ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization **8**(2); 95–105
doi:10.1017/S1479262109990220

Marker-assisted reduction of redundancy in a genebank collection of cultivated lettuce

R. van Treuren[1]*, E. C. de Groot[1], I. W. Boukema[1], C. C. M. van de Wiel[2] and Th. J. L. van Hintum[1]

[1]*Centre for Genetic Resources, The Netherlands, Wageningen University and Research Centre, PO Box 16, 6700 AA Wageningen, The Netherlands and* [2]*Plant Research International B.V., Business Unit Biodiversity and Breeding, Wageningen University and Research Centre, PO Box 16, 6700 AA, Wageningen, The Netherlands*

Received 28 September 2009; Accepted 30 November 2009 – First published online 5 January 2010

Abstract

To reduce the level of redundancy in a collection of cultivated lettuce, data from 160 amplified fragment length polymorphism (AFLP) fragments and 10 polymorphic microsatellites were used in combination with passport data and morphological data, the latter obtained from an experimental field trial performed for verification purposes. Based on the observed distribution of the number of marker differences between and within accessions, a minimum of three AFLP differences and two microsatellite differences were regarded as levels warranting distinction between accessions in the redundancy analysis. The strategy followed in the redundancy analysis was mainly based on the confirmation of duplication by each of two independently generated data sources. The molecular data were used for the validation as well as the identification of potential duplicates, revealing a total number of 198 redundancies, corresponding to 12.9% of the total collection. Trueness to type, number of characterization and evaluation data, and collection management considerations, such as available seed quantities and germination percentages, were used as primary, secondary and tertiary criteria to decide which accession from duplication groups to maintain in the collection. Removal of accessions showed negligible effects on total collection diversity, as quantified for AFLPs and microsatellites, characterization and evaluation traits and resistance profiles against downy mildew pathotypes, indicating that the applied strategy was effective.

Keywords: genebank efficiency; germplasm collection; *Lactuca sativa*; molecular markers; rationalization

Introduction

Since the establishment of crop genetic resources collections for present and future utilization, acquisition of germplasm has been a key activity of genebanks. Many collections started as working collections for taxonomic studies and plant-breeding activities, and were subsequently expanded with obsolete varieties and materials obtained from expeditions in natural distribution areas of crop-related wild relatives. Moreover, it has been common practice to extend collections by inclusion of material from colleague genebanks. It is estimated by the Food and Agriculture Organization of the Untied Nations that on a global scale over six million accessions are stored in approximately 1500 collections (FAO, 2009). However, it has also been recognized since long that only a limited share of the genebank accessions conserved worldwide are to be considered unique (Plucknett *et al.*, 1987; FAO, 2009). For example, based on an analysis of passport data, it has been shown that if the lettuce collections of the Centre for Genetic Resources, The Netherlands (CGN),

* Corresponding author. E-mail: robbert.vantreuren@wur.nl

the Institut für Pflanzengenetik und Kulturpflanzenforschung, the Horticulture Research International and the Western Regional Plant Introduction Station would be combined, about 40% of the material would be redundant (van Hintum, 2000). Because of the high number of accessions and presumed high levels of duplication, both within and between collections, many genebanks have become more concerned with the assessment of diversity within their collections and with the improvement of the efficiency of genetic resources management, than with the acquisition of material.

The European Cooperative Programme for Plant Genetic Resources (ECPGR) has started to create A European Genebank Integrated System (AEGIS) for plant genetic resources for food and agriculture, aiming at the conservation of the genetically unique and important accessions for Europe and seeking ways to secure their availability for breeding and research (ECPGR, 2009). The rationale of AEGIS is the creation of a set of virtual crop collections consisting of genetically unique accessions within the European collections that will be stored and maintained by the holding genebank under agreed minimum standard conditions, and distributed freely. One of the key elements of AEGIS is the identification of the most appropriate accession among potential duplicates that subsequently will be designated as an AEGIS accession, after which genebanks may decide to remove redundant accessions from their collection in order to increase efficiency. Prior to comparisons among collections, in-house analysis of diversity and identification of redundancies will obviously facilitate the assignment of most appropriate accessions.

Identification of potential duplicates is often carried out by examining the passport data of accessions (van Hintum and Knüpffer, 1995; Greene and Pederson, 1996). For example, passport data may show that accessions are conserved under identical, similar or synonymous names, that accessions have a common genetic background according to pedigree information, or that samples were collected from the same geographic locality. However, accessions may be conserved by genebanks under incorrect names, varieties carrying identical names may have been maintained by different breeding companies under different selection regimes, or genetically distinct samples may have been collected from a highly heterogeneous population. Therefore, potential duplicates based on passport data need to be validated by additional data sources in order to avoid unjustified decisions. In that context, molecular marker technologies are nowadays increasingly being used to assess similarity (Van Treuren and van Hintum, 2003; Spooner *et al.*, 2005).

The advantage of molecular marker technologies is their high potential to detect genetic differences, not only between samples from distinct accessions but often also from within accessions. The level of intra-accession diversity that can be expected generally declines with increasing levels of autogamy. However, even for highly autogamous crops such as barley, intra-accession variation is not uncommon (Van Treuren and van Hintum, 2001). Other factors affecting the observed intra-accession variation include the type of material studied, i.e. cultivars, landraces or wild populations, insufficient purification during cultivar development, mutations at marker loci and errors in determining the marker profiles. Moreover, the genetic integrity of accessions may have been affected during *ex situ* maintenance, for example as a result of genetic drift, unintentional selection or contamination during regeneration. All these factors cause potential duplicate accessions to be rarely completely identical. Therefore, a key issue in redundancy studies using molecular data is the level of inter-accession variation that is acceptable to confirm the duplicate status of accessions. This can be expected to vary between studies depending on the type of material investigated and the resolving power of the markers used. Analysis of molecular variance (AMOVA) in an amplified fragment length polymorphism (AFLP) study in flax (Van Treuren *et al.*, 2001) and bootstrap procedures in a microsatellite study in barley (Lund *et al.*, 2003) have been used to distinguish statistically homogeneous groups. However, AMOVA is not very useful in case of homogeneous accessions, while in the barley study, intra-accession variation was not considered in the analysis.

An entire collection of cultivated lettuce and its wild relatives was previously screened with AFLPs and microsatellites (van Hintum, 2003), while earlier analysis based on passport data had revealed a potential redundancy of about 8% (van Hintum, 2000). Here, we evaluated the distribution of molecular diversity between and within accessions of this collection of cultivated lettuce in order to obtain threshold values for inter-accession variation, which were subsequently used in a redundancy analysis aiming at reduction of the size of the collection.

Material and methods

Study material

Cultivated lettuce (*Lactuca sativa* L.) is a self-pollinated species that exhibits a wide diversity of varieties (Lebeda *et al.*, 2007; Mou, 2008). The present study was carried out on the cultivated part of the lettuce collection of the CGN that at the start of the redundancy analysis in 2007 consisted of 2571 accessions, of which 1540 belonged to cultivated lettuce and the remaining 1031 to crop wild relatives. CGN's lettuce collection is fairly

well described with passport, characterization and evaluation data (Boukema *et al.*, 1990). The majority of cultivated accessions have been classified according to the generally recognized crop types butterhead, crisphead, cutting, cos, stalk, oilseed and Latin lettuce (Lebeda *et al.*, 2007). During field trials organized by CGN for characterization purposes, cultivated accessions have also been classified according to cultivar type, which represent cultivars with similar morphological characteristics. Cultivar types have been assigned to accessions in cooperation with the representative of the Netherlands Inspection Service for Horticulture responsible for lettuce variety registration, and with Dutch lettuce breeders. Currently, the CGN lettuce collection comprises 452 different cultivar types.

Redundancy analysis

For the redundancy analysis, a strategy was followed that is generally applicable and that is based on the confirmation of duplication by two independent data sources (Fig. 1). The procedure starts with the pairwise comparison of accessions, in which the passport data are used as a first data source to identify or reject potential duplicates. In case of potential duplication, a second data source is used to confirm or reject the duplication status of accessions. In case of rejection, a third data source is used to validate the potential duplication status. Finally, decisions about the removal of redundancies are made when the duplication status of accessions was shown by two data sources. In the present study, molecular data were used as second data source and results from a morphological field trial as third data source to validate potential duplicates based on passport data. However, passport data may be incomplete or incorrect, which may leave true duplicates undetected. Therefore, the molecular data were also used to initially identify potential duplicates, after which the morphological field trial was used for confirmation of the redundancy status (Fig. 1).

In the present study, a few refinements to the strategy outlined in Fig. 1 were made. First, in case potential duplicates based on passport data were rejected by the

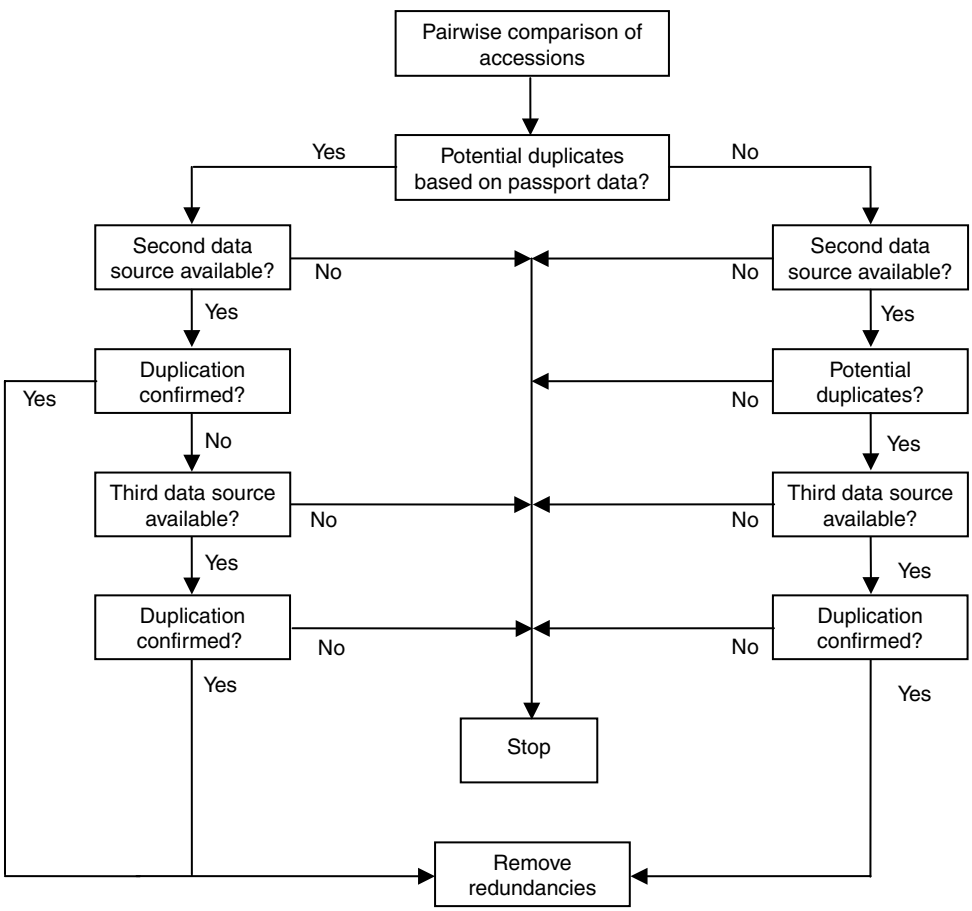

Fig. 1. Flow chart of the methodology used in the redundancy analysis to identify and validate potential duplicate accessions.

molecular data, accessions were only included in the experimental field trial when the initial duplication status had not been based on the matching of widely used accession names, such as for example 'Balady' (= landrace), 'Alface' (= lettuce), 'Outdoor lettuce' or 'Greenhouse lettuce' because of the high probability of false positives in such cases. Second, potential duplication groups based on passport data that were rejected by the molecular data and confirmed by the morphological field trial were considered for removal of redundancies only when the number of molecular differences between accessions did not exceed those within accessions, because such cases may represent true genetic differences instead of one accession being a subset of the other, as for example in the case of contamination. Third, potential duplicates based on the molecular data were included in the morphological field trial only when one or both of the compared accessions had a widely used name or the name was missing, and their cultivar types were indifferent. Moreover, molecularly similar accessions that differed in registered crop type were also included to check for potential documentation errors. Fourth, before the removal of redundancies, all remaining duplication groups were re-examined by a crop expert as a final check on false positives leading to unjustified removal of accessions. Apart from the identification of candidate accessions for removal from the collection, our approach allowed for review of the passport data where appropriate in case of identified errors.

Molecular data

Within the framework of the European Union (EU) project 'Molecular markers for genebanks', CGN's entire lettuce collection, which at the time of tissue sampling in 1997 consisted of 2323 accessions, was characterized with AFLPs and microsatellites. For cultivated lettuce, comprising 1419 accessions, two randomly chosen individuals per accession were screened for three AFLP primer combinations according to the methods of Vos *et al.* (1995) and for ten microsatellites (Van de Wiel *et al.*, 1999) according to the methods described by Van Treuren *et al.* (2008). The molecular analyses in cultivated lettuce yielded 160 polymorphic AFLP fragments and 150 different microsatellite alleles. More detailed information about the molecular analyses can be found in van Hintum (2003) and Jansen *et al.* (2006).

Morphological characterization

Potential duplicates, for which the analysis of passport and molecular data revealed incongruent results, were planted side by side in an experimental field trial in 2008 and were compared morphologically. A total number of 283 accessions were involved in the experiment, including 23 standard references. The field trial was conducted at the experimental field station 'Nergena' located near Wageningen. Sowing was performed on the 14th of March 2008 and 30 plants per accession were transferred to the field on the 23rd of April 2008. Plants were arranged in two rows per accession with 30 cm inter-plant distance. Accessions were scored for 24 characters, largely resembling the descriptor list presented by Boukema *et al.* (1990), and including all ECPGR minimum descriptors developed for cultivated lettuce (Thomas *et al.*, 2005). Accessions from potential duplicate groups were independently investigated for morphological similarity by CGN, the former curator of CGNs lettuce collection, and the former representative of the Netherlands Inspection Service for Horticulture responsible for lettuce variety registration. The duplicate status of accessions was rejected in case of lacking consensus between the evaluators.

Data analyses

Accessions with identical or similar names were considered potential duplicates based on passport data and were identified by sorting accessions in alphabetic order of their principal name. In addition, matching Plant Introduction numbers and published synonyms (Rodenburg, 1960) were used to find potential duplicates based on passport data.

To evaluate the distribution of molecular variation within the collection, the number of AFLP and simple sequence repeat (SSR) scores that were different between plants was quantified at different levels of organization, i.e. (1) between accessions of different crop type; (2) between accessions of the same crop type; (3) between accessions of the same cultivar type; (4) between accessions with similar names; (5) between plants of the same accession. For each of the five organizational levels, 100,000 pairwise comparisons of randomly chosen samples were made and frequency distributions were constructed of the number of AFLP and SSR differences, respectively. For each organizational level, a binomial function was fitted to the observed distribution. This was performed by simulating *P*-values from 0.000 to 1.000 in order to find the value that corresponded with the smallest sum of squared differences between expected and observed frequencies of molecular differences. These analyses were carried out using a programme written in Turbo Pascal.

To quantify the effect of removing redundant accessions on collection diversity, variability prior to

rationalization was compared to that of the reduced collection. For this purpose, four parameters were used, i.e. the number of polymorphic AFLP fragments, the number of microsatellite alleles, the number of character states for 26 characterization and evaluation traits recorded in previous experiments, and the number of resistance profiles against 28 pathotypes of downy mildew (*Bremia lactucae*) that had been determined by Dutch breeding companies during 2000–2004.

Results

Distribution of molecular diversity

The distribution of the number of AFLP and microsatellite differences between plants at different organizational levels within the investigated collection of cultivated lettuce is presented in Fig. 2. As had been expected, the level of similarity increased with lower organizational levels. A major shift in distribution shape towards higher similarity was observed between samples belonging to the same crop type and samples belonging to the same cultivar type, indicating that cultivar type is a useful concept for structuring a lettuce collection. The distribution of molecular differences between samples from accessions with similar names (potential duplicates based on passport data) showed a high frequency of molecularly identical samples and samples with very low numbers of marker differences. However, for part of the pairwise compared accessions, a large number of molecular differences were observed at this organizational level, indicating that duplicate groups identified based on passport data may contain false positives. Large differences could even be observed between samples from the same accession. Although the majority of samples appeared identical at this organizational level, marker differences were observed for more than 30% of the comparisons for AFLPs and for more than 20% of the comparisons for microsatellites.

The occurrence of intra-accession variation demanded that the redundancy analysis should build on a predetermined minimum level of inter-accession variation. For this purpose, a binomial function was fitted to the observed distributions (Table 1). In line with the distributions observed in Fig. 2, the probability of detecting a marker difference decreased with lower organizational levels. These probabilities were higher for microsatellites than for AFLPs due to the higher degree of polymorphism at individual loci. For example, the probability of detecting a marker difference between samples from the same accession is ten times as high for microsatellites compared to AFLPs. However, lower probabilities of detecting differences at individual loci are more than compensated by the higher multiplex ratio of AFLPs, as can be seen in the expected probabilities of finding a specified number of marker differences between samples from the same accession (Table 1). These data showed that the probability of finding a single marker difference within an accession was relatively high, i.e. 0.240 for AFLPs and 0.167 for microsatellites. For microsatellites, the probability of detecting two differences was only 1.5%, while for AFLPs still a value of 4.0% was estimated, nearly three times as high compared to microsatellites. Based on these findings, a maximum of two AFLP differences and a single microsatellite difference were regarded acceptable levels of inter-accession variation in the redundancy analysis.

Redundancy analysis

The redundancy analysis in the present study revealed a total number of 198 redundancies, corresponding with 12.9% of the collection of cultivated lettuce in 2007. The majority of redundancies carried identical or synonymous accession names, while in one case an accession carrying the general name 'Mestnyi', which means 'local', could be matched with a known cultivar. The value of using the molecular data to identify potential duplicates was demonstrated by the identification of four redundancies carrying an incorrect accession name and 51 redundancies for which the accession name was absent (Table 2(a)). Concerning the population type of the identified redundancies, the majority of accessions were registered as cultivars. In addition, in nine cases, accessions registered as landraces were considered duplicates of known cultivars (Table 2(b)). For example, the earlier mentioned accession Mestnyi (CGN04661), registered as a landrace collected in Armenia, could not be distinguished molecularly or phenotypically from accession CGN05237 representing the well known old French cultivar 'Romaine Verte Maraîchère'. This indicated that assigning the population type landrace to collected materials should be carefully considered since such accessions might well represent modern varieties.

For seven redundancies, clear off-types observed during the morphological field trial resolved the initially contradicting results between passport and molecular data. These off-types were most probably caused by seed contamination, resulting in 'compound duplication' in terms of van Hintum and Knüpffer (1995). For example, one of the two plants sampled from accession CGN04718 with principal name 'Passion Verte d'Hiver de Châteaurenard' could not be distinguished molecularly from the plants sampled from accession CGN04717 registered under identical name, whereas the other

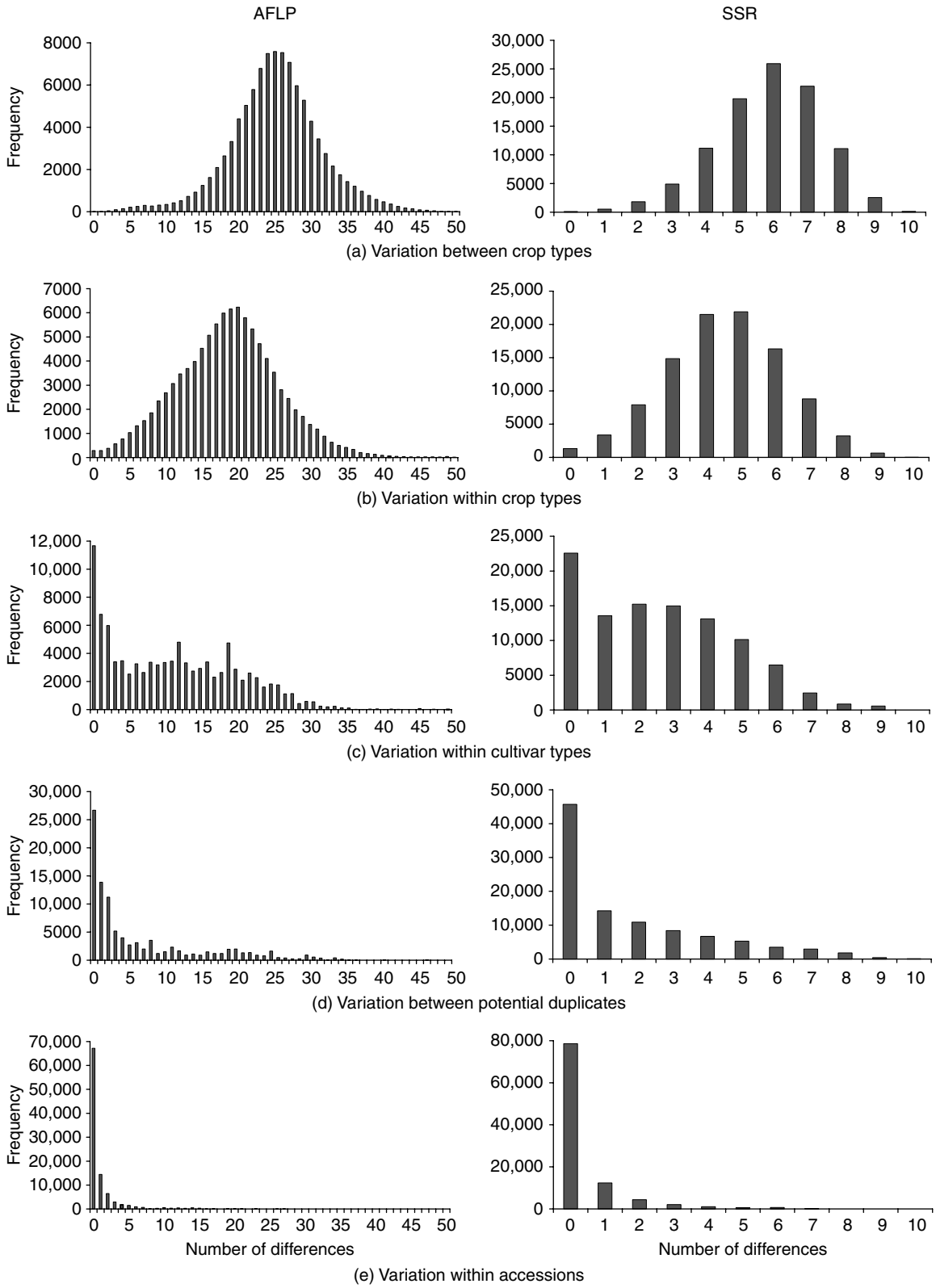

Fig. 2. Distribution of the number of AFLP and microsatellite (SSR) differences between plants at different organizational levels, i.e. (a) between accessions of different crop type, (b) between accessions of the same crop type, (c) between accessions of the same cultivar type, (d) between accessions with similar names and (e) between plants of the same accession. Each distribution is based on 100,000 pairwise comparisons of randomly chosen samples.

Table 1. Single-locus probability of detecting a difference between samples at five organizational levels for the investigated amplified fragment length polymorphism (AFLP) and simple sequence repeat (SSR) markers, respectively

	AFLP	SSR
Different crop type	0.152 (0.001)	0.596 (0.000)
Same crop type	0.117 (0.009)	0.460 (0.001)
Same cultivar type	0.101 (0.056)	0.259 (0.063)
Same accession name	0.008 (0.070)	0.062 (0.061)
Same accession	0.002 (0.013)	0.020 (0.004)
$P(X = 0)$	0.716	0.817
$P(X = 1)$	0.240	0.167
$P(X = 2)$	0.040	0.015
$P(X = 3)$	0.004	0.001

Probabilities were estimated by fitting a binomial function to the observed distribution of the number of marker differences. The sum of squared differences between observed and expected frequencies is given between brackets. For samples from the same accession, the probability of observing x differences among all marker loci is denoted by $P(X = x)$.

sample from CGN04718 showed 20 AFLP and 6 SSR differences. This deviating plant could be matched molecularly with the samples from accession CGN05783 with the principal name 'Luniek', indicating that CGN04718 had most likely been contaminated with seeds from CGN05783.

Rationalization approach

When duplication groups have been validated and redundancies identified, it has to be decided which accession from each duplication group to maintain in the collection and which accessions to discard, or alternatively to merge the accessions involved. Considering the autogamous nature of lettuce, it was decided in the present study never to merge accessions, but rather to choose the most authentic accession for maintenance. Trueness to type in case of cultivars was used as the primary criterion to decide which accession to maintain. For this purpose, previous experimental data and observations made during the new characterization trial were used, including data on heterogeneity of the accessions and the occurrence of off-types. Also completeness of the passport data was taken into account, in particular with respect to documentation on the variety name and on the origin of the accession. As a secondary criterion, the number of collected characterization and evaluation data was used, the latter including resistance profiles against 28 pathotypes of downy mildew, determined previously by Dutch breeding companies for part of the collection. Management considerations formed the tertiary criterion, including available seed quantities, date of last regeneration and last recorded germination percentage. In total, 198 identified redundancies were removed from the active and base collection.

Table 2. Number of redundancies identified in the present study, classified according to the involved status of the accession name (a) and to the involved population type (b)

	Number
(a) Accession name	
Identical	88
Synonymous	54
General	1
Incorrect	4
Missing	51
Total	198
(b) Population type	
Cultivars	144
Landraces	15
Breeding materials	2
Unknown	8
Cultivar/landrace	9
Cultivar/unknown	16
Landrace/unknown	4
Total	198

To evaluate the effect of the rationalization process on collection diversity, the level of variation was quantified before and after reduction of the collection (Table 3). Only a single AFLP polymorphism and a single microsatellite allele were lost, while all character states for characterization and evaluation traits were maintained. The single lost resistance profile against downy mildew involved an intermediate resistance type for pathotype BL19, but 83 accessions highly resistant against this pathotype still formed part of the remaining collection. Therefore, the effects of the rationalization process on collection diversity were considered negligible.

Discussion

Application of molecular data in redundancy studies

Molecular markers are increasingly being used in germplasm characterization, such as for the purpose of identification and validation of duplicate accessions. However, the maximum extent of inter-accession variation that is still considered acceptable to identify accessions as duplicates is still a major issue (Van Treuren *et al.*, 2001; Lund *et al.*, 2003). In the present study, a maximum of two AFLP differences, out of 160 potential polymorphisms, and a single microsatellite difference, out of ten potentially different loci, were regarded acceptable levels based on the observed distribution of the number of marker differences between and within

Table 3. Effect of removing redundant accessions on the diversity of the investigated collection of cultivated lettuce

	Before rationalization	After rationalization
Number of accessions	1540	1342
Number of polymorphic AFLP fragments	160	159
Number of microsatellite alleles	150	149
Number of character states for characterization and evaluation traits	108	108
Number of resistance profiles against downy mildew pathotypes	224	223

accessions. However, these threshold values should not be treated as general guidelines because they are dependent, among others, on the type and number of markers used and the variability of the material studied.

Molecular differences between samples were calculated as the absolute number of differences at AFLP and microsatellite loci. The use of this parameter may cause questionable decisions on similarity for datasets containing large numbers of missing values, in which case the proportion of differences would be more appropriate. In the present study, however, the percentage missing values was only 3.2% on average per marker locus. Moreover, genetic similarity of samples was inferred only if both the AFLP and microsatellite differences did not exceed the threshold values. In addition, decisions on redundancy were not taken solely on the basis of molecular data, but had to be in line with those from an independent second data source, i.e. passport or morphological data. It is therefore considered unlikely that missing values within the molecular dataset significantly affected the obtained results.

The threshold values for genetic similarity were determined based on the entire dataset and applied to all sample comparisons, irrespective of the population type involved. If a collection consists of material that strongly varies in intra-accession variation, as for example could be the case between landraces and modern cultivars, differential threshold values are more appropriate. The landraces investigated in the present study showed less than three AFLP differences between samples from the same accession in 75% of the comparisons, while less than two differences for microsatellites were found in 80%. For the total dataset, these figures were 88 and 91%, respectively. Given these limited differences and considering that 84.5% of the samples with known population type consisted of cultivars and breeding materials, while only 15.5% were registered as landraces, severe inflation by the landraces of the threshold values as used for cultivar comparisons was considered unlikely. However, the use of uniform threshold values may have caused slight underestimation of the level of redundancy among landraces.

Molecular data appeared useful in the present study, not only to identify and validate duplicates, but also to monitor maintenance of the genetic integrity of accessions. The genetic integrity of accessions may be compromised by a variety of factors, including gene flow among accessions. Molecular matching of off-type plants with plants from other accessions suggested that seed contamination in CGNs lettuce collection does occur. Although lettuce is predominantly self-fertilizing, the observation of highly heterozygous plants suggested that gene flow by cross-pollination also takes place. An example is presented for accession CGN16238 of which one plant was found to be heterozygous at six out of the ten investigated microsatellite loci, while the second examined plant was homozygous at all loci for alleles observed in plant 1 (Table 4). Such high levels of heterozygosity are unexpected under strict self-fertilization, but rather suggest contamination by cross-pollination. Therefore, it was investigated whether a potential pollination source could be identified among the examined lettuce accessions using a paternity exclusion analysis (Van Treuren *et al.*, 2006). This analysis revealed two (duplicate) potential pollination sources, assuming that CGN10956 and CGN16236 contain either allele 257 or a null allele at locus D109 (Table 4). Examination of the regeneration history of the involved accessions showed that CGN16236 and CGN16238 have been rejuvenated adjacently in 1993. Low frequency off-types may have only temporary effects on the genetic integrity of accessions because of their low probability of being included in the next seed multiplication trial and the probability that they are recognized and removed during regeneration.

Rationalization strategy

In rationalizing collections, the risk of removing false positives from a collection should be minimized as much as possible in order to maintain collection diversity. In the present study, effects of the rationalization process on collection diversity as measured by AFLP,

Table 4. Microsatellite scores of the two molecularly investigated plants of accession CGN16238 and those of the potential pollination sources of highly heterozygous plant 2 of CGN16238, remaining after a paternity exclusion analysis

	CGN16238		CGN10956		CGN16236	
SSR	Plant 1	Plant 2	Plant 1	Plant 2	Plant 1	Plant 2
A001	181	150:181	150	150	150	150
A004	180	180	x	x	180	x
B101	187	187	187	187	187	187
B104	169	169:187	187	187	187	187
D103	266	266:269	269	269	269	269
D106	188	167:188	167	167	167	167
D108	154	118:154	118	118	118	118
D109	257	257	300	x	291	x
E003	206	197:206	197	197	197	197
E011	257	257	257	257	257	257

Missing values are denoted by 'x'.

microsatellites, character states for characterization and evaluation traits, and resistance profiles against downy mildew pathotypes appeared negligible, indicating that a sufficiently cautious approach was followed by focusing on the most obvious redundancies. This cautious approach may have left true duplicates undetected, suggesting that the redundancy level observed may have been underestimated. Apart from the undetected duplicates, also overrepresentation of certain groups as compared to other groups of accessions could be considered redundancy. This type of overrepresentation may easily occur in lettuce as new varieties may be developed by introgression of a limited number of desirable traits into existing cultivars. As a result, such new varieties display a high genetic similarity with their predecessor (Van Treuren *et al.*, 2008). This type of redundancy might be reduced by structuring of the collection into hierarchical units, after which overrepresented units can be identified and its composition optimized by selecting and maintaining those accessions that collectively contain the highest level of genetic diversity (Van Treuren *et al.*, 2009). Such selections are unlikely to include highly similar accessions, such as undetected true duplicates and essentially derived varieties.

Different strategies can be followed to remove redundancies from a collection. Firstly, redundancies may be lumped into a single accession, which maintains the total genetic variation, at least in the first generation (van Hintum *et al.*, 1996). This can be considered a positive effect because it avoids loss of alleles. It may, however, also cause maintenance of contaminants. Secondly, as an alternative strategy, redundancies may be discarded from a collection, which has the advantage that allele frequencies are not influenced in the accession that is maintained. Lumping of accessions is considered more appropriate to outbreeding species, while removal of accessions from a collection is more appropriate for self-fertilizing crops, such as lettuce (Sackville Hamilton *et al.*, 2002).

The redundant lettuce accessions identified in the present study were 'archived', which means that the material remains stored, but does no longer form an integral part of the active, nor of the base collection. Compared to discarding samples, the advantage of the archive status is that the material is still present and its status can be reconsidered if necessary as long as its viability lasts (Engels and Visser, 2003).

In case, duplications can be traced back to errors during genebank management procedures, such as mislabelling or contamination of accessions; restoration of the authentic accession can be considered. This may be achieved by requesting new seed lots in case the material in question is still available elsewhere, or by returning to older genebank seed lots provided that such samples correctly represent the intended accession. An example in the present study was given by CGN05024, registered as an unnamed cos accession of the cultivar type Romaine Verte Maraîchère, which could molecularly and phenotypically not be distinguished from butterhead accessions CGN05875, CGN05973 and CGN11431 all belonging to the cultivar type 'Passe-Partout'. Because also the seed colour of the recent samples of CGN05024 appeared incorrect, an older seed lot displaying the correct seed colour was considered to represent the right accession, and this seed lot was regenerated and verified in order to restore the accession.

Economic aspects of redundancy studies

The redundancy analysis in the present study was made possible by the availability of molecular data.

However, such comprehensive molecular datasets are lacking for most genebank collections. It cannot be recommended to invest substantially in the molecular screening of entire collections solely for the purpose of reducing redundancy. A possibility to reduce necessary investments is to restrict the molecular screening to the validation of potential duplicate groups constituted on the basis of available data sources, such as passport information. Another option is to screen small numbers of plants per accession from certain parts of a collection where redundancy is expected to be highest, after which accessions from resulting duplication groups can be investigated in more detail with larger sample sizes. This approach, which reduces the probability of false positives at the expense of leaving true duplicates undetected, has been followed in part of CGNs potato collection (McGregor *et al.*, 2002; Van Treuren *et al.*, 2004).

Total regeneration costs at CGN are approximately 330 € per lettuce accession, which means that by removing the 198 identified redundancies from the collection, 65,340 € are saved per regeneration. A manifold of these savings would have been necessary to obtain the molecular dataset as used in the present study. Whether the benefits of a reduced collection outweigh the necessary investments to identify redundancies will depend on the number and type of markers studied, the regeneration costs of the crop concerned, the final level of redundancy identified and the timeframe within which return of investments is expected (Van Treuren *et al.*, 2001, 2004; Van Treuren and Bas, 2008). Many of the commonly used molecular techniques, such as AFLP and microsatellites, are still costly to apply. However, high throughput sequencing and single nucleotide polymorphism genotyping using microarray techniques continue to drop in price and can be expected to become within reach of genebanks in the near future.

Acknowledgements

The authors would like to thank Nico van Marrewijk for the validation of potential duplicate accessions during the morphological field trial, which was partly subsidized by the EU GENRES project 'Leafy vegetables germplasm, stimulating use' (AGRI-2006-0262). The molecular data were generated within the framework of the EU project 'Molecular markers for genebanks: Application of marker technology for the improvement of *ex situ* germplasm conservation methodology' (Bio4 CT 9 PL96.2062). The project was also partly financed by the Dutch Ministry of Agriculture, Nature and Food Quality. We are also grateful to Bert Visser and two anonymous reviewers for their comments on an earlier version of the paper.

References

Boukema IW, Hazekamp Th and van Hintum ThJL (1990) *The CGN Collection Reviews: The CGN Lettuce Collection*. Wageningen: Centre for Genetic Resources, The Netherlands (CGN).

ECPGR (2009) *A Strategic Framework for the Implementation of a European Genebank Integrated System (AEGIS). A Policy Guide. European Cooperative Programme for Plant Genetic Resources (ECPGR)*. Rome: Bioversity International.

Engels JMM and Visser L (2003) *A Guide to Effective Management of Germplasm Collections. IPGRI Handbook for Genebanks No. 6*. Rome: International Plant Genetic Resources Institute.

FAO (2009) *The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture*. Rome: Food and Agriculture Organization of the United Nations.

Greene SL and Pederson GA (1996) Eliminating duplicates in germplasm collections: a white clover example. *Crop Science* 36: 1398–1400.

Jansen J, Verbakel H, Peleman J and van Hintum ThJL (2006) A note on the measurement of genetic diversity within genebank accessions of lettuce (*Lactuca sativa* L.) using AFLP markers. *Theoretical and Applied Genetics* 112: 554–561.

Lebeda A, Ryder EJ, Grube R, Doležalová I and Krístková E (2007) Lettuce (*Asteraceae*; *Lactuca* spp.). In: Singh RJ (ed.) *Genetic Resources, Chromosome Engineering, and Crop Improvement, Vegetable Crops*. Vol. 3. Boca Raton: CRC Press, pp. 377–472.

Lund B, Ortiz R, Skovgaard IM, Waugh R and Anderson SB (2003) Analysis of potential duplicates in barley genebank collections using re-sampling of microsatellite data. *Theoretical and Applied Genetics* 106: 1129–1138.

McGregor CE, van Treuren R, Hoekstra R and van Hintum ThJL (2002) Analysis of the wild potato germplasm of the series *Acaulia* with AFLPs: implications for *ex situ* conservation. *Theoretical and Applied Genetics* 104: 146–156.

Mou B (2008) Lettuce. In: Prohens J and Nuez F (eds) *Handbook of Plant Breeding, Vegetables I: Asteraceae, Brassicaceae, Chenopodiaceae, and Cucurbitaceae*. Vol. I. New York: Springer, pp. 75–116.

Plucknett DL, Smith HJH, Williams JT and Anishetty NM (1987) *Gene Banks and the World's Food*. Princeton: Princeton University Press.

Rodenburg CM (1960) *Varieties of Lettuce. An International Monograph*. Zwolle: N.V. Uitgevers-Maatschappij W.E.J. Tjeenk Willink.

Sackville Hamilton NR, Engels JMM, van Hintum ThJL, Koo B and Smale M (2002) *Accession Management. Combining or Splitting Accessions as a Tool to Improve Germplasm Management Efficiency. IPGRI Technical Bulletin No. 5*. Rome: International Plant Genetic Resources Institute.

Spooner D, van Treuren R and de Vicente MC (2005) *Molecular Markers for Genebank Management. IPGRI Technical No. 10*. Rome: International Plant Genetic Resources Institute.

Thomas G, Astley D, Boukema IW, Daunay MC, Del Greco A, Díez MJ, van Dooijeweert W, Keller J, Kotlińska T,

Lebeda A, Lipman E, Maggioni L and Rosa E (2005) *Report of a Vegetables Network. Joint Meeting with an* ad hoc *Group on Leafy Vegetables*. Skierniewice, Poland, 22–24 May 2003. Rome: International Plant Genetic Resources Institute.

Van de Wiel CCM, Arens P and Vosman B (1999) Microsatellite retrieval in lettuce (*Lactuca sativa* L.). *Genome* 42: 139–149.

van Hintum ThJL (2000) Duplication within and between germplasm collections. III. A quantitative model. *Genetic Resources and Crop Evolution* 47: 507–513.

van Hintum ThJL (2003) Molecular characterization of a lettuce germplasm collection. In: van Hintum ThJL, Lebeda A, Pink D and Schut JW (eds) *Eucarpia Leafy Vegetables 2003, Proceedings of the Eucarpia Meeting on Leafy Vegetables, Noordwijkerhout, The Netherlands, 19–21 March 2003*. Wageningen: Centre for Genetic Resources, The Netherlands (CGN), pp. 99–104.

van Hintum ThJL and Knüpffer H (1995) Duplication within and between germplasm collections. I. Identification of duplication on the basis of passport data. *Genetic Resources and Crop Evolution* 42: 127–133.

van Hintum ThJL, Boukema IW and Visser DL (1996) Reduction of duplication in a *Brassica oleracea* germplasm collection. *Genetic Resources and Crop Evolution* 43: 343–349.

Van Treuren R and Bas N (2008) Perennial kales: collection rationalization and genetic relatedness to other *Brassica oleracea* crop types. *Genetic Resources and Crop Evolution* 55: 203–210.

Van Treuren R and van Hintum ThJL (2001) Identification of intra-accession genetic diversity in selfing crops using AFLP markers: implications for collection management. *Genetic Resources and Crop Evolution* 48: 287–295.

Van Treuren R and van Hintum ThJL (2003) Marker-assisted reduction of redundancy in germplasm collections: genetic and economic aspects. *Acta Horticulturae (ISHS)* 623: 139–149.

Van Treuren R, van Soest LJM and van Hintum ThJL (2001) Marker-assisted rationalisation of genetic resources collections: a case study in flax using AFLPs. *Theoretical and Applied Genetics* 103: 144–152.

Van Treuren R, Magda A, Hoekstra R and van Hintum ThJL (2004) Genetic and economic aspects of marker-assisted reduction of redundancy from a wild potato germplasm collection. *Genetic Resources and Crop Evolution* 51: 277–290.

Van Treuren R, Goossens P and Ševčíková M (2006) Variation in effective pollination rates in relation to the spatial and temporal distribution of pollen release in rejuvenated perennial ryegrass. *Euphytica* 147: 367–382.

Van Treuren R, van Hintum ThJL and van de Wiel CCM (2008) Marker-assisted optimization of an expert-based strategy for the acquisition of modern lettuce varieties to improve a genebank collection. *Genetic Resources and Crop Evolution* 55: 319–330.

Van Treuren R, Engels JMM, Hoekstra R and van Hintum ThJL (2009) Optimization of the composition of crop collections for *ex situ* conservation. *Plant Genet Resources: Characterization and Utilization* 7: 185–193.

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