whereas fiber use likely came later. However, these results also could be interpreted to show that stronger selection was applied or required to fix the oil character, with fiber types being ancestral. Our IRAP data, which did not separate the flax collection by virtue of use type, consonant with the results of Everaert et al. (2001), Wiesnerova and Wiesner (2004), and Diederichsen and Fu (2006), is consistent with fiber types being at least as ancient as oil types.

The lack of resolution with IRAP markers for use type might be attributed to the portion of the genome sampled because retrotransposons, especially those sampled by methods like IRAP, tend to be clustered in intergenic regions at least in some genomes. Thus, the visualized IRAP fragments might not be tightly linked to portions of the genome carrying the specific genes governing traits for use type. This finding is reminiscent of our recent study on the large pea germplasm, although it was made with a locus-specific, co-dominant retrotransposon method that does not require two retrotransposons to be near to each other. In that study, no specific assignment to morphological types could be made (Jing et al. 2010; Smýkal et al. 2011). Nevertheless, in the current study, BAPS analysis showed that some clusters contain more fiber than oil flax types (Table S4, Online Resource S3). As mentioned earlier, fiber use had more antiquity than oil use, although there is no general concensus on this (Allaby et al. 2005). This contrasts with AFLP-based analyses, where distinctions were obtained, although no specific AFLP fragments could be associated with a given morphotype (van Treuren et al. 2001; Vromans 2006). A comparable number of fragments (141 per accession) was amplified in our and Vroman's AFLP-based (2006) or RAPD-based studies by Fu (2005). The differing results may be result from the genic and generally neutral nature of retrotransposon and also RAPD markers, in contrast to AFLP, which does not distinguish genic from non-genic polymorphisms. Alternatively, accession choice or number, as seen for pea (Smýkal et al. 2008a, b, 2011), could explain the differences. In our AMOVA analysis, 23.45 and 76.55% of the variation is among and within fiber and oil populations, respectively, very close to that for AFLP (22.7 and 77.3%; Vromans 2006).

Genetic diversity in cultivated flax

Genetic diversity is a commonly thought to narrow as a consequence of plant breeding (Tanksley and McCouch 1997), and low diversity was seen in cultivated flax during variety testing for distinctness, uniformity and stability (DUS) criteria using morphological descriptors (Everaert et al. 2001; Diederichsen and Fu 2006). Narrow germplasm limits sources of novel alleles and haplotypes for trait improvement; Vromans (2006) concluded that further decreases in the genetic variation of fiber flax will result if only modern cultivars are used in future breeding. Our data support this view and show that the polymorphism of IRAP loci decreases from landraces to commercial varieties. Since we analyzed only 10 accessions from diverse wild *Linum* species and, crucially, only one accession represented the wild progenitor, pale flax (*L. bienne*), our data cannot at present provide support for the idea that introgression from wild material could widen the genetic base. Efforts have begun in this direction; recently 34 pale flax accessions from Turkey for their diversity patterns, which showed significant associations with geographic distances and elevation differences (Uysal et al. 2010). Comparable studies of wild progenitors and cultivated germplasm are limited but needed both for understanding of domestication and use in breeding (Smýkal et al. 2011).

Residual variation nevertheless persists within cultivars. This intra-accession variation is common in landraces and even varieties of all crops (Smýkal 2006; Smýkal et al. 2008b; Cieslarová et al. 2010), being higher in open-pollinating species. A flax cultivar may consist of a number of phenotypically similar F5 or F6-derived sister lines that, at the molecular level, still may differ for a small proportion of the genome (van Treuren et al. 2001; Vromans 2006). Moreover, especially for landraces, such variation might be more pronounced (Diederichsen and Raney 2008) and could be partly the result of transposable element activity restricted to local adaptive process (Belyayev et al. 2010). The small, but detectable intra-accession variation found by IRAP here as well as earlier by RAPD and AFLP (Fu et al. 2003; van Treuren et al. 2001; Vromans 2006) is consistent with this structure. Underlying the heterogeneity is outcrossing (Everaert et al. 2001; van Treuren et al. 2001), which has been estimated at less than 6%, although higher values levels be observed due to phenotypic variation in flower morphology (Williams 1988). To capture this genetic heterogeneity within genebank accessions (Smýkal et al. 2008a; Cieslarová et al. 2010), a bulk approach is commonly used for molecular analysis, including flax germplasm (Fu 2005). Thus, in our study, we have used a sample size of 10 plants per accession, comparable to that in previous studies in flax (Fu 2005; van Treuren et al. 2001; Vromans 2006) used dominant RAPD or AFLP

markers. The issue of bulk versus single plant assessment in plant genetic resources is still current, because while a single plant will not capture the genetic (and phenotypic) makeup of heterogeneous accessions, the cost of analysis may be decisive when faced with hundreds or thousands of accessions. As an alternative, many genetic diversity assessments have involved genotyping a representative DNA sample prepared from bulked tissue from several individuals of an accession because this bulking strategy increases the efficiency of germplasm characterization. As cited in Kwon et al. (2010), Fu et al. (2003) summarized that from the 81 papers published during 1997–2002, approximately an equal number of researchers fell into the following three categories: (1) using more than 10 plants per bulk; (2) using 6–10 plants per bulk; (3) using less than six plants per bulk. There are several published studies to compare individual versus bulk analysis using not only co-dominant (Eschholz et al. 2008), but also dominant markers such as AFLP. The latter showed that although bulking could reveal up to 20% less variation, genetic relationships inferred by bulking were largely the same as those from single plant (Fu et al. 2006). It is highly likely, depending on the bulk size the complexity of the amplification target set in the genome, and competition between templates in the PCR reaction, that low-frequency markers will be amplified from the bulked DNA samples, but the genetic diversity and relationship among accessions should be based on common alleles rather than on rare alleles (Kwon et al. 2010).

Bayesian germplasm clustering

As traditional methods for characterizing genetic population structure rely on a priori grouping of individuals, Bayesian methods were introduced to avoid this limitation by using linkage and Hardy–Weinberg disequilibrium to decompose a sample of individuals into genetically distinct groups (Latch et al. 2006). We have chosen BAPS software (Corander and Marttinen 2006), because STRUCTURE (Pritchard et al. 2000) often experiences difficulties in identifying K values in domestic and wild species, especially for large datasets (Jing et al. 2010; Smýkal et al. 2011). Moreover, as shown in the comparative study of Latch et al. (2006) and Smýkal et al. (2011), BAPS performs very well at low levels of population differentiation, with an F_{ST} lower than 0.03. This is often case in germplasm collections, which are not true populations. In our study, flax germplasm differentiation was low (G_{ST} = of 0.053), supported by AMOVA results. Despite this, the conclusive separation of 708 analyzed flax accessions into 20 clusters was achieved with a relatively even distribution of accessions (from 12 to 85 per cluster, mean 35). Although, there was no strict partitioning according to origin, accession or user type, some clusters contained a larger proportion of specific accessions than others. As shown here and previously for pea (Smýkal et al. 2008a, b, 2011), BAPS provides a powerful analytical tool for germplasm structure that is computationally efficient, suits complex data sets (Latch et al. 2006), accommodates spatial models of genetic populations, and addresses admixture inference (Corander et al. 2008). The later could not be tested in our dataset, as IRAP is an essentially dominant marker system, although some co-dominant character is seen (Smýkal 2006; Kalendar et al. 2010 and unpublished data). Moreover, using bulked samples would hamper admixture estimation as this is very sensitive under these circumstances. Critically, our study focused on estimating the genetically divergent groups, i.e., genetic mixture estimation, not admixture estimation. Nevertheless, each accession in the examined dataset produced a unique fingerprint, despite bulking. Consequently, Bayesian analysis provides valuable information about partitioning of genetic diversity within flax germplasm.

In conclusion, our results provide bases for better flax germplasm management, core collection establishment, and exploration of diversity in breeding. This study demonstrates the utility of retrotransposons, a dominant and ubiquitous part of eukaryotic genomes as well as model-based Bayesian analysis, for diversity studies. While the flax germplasm was not partitioned according to origin or use, the identification of eight clusters containing high proportions (70–100%) of commercial cultivars, two clusters of landraces, and three clusters of breeding lines will help in the construction of a core collection. We have not aimed here to specify which accessions in particular should be assigned to a core collection; this task awaits inclusion of morphological and digital image analyses of phenotypic traits as well as agronomic data.

Acknowledgments This work was financially supported by the Ministry of Education of Czech Republic research project MSMT 2678424601. The technical support of Mrs. L. Vítámvásová and E. Fialová, as well as the laboratory support of Anne-Mari Narvanto and Ursula Lönnqvist of the Institute of Biotechnology, University of Helsinki, are gratefully acknowledged. P.S. acknowledges the OECD fellowship AGR/PR(2006)1-JA00035882, within which flax LTR sequences and IRAP primers were generated in the Schulman laboratory.

References

Allaby RG, Peterson GW, Merriwether DA, Fu YB (2005) Evidence of the domestication history of flax (*Linum usitatissimum* L.) from genetic diversity of the *sad2* locus. Theor Appl Genet 112:58–65

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new

generation of protein database search programs. Nucleic Acids Res 17:3389–3402
Antonius-Klemola K, Kalendar R, Schulman AH (2006) TRIM retrotransposons occur in apple and are polymorphic between varieties but not sports. Theor Appl Genet 112:999–1008
Belyayev A, Kalendar R, Brodsky L, Nevo E, Schulman AH, Raskina O (2010) Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. Mob DNA 1:6
Brown AHD, Spillane C (1999) Implementing core collections—principles, procedures, progress, problems and promise. In: Johnson RC, Hodgkin T (eds) Core collections for today and tomorrow. IPGRI, Rome
Chen Y, Schneeberger RG, Cullis CA (2005) A site-specific insertion sequence in flax genotrophs induced by environment. New Phytol 167:171–180
Chen Y, Lowenfeld R, Cullis CA (2009) An environmentally induced adaptive (?) insertion event in flax. Int J Genet Mol Biol 3:38–47
Chennaveeraiah MS, Joshi KK (1983) Karyotypes in cultivated and wild species of *Linum*. Cytologia 48:833–841
Cieslarová J, Smýkal P, Dočkalová Z, Hanáček P, Procházka S, Hýbl M, Griga M (2010) Molecular evidence of genetic diversity changes in pea (*Pisum sativum* L.) germplasm after long-term maintenance. Genet Resour Crop Evol. doi:10.1007/s10722-010-9591-3
Cloutier S, Niu Z, Datla R, Duguid S (2009) Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). Theor Appl Genet 119:53–63
Corander J, Marttinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. Mol Ecol 15:2833–2843
Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinform 9:539
Cullis CA (2005) Mechanisms and control of rapid genomic changes in flax. Ann Bot 95:201–206
Diederichsen A (2007) Ex situ collections of cultivated flax (*Linum usitatissimum* L.) and other species of the genus *Linum* L. Genet Resour Crop Evol 54:661–678
Diederichsen A, Fu YB (2006) Phenotypic and molecular (RAPD) differentiation of four infraspecific groups of cultivated flax (*Linum usitatissimum* L. subsp. *usitatissimum*). Genet Resour Crop Evol 53:77–90
Diederichsen A, Hammer K (1995) Variation of cultivated flax (*Linum usitatissimum* L. subsp. *usitatissimum)* and its wild progenitor pale flax (subsp. *angustifolium* (Huds.) Thell.). Genet Resour Crop Evol 42:263–272
Diederichsen A, Raney JP (2008) Pure-lining of flax (*Linum usitatissimum* L.) genebank accessions for efficiently exploiting and assessing seed character diversity. Euphytica 164:255–273
Durrant A (1962) The environmental induction of heritable changes in *Linum*. Heredity 17:27–61
Eschholz TW, Stamp P, Peter R, Hund A (2008) Genetic diversity of Swiss maize (*Zea mays* L. ssp *mays*) assessed with individuals and bulks on agarose gels. Genet Resour Crop Evol 57:71–84
Evans GM, Durrant A, Rees H (1966) Associated nuclear changes in the induction of flax genotrophs. Nature 212:697–699
Everaert I, de Riek J, de Loose M, van Waes J, van Bockstaele E (2001) Most similar variety grouping for distinctness evaluation of flax and linseed (*Linum usitatissimum* L.) varieties by means of AFLP and morphological data. Plant Var Seeds 14:69–87
Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491
Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50
Frankel OH, Brown AHD (1984) Plant genetic resources today: a critical appraisal. In: Holden JHW, Williams JT (eds) Crop genetic resources: conservation & evaluation. George Allan & Unwin, London, pp 249–257
Fu YB (2002) Redundancy and distinctness in flax germplasm as revealed by RAPD. Plant Genet Resour 4:117–124
Fu YB (2005) Geographic patterns of RAPD variation in cultivated flax. Crop Sci 45:1084–1091
Fu YB, Allaby RG (2010) Phylogenetic network of *Linum* species as revealed by non-coding chloroplast DNA sequences. Genet Resour Crop Evol 57:667–677
Fu YB, Diederichsen A, Richards KW, Peterson G (2002a) Genetic diversity within a range of cultivars and landraces of flax (*Linum usitatissimum* L.) as revealed by RAPDs. Genet Resour Crop Evol 49:167–174
Fu YB, Peterson G, Diederichsen A, Richards KW (2002b) RAPD analysis of genetic relationships of seven flax species in the genus *Linum* L. Genet Resour Crop Evol 49:253–259
Fu YB, Guerin S, Peterson GW, Diederichsen A, Rowland GG, Richards KW (2003) RAPD analysis of genetic variability of regenerated seeds in the Canadian flax cultivar CDC Normandy. Seed Sci Technol 1:207–211
Fu YB, Legge WG, Clarke JM, Richards KW (2006) Genetic variation and relationships of pedigree-known oat, wheat, and barley cultivars releaved by bulking and single-plant sampling. Genet Resour Crop Evol 53:1153–1164
Gill KS, Yermanos DM (1967) Cytogenetic studies on the genus *Linum* L. Hybrids among taxa with 15 as the haploid chromosome number. Crop Sci 7:623–627
Goldsbrough PB, Cullis CA (1981) Characterization of the genes for ribosomal RNA in flax. Nucleic Acids Res 9:1301–1309
Grandbastien MA (1998) Activation of plant retrotransposons under stress conditions. Trends Plant Sci 3:181–187
Haggans CJ, Hutchins AM, Olson BA, Thomas W, Martini MC, Slavin JL (1999) Effect of flaxseed consumption on urinary estrogen metabolites in postmenopausal women. Nutr Cancer 33:188–195
Hickey M (1988) 100 families of flowering plants, 2nd edn. University Press, Cambridge
Hilakivi-Clarke L, Clarke R, Lippman M (1999) The influence of maternal diet on breast cancer risk among female offspring. Nutrition 15:392–401
Huo H, Conner JA, Ozias-Akins P (2009) Genetic mapping of the apospory-specific genomic region in *Pennisetum squamulatum* using retrotransposon-based molecular markers. Theor Appl Genet 119:199–212
International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768
Jing R, Vershinin A, Grzebyta J, Shaw P, Smýkal P, Marshall D, Ambrose MJ, Ellis THN, Flavell AJ (2010) The genetic diversity and evolution of field pea (*Pisum*) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis. BMC Evol Biol 10:44
Kalendar R, Schulman AH (2006) IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nat Protoc 1:2478–2484
Kalendar R, Tanskanen J, Chang W, Antonius K, Sela H, Peleg O, Schulman AH (2008) *Cassandra* retrotransposons carry independently transcribed 5S RNA. Proc Natl Acad Sci USA 105:5833–5838

Kalendar R, Lee D, Schulman AH (2009) FastPCR software for PCR primer and probe design and repeat search. Genes Genomes Genomics 3:1–14

Kalendar R, Antonius K, Smýkal P, Schulman AH (2010) iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. Theor Appl Genet 121:1419–1430

Kulpa W, Danert S (1962) Zur Systematik von *Linum usitatissimum* L. (On the systematics of *Linum usitatissimum* L.) Kultupflanze. Beiheft 3:341–388

Kvavadze E, Bar-Yosef O, Belfer-Cohen A, Boaretto E, Jakeli N, Matskevich Z, Meshveliani T (2009) 30,000-year-old wild flax fibers. Science 325:1359

Kwon SJ, Hu J, Coyne CJ (2010) Genetic diversity and relationship among faba bean (*Vicia faba* L.) germplasm entries as revealed by TRAP markers. Plant genetic resources: characterization and utilization. doi:10.1017/S1479262110000201

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conserv Genet 2:295–302

Luck JE, Lawrence GJ, Finnegan EJ, Jones DA, Ellis JG (1998) A flax transposon identified in two spontaneous mutant alleles of the L6 rust resistance gene. Plant J 3:365–369

McDill J, Repplinger M, Simpson BB, Kadereit JW (2009) The phylogeny of *Linum* and *Linaceae* subfamily *Linoideae*, with implications for their systematics, biogeography, and evolution of heterostyly. Syst Bot 34:386–405

Menotti-Raymond M, David VA, Pflueger SM, Lindblad-Toh K, Wade CM, O'Brien SJ, Johnson WE (2008) Patterns of molecular genetic variation among cat breeds. Genomics 1:1–11

Muravenko OV, Yurkevich OY, Bolsheva NL et al (2009) Comparison of genomes of eight species of sections *Linum* and *Adenolinum* from the genus *Linum* based on chromosome banding, molecular markers and RAPD analysis. Genetica 135:245–255

Nei M (1973) Analysis of gene diversity in subdivided 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

Paterson AH, Bowers JE, Bruggmann R et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Petit M, Guidat C, Daniel J et al (2010) Mobilization of retrotransposons in synthetic allotetraploid tobacco. New Phytol 186:135–147

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

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

Rohlf F (2006) NTSYSpc: numerical taxonomy system (ver. 2.2). Exeter Publishing, Ltd, Setauket

Schneeberger R, Cullis CA (1991) Specific DNA alterations associated with the environmental induction of heritable change in flax. Genetics 128:619–630

Schulman AH, Flavell AJ, Ellis TH (2004) The application of LTR retrotransposons as molecular markers in plants. Methods Mol Biol 260:145–173

Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res 10:908–915

Smith JSC, Chin EC, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S, Ziegle J (1997) An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.) comparisons with data from RFLPS and pedigree. Theor Appl Genet 95:163–173

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

Smýkal P, Hýbl M, Corander J, Jarkovský J, Flavell AJ, 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

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, Kenicer G, Flavell AJ, Kosterin O, Redden RJ, Ford R, Zong X, Coyne CJ, Maxted N, Ambrose MJ, Ellis THN (2011) Phylogeny, phylogeography and genetic diversity of the *Pisum* genus. Plant Genetic Resources. doi:10.1017/S147926211000033X

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

Tanhuanpää P, Kalendar R, Schulman AH, Kiviharju E (2008) The first doubled haploid linkage map for cultivated oat. Genome 51:560–569

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

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Uysal H, Fu YB, Kurt O, Peterson GW, Diederichsen A, Kusters P (2010) Genetic diversity of cultivated flax (*Linum usitatissimum* L.) and its wild progenitor pale flax (*Linum bienne* Mill.) as revealed by ISSR markers. Genet Resour Crop Evol 57:1109–1119

van Treuren R, van Soest LJM, van Hintum TJL (2001) Marker-assisted rationalisation of genetic resource collections: a case study in flax using AFLPs. Theor Appl Genet 103:144–152

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

Vromans J (2006) Molecular genetic studies in flax (*Linum usitatissimum* L.). Ph.D. Thesis, Wageningen University, The Netherlands

Vukich M, Schulman AH, Giordani T, Natali L, Kalendar R, Cavallini A (2009) Genetic variability in sunflower (*Helianthus annuus* L.) and in the Helianthus genus as assessed by retrotransposon-based molecular markers. Theor Appl Genet 119:1027–1038

Wiesnerova D, Wiesner I (2004) ISSR-based clustering of cultivated flax germplasm is statistically correlated to thousand seed mass. Mol Biotechnol 26:207–214

Williams IH (1988) The pollination of linseed and flax. Bee World 69:145–152

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

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

Zohary D, Hopf M (2000) Domestication of plants in the Old World: the origin and spread of cultivated plants in West Asia, Europe and the Nile Valley. Oxford University Press, Oxford, p 316