Plant Science 184 (2012) 54–62

Contents lists available at SciVerse ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Empirical evaluation of DArT, SNP, and SSR marker-systems for genotyping, clustering, and assigning sugar beet hybrid varieties into populations

Ivan Simko[a,*], Imad Eujayl[b], Theo J.L. van Hintum[c]

[a] United States Department of Agriculture, Agricultural Research Service, U.S. Agricultural Research Station, 1636 E. Alisal St., Salinas, CA 93905, USA
[b] United States Department of Agriculture, Agricultural Research Service, Northwest Irrigation and Soils Research Laboratory, 3793 N. 3600 E. Kimberly, ID 83341, USA
[c] Centre for Genetic Resources, The Netherlands (CGN), Wageningen University and Research Centre, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 22 October 2011
Received in revised form
11 December 2011
Accepted 13 December 2011
Available online 19 December 2011

Keywords:
Molecular markers
Genotypic diversity
Population structure
Assignment into a population
Data resolution
Sugar beet

ABSTRACT

Dominant and co-dominant molecular markers are routinely used in plant genetic research. In the present study we assessed the success-rate of three marker-systems for estimating genotypic diversity, clustering varieties into populations, and assigning a single variety into the expected population. A set of 54 diploid sugar beet (*Beta vulgaris* L. ssp. vulgaris) hybrid varieties from five seed companies was genotyped with 702 Diversity Array-Technology (DArT), 34 Single Nucleotide Polymorphisms (SNP), and 30 Simple Sequence Repeats (SSR) markers. Analysis of the population structure revealed three well-defined populations and clustering of varieties that generally correlates with their seed company origin. Two populations each contained varieties from two different seed companies indicating genetic similarity of this material. The third population was comprised only of varieties from a single seed company. Analysis of the SSR and SNP datasets indicates that some of the hybrid varieties likely have a common (or very closely related) parent.

Comparison of the three marker-systems revealed substantial differences in the number of loci needed for analyses. Varietal clustering required approximately 1.8–2 × more SSR, 3–4.5 × more SNP, and 4.8 × more DArT markers than were required for detection of genotypic diversity. When marker-systems were compared across different types of analyses per locus success-rate was the highest for the SSR and the lowest for the DArT markers. Generally, about 1.4–3 × more SNPs, and 4.9–13.3 × more DArTs then SSRs were needed to achieve the 100% success-rate. However, using only DArT markers with a high level of polymorphism decreased the number of DArT loci needed for analyses by 38–61%. Results from the present work provide a premise to selecting the type(s) and number of markers that are needed for genetic diversity analysis of sugar beet hybrid varieties.

Published by Elsevier Ireland Ltd.

1. Introduction

Detection of population structure, assignment of an individual into a population, and assessment of genetic variation in both domesticated and wild species are frequently used in plant breeding, germplasm classification in gene banks, investigation of evolutionary processes, and other research areas [1]. There are several types of molecular markers that are commonly used in plant genetic analyses such as assessment of population structure, linkage map construction, mapping alleles for desirable traits, marker-assisted selection, and fingerprinting of varieties. These markers are classified as either co-dominant or dominant, depending on their ability to distinguish allelic status of a heterozygote from a dominant homozygote. Results of modeling show that fewer co-dominant than dominant markers are needed to attain the same estimate of genetic diversity [2] and clustering of accessions [3]. To evaluate empirically the success-rate of different molecular marker-systems in detecting genetic diversity, resolving population structure, and assigning individuals into populations, we have examined diploid hybrid varieties from five sugar beet breeding companies that were genotyped with three molecular marker-systems namely; SSR, SNP, and DArT.

SSR (microsatellites) are short tandem repeats in DNA that are present in genomes of all analyzed eukaryotic organisms [4]. Because of their high reproducibility, multiallelism, and co-dominant inheritance, SSRs are frequently used in plant genetics [5].

Abbreviations: AUC, area under the curve; DArT, diversity array technology; DArT-HP, Diversity Array Technology markers with Highest Polymorphism; DR, data resolution statistics; EST, Expressed Sequence Tag; PCR, Polymerase Chain Reaction; PIC, polymorphism information content; SNP, Single Nucleotide Polymorphism; SSR, Simple Sequence Repeats (microsatellites).

* Corresponding author. Tel.: +1 831 755 2862; fax: +1 831 755 2814.
E-mail addresses: ivan.simko@ars.usda.gov (I. Simko), imad.eujayl@ars.usda.gov (I. Eujayl), theo.vanhintum@wur.nl (T.J.L. van Hintum).

0168-9452/$ – see front matter. Published by Elsevier Ireland Ltd.
doi:10.1016/j.plantsci.2011.12.009

SNP is a single-point mutation of DNA in which one nucleotide in a particular locus is substituted with another one. SNPs are often used in genome-wide association studies as they are highly abundant in genomes, amenable to high-throughput screening, co-dominant, and usually biallelic.

DArT is a genotyping system based on microarraying. The method detects several hundred of polymorphic fragments (DArT markers) in the genome in a single analysis [6]. Unlike SSR- and SNP-based markers, DArT are dominant markers scored as either present or absent, thus providing less genetic information for a given locus. Hurtado et al. [7] observed that 251 DArTs and 36 SSRs (approximately 7:1 ratio) generated broadly similar clustering patterns in 436 varieties of cassava; however, greater genetic differentiation was revealed with SSR markers.

The objectives of the present research were: (1) to compare genotypic diversity assessed by DArT, SSR, and SNP marker-systems; (2) to evaluate the success-rate of these marker-systems for clustering varieties into populations and (3) for assignment of a single variety into a population. This kind of empirical analysis should reveal the relative success-rate of different marker-systems in resolving population structure of cultivated sugar beet and for fingerprinting of varieties. New information about performance of DArT markers is particularly vital, because there are limited studies that compare this type of molecular marker with other marker-systems.

2. Materials and methods

2.1. Plant material

Seeds of 54 diploid hybrid varieties (Supplementary file 1) were obtained from five commercial sugar beet seed companies: 8 from American Crystal Hybrid Seeds Inc. (Eden Prairie, MN, USA), 15 from Betaseed Inc. (Shakopee, MN, USA), 10 from Holly Hybrids (Sheridan, WY, USA), 18 from Hilleshög (Longmont, CO, USA), and 3 from SeedEx (Fargo, ND, USA). Plants from all varieties were grown in a greenhouse and total genomic DNA was extracted from freeze-dried leaves of a randomly selected single plant using Qiagen DNeasy Kit (QIAGEN, Valencia, CA, USA). This single plant was considered a typical representative of the respective variety, provided that the varieties are hybrids developed by crossing two near homozygous lines. In reality, sugar beet hybrid varieties are not completely homozygous and contain a degree of heterozygosity. Genotyping with the three marker-systems was carried out on aliquots of DNA originating from the same extraction.

2.2. Molecular markers

2.2.1. SSR markers

Thirty SSR markers used for genotyping comprised 17 unlinked genomic SSRs with known map position [8] and 13 EST-SSRs developed from the sugar beet GenBank EST database at NCBI (http://www.ncbi.nlm.nih.gov/nucest?term=(sugar%20beet)%20 AND%20“Beta%20vulgaris”[porgn:__txid161934]) (Supplementary files 2 and 3). PCR was performed with M-13 tailed forward primers that were labeled with FAM, PET, or NED dyes (Applied Biosystems, Foster City, CA, USA). The PCR reaction mixture (20 μl) consisted of approximately 20 ng template DNA, 1× PCR buffer (Applied Biosystems, Foster City, CA, USA), 2.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.2 μM of forward labeled primer, 0.4 μM of M-13, 0.2 μM of reverse primer, and 0.5 U of AmpliTaq Gold (Applied Biosystems, Foster City, CA, USA). The thermocycling conditions included an initial denaturing period of 5 min at 95 °C, followed by 35 cycles of 95 °C for 50 s, annealing at 58 °C or 60 °C for 50 s, extension at 72 °C for 90 s, and a final extension period for 10 min at 72 °C. PCR products were electrophoresed using ABI 3100 Genetic Analyzer following the manufacturer’s protocol (Applied Biosystems, Foster City, CA, USA).

2.2.2. SNP markers

Thirty-four SNP markers (Supplementary file 2) with known genetic map positions [9] were used for genotyping, which was performed by TraitGenetics GmbH (Gatersleben, Germany). Genotyping was carried out using the 48plex SNPlex system (Applied Biosystems, Foster City, CA, USA) according to the manufacturer’s recommendations. SNPlex genotyping was based on an oligonucleotide ligation assay to differentiate between the two SNP alleles, followed by PCR using universal primers. PCR products were then hybridized to DNA probes carrying fluorescent dyes. Labeled fragments were detected with ABI 3730XL capillary sequencer (Applied Biosystems, Foster City, CA, USA).

2.2.3. DArT markers

Genotyping with DArT markers was performed by Diversity Array Technology Pty, Ltd. (Yarralumla, Australia). Marker development was based on the protocol of Jaccoud et al. [10]. The genome complexity reduction was carried out through digestion of total genomic DNA with the PstI/BstNII combination of restriction enzymes, ligation of enzyme adaptors, and amplification of adaptor-ligated fragments [6]. Four libraries were constructed using the representations produced by PstI/BstNII digestion. Seven hundred and two DArT markers that showed polymorphism on the set of 54 varieties were scored as present or absent (1/0) using bimodal distribution of relative signal intensity.

2.2.4. DArT-HP

Considering that SNP markers are used by TraitGenetics for commercial genotyping of sugar beet accessions, it is likely that these markers have been selected for high level of polymorphism during their development. Mapped genomic SSRs [8] may also have been subjected to a selection procedure leading to a higher marker polymorphism. Therefore, for comparison purposes, we performed similar selection on the bulk of DArT markers. One hundred markers with the highest level of polymorphism (DArT-HP) were selected from the total number (702) of scored DArT markers. Polymorphism of markers was estimated using polymorphism information content (PIC) formula:

$$\text{PIC} = 1 - \sum_{i=1}^{n} p_i{}^2$$

where n is the number of alleles and p_i is the frequency of the ith allele of the evaluated locus [11]. DArT-HP markers were used separately in all analyses to observe an effect of high polymorphism on performance of dominant markers.

2.3. Data analysis

2.3.1. Consistency of molecular marker datasets

Data resolution (DR) statistics were used to evaluate quality of marker datasets, particularly the internal consistency of the data [12]. For all marker-systems we used the simple matching distance averaged over all markers with at least two non-missing observations. DR values can be in the range from 0 to 1, higher values indicating higher internal consistency of the data. All data analyses were performed with tailor-made software programs written in visual basic for applications in a MS-Excel environment [12]. The number of replications was set to 10,000.

2.3.2. Estimating the optimal number of populations and assigning varieties into populations – analyses with the complete set of markers

To detect the population structure and to assign individuals into populations, 766 markers from the three marker-systems (30 SSRs, 34 SNPs, and 702 DArTs) were combined into a single set. Two computer programs were employed to carry out independent analyses of population structure: STRUCTURE 2.3 [13,14] and GenoDive v.2.0b20 [15]. For analysis with STRUCTURE recessive alleles (scored as 0) were indicated for DArT markers [13] as described in the documentation for STRUCTURE software, version 2.3. Five runs of STRUCTURE were carried out by setting the number of populations (K) from 1 to 10. For each run, the number of iterations and burn-in period iterations were both set to 100,000. The optimal number of populations was estimated using the *ad hoc* statistic (ΔK), which is based on the rate of change in the log probability of data between successive K values [16]. All analyses were performed with the model that assumes admixture. GenoDive was used to perform clustering analyses using k-means approach. The clustering was carried out using a matrix of Euclidian distances between varieties, based on the allele frequencies. Analyses were run for 1 to 10 clusters (K) and simulated annealing with 100,000 steps. The optimal number of clusters (populations) was determined using pseudo-F statistics [15,17]. Both STRUCTURE and GenoDive identified the same optimal number of populations (K) with the identical assignment of individuals into populations. The results obtained with 766 markers were assumed to reveal the true (expected) population structure and all analyses with a subset of these markers were compared to it.

2.3.3. Estimating the optimal number of populations and assigning varieties into populations – analyses with a subset of markers

The subsequent analyses of population assignment, clustering of varieties, and population structure were carried out with a subset of 5, 10, 15, 20, 25, 30, 34, 50, 75, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, and 702 randomly sampled or total number of DArT markers. Sampling of loci was the same for all marker-systems with the exception that the maximum number of analyzed loci was restricted to 100 for DArT-HP, 34 for SNP, and 30 for SSR due to limited number of available markers. In all analyses the expected assignment of varieties into populations was based on the STRUCTURE and k-means (GenoDive) clustering results with 766 markers described in Section 2.3.2. In a statistical phrasing, assigning an individual to some known clusters is a supervised clustering problem [3,18], while assigning individuals into clusters that are defined *a posteriori* is an unsupervised clustering problem [18,19]. To measure the similarity between the expected and estimated population assignment under supervised clustering we used index:

$$S = \frac{a}{n}$$

where a is the number of varieties placed into the correct populations and n is the total number of varieties. Because labeling of populations in STRUCTURE and GenoDive is arbitrary, we computed S for each of the $K!$ possible permutations of the population labels [20] and recorded the maximum S_{max} across permutations. The proportion of varieties correctly placed into the expected populations (S_{max}) was expressed as a percentage ($S_{max} \times 100$) and is referred to as 'success-rate' throughout the text. All analyses were carried out ten times and mean values were calculated for each combination of marker-system and the number of loci.

2.3.4. Population structure

Analysis of population structure was performed with STRUCTURE 2.3 software [13,14] as described in Sections 2.3.2 and 2.3.3. The number of populations was set to be equal to the expected number of populations determined previously with 766 markers. All analyses were performed with the model that assumes admixture. Assignment of varieties into populations was based on a highest membership probability criterion (q value) calculated by STRUCTURE. This analysis is called '*structure*' throughout the text.

2.3.5. Clustering of varieties

K-means clustering of varieties was carried out with GenoDive v.2.0b20 [15] as described in Sections 2.3.2 and 2.3.3; however, analyses were performed only for the expected number of clusters previously determined with 766 markers. This analysis is called '*clustering*' throughout the text.

2.3.6. Population assignment

Population assignment was performed with GenoDive v.2.0b20 [15]. This analysis assigns an individual accession into a population by calculating the likelihood that the variety's genotype is found in the population [21]. To avoid bias of assignment allele frequencies in each population were calculated without the targeted individual. Frequencies of alleles that were equal to zero were replaced with the frequency of 0.005 as recommended in the documentation for GenoDive. This analysis is called throughout the text as '*assignment*'.

2.3.7. Genotypic diversity

The software MultiLocus ver. 1.3b [22] was used to estimate a number of different genotypes that can be identified in a set of 54 varieties with a step-wise increasing number of marker-loci. This analysis shows if scoring more loci is likely to increase the number of identified genotypes, or whether one has reached a plateau. Loci were sampled at random from 1 to $m - 1$, where m is the total number of marker-loci for the particular marker-system. One thousand samplings were performed for each combination of marker-system and loci number. The number of different genotypes that were identified in the analysis was converted into the percent scale relative to the total number of varieties. For simplicity and consistency with other analyses, the percent of identified genotypes is also referred to as success-rate of analysis. All analyses with MultiLocus were performed ten times and mean values were calculated. This analysis is called '*genotyping*' throughout the text.

2.3.8. Combining success-rates and statistical analyses

Success-rates of *structure*, *clustering*, *assignment*, and *genotyping* were calculated for several subsets of randomly sampled markers. To combine success-rates from multiple subsets into a single overall score, we calculated area under the curve (AUC) through a simple midpoint (trapezoidal) rule:

$$\mathrm{AUC} = \sum_{i=1}^{n-1} \frac{a_i + a_{i+1}}{2} \times (m_{i+1} - m_i)$$

where a_i is a success-rate at the ith subset, m_i is the number of markers at the ith subset, and n is the total number of subsets. When two analyses are compared, a higher AUC value indicates a higher success-rate across n subsets of markers. One-way ANOVA (analysis of variance), two-way ANOVA, and Tukey–Kramer HSD for significance of differences between AUCs were calculated with JMP 6.0.3 (SAS Institute, Cary, NC, USA).

2.3.9. Estimating the number of marker-loci needed to achieve 100% success-rate

Not all marker-systems reached 100% success-rate when available loci were used for analyses. To estimate the number of marker-loci that would be needed to achieve 100% success-rate, we optimized parameters for the

$$y = a \times \left[1 - \left(1 + \left(\frac{x}{b}\right)^{c}\right)^{-d}\right]$$

function that approximates success-rate curves. In this formula x is the number of marker-loci, y is the success-rate, and a, b, c, and d are parameters that need to be optimized to achieve the best fit of the curve to the series of observed datapoints. Optimization of parameters was carried out with the program pro Fit 6.1.16 (QuantumSoft, Uetikon am See, Switzerland). The estimated number of loci was rounded to the nearest whole number.

2.3.10. *Testing the possibility of a common parent*

To explore the possibility of some hybrids being half-sibs, a chance of having a common parent was estimated for all pairs of varieties. This estimate was based on the assumption that any two varieties may have a common parent only if they have at least a single allele in common at all loci. All varieties were compared with each other and for every pair of varieties the number of marker-loci supporting the assumption of a common parent was counted. This value was divided by the total number of marker-loci comparisons between the two varieties, resulting in a frequency of support. The frequencies of support were calculated separately from the SSR and SNP datasets and multiplied, yielding a combined score:

$$FS = \frac{C_{SSR}}{N_{SSR}} \times \frac{C_{SNP}}{N_{SNP}}$$

where *FS* is the frequency of support for a pair of varieties; C_{SSR} and C_{SNP} is the number of SSR and SNP marker-loci in which the two varieties have at least one allele in common; and N_{SSR} and N_{SNP} is the total number of SSR and SNP marker-loci comparisons between the two varieties. To account for possible scoring errors, the cut-off for the frequency of support was chosen at 0.92, which corresponds to two marker-loci (either both SNP or both SSR, or one of each) not supporting the possibility of a common parent.

3. Results

3.1. *Marker polymorphism and data resolution*

Seven hundred and sixty-six polymorphic markers were used to genotype 54 diploid sugar beet varieties from five commercial seed companies. The total number of alleles per SSR locus ranged from 2 to 9, with the average of 4.3. All SNP loci showed biallelic status. Thirty SSR markers have the mean PIC of 0.59, 34 SNP markers have the mean PIC of 0.41, and 702 DArT markers have the mean PIC of 0.28. When markers were grouped into bins based on increasing level of polymorphism, the highest frequency of SSR, SNP, and DArT markers were found in bins 0.61–0.70, 0.41–0.50, and 0.21–0.30, respectively (Fig. 1). A group of 100 DArT markers with the highest level of polymorphism (DArT-HP) has the mean PIC of 0.47, and all markers are grouped in the 0.41–0.50 bin (data not shown).

The DR of the 30 SSRs is 0.454, of the 34 SNPs is 0.266, of the 702 DArTs is 0.830, and of the 100 DArT-HP is 0.782 (Fig. 2). Fitting the curve based on the following function:

$$\mathrm{DR} = \frac{n}{n + nc}$$

where *n* is the number of marker-loci, shows that 122 DArT marker-loci are needed to get the same DR as the 30 SSRs, and 53 DArTs are needed to get the level as the 34 SNPs. Approximately 142 DArT-HP loci are estimated to have the same DR value as 702 DArTs.

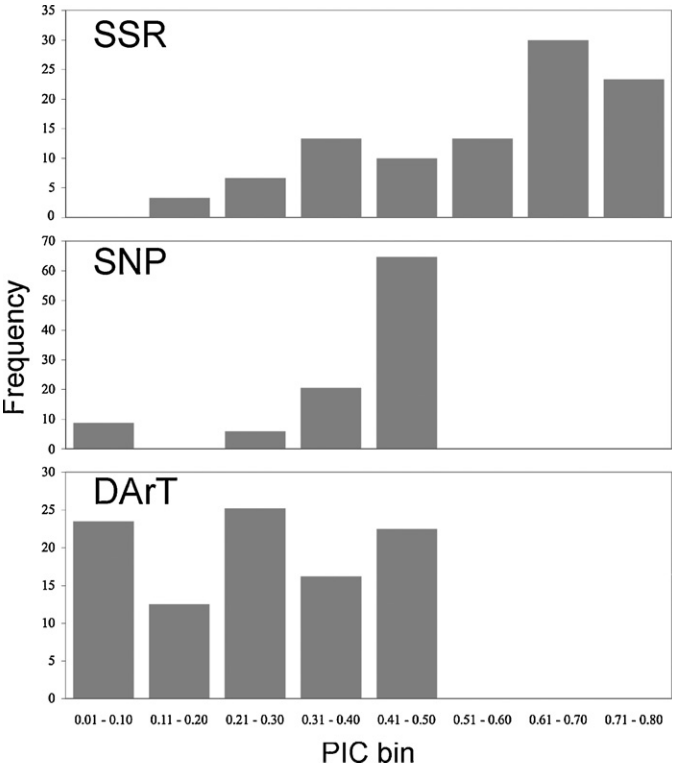

Fig. 1. Distribution of polymorphism information content (PIC) values for 30 SSR, 34 SNP, and 702 DArT markers used for genotyping 54 varieties of sugar beet. The mean PIC value for SSRs is 0.59, for SNPs is 0.41, and for DArTs is 0.28. Distribution is not shown for 100 DArT-HP markers (PIC = 0.47), because all markers are grouped in the 0.41–0.50 bin.

3.2. *Population structure*

Analysis of the 54 varieties with 766 markers suggests that the most likely number of populations is three (Fig. 3). Analysis with both STRUCTURE and GenoDive programs grouped varieties into the identical populations. The first population comprises 23 varieties from two breeding companies: 15 varieties from Betaseed and

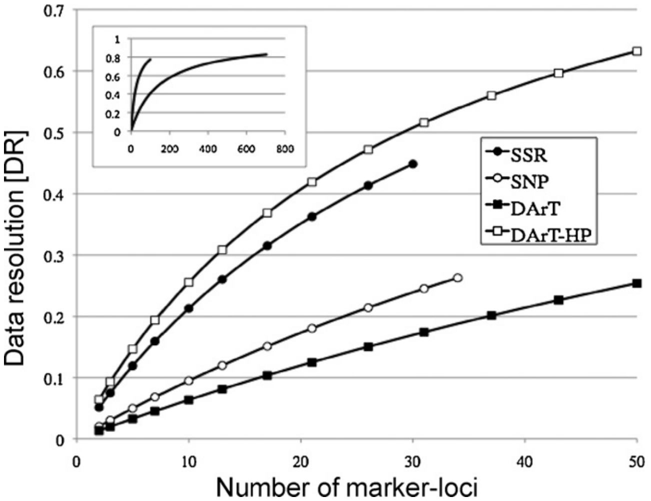

Fig. 2. Data resolution (DR) curve for the SSR (closed circle), SNP (open circle), DArT (closed square), and DArT-HP (open square) marker-systems. The maximum DR value of 30 SSRs, 34 SNPs, 702 DArTs, and 100 DArTs-HP are 0.454, 0.266, 0.830 and 0.782, respectively. For better resolution, data are shown only for 50 marker-loci. Small insert at the top left corner shows data resolution curves for all DArT (longer curve) and DArT-HP (shorter curve) marker-loci.

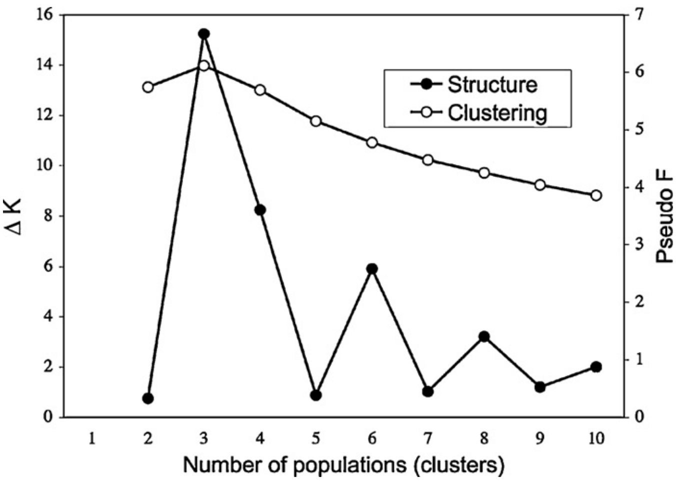

Fig. 3. The optimal number of populations (clusters) estimated from STRUCTURE (left axis – Δ*K* for structure), and GenoDive (right axis – pseudo-F for clustering). Analyses were performed on 54 sugar beet varieties genotyped with 702 DArT markers, 30 SSR markers, and 34 SNP markers. Peaks for both approaches indicate that the optimal number of populations is three (*K* = 3).

8 varieties from American Crystal Hybrid Seeds. The second population contains 21 varieties from three breeding companies: 18 from Hilleshög, 3 from SeedEx, and a single accession from Holly Hybrids. The third population includes only 9 varieties from Holly Hybrids (Fig. 4). In preliminary analyses we also tested the 'No Admixture' model in STRUCTURE (data not shown). Grouping of accessions was identical to the admixture model. The only noticeable difference was that under no admixture model all estimates of *q* reached the value of 1.

3.3. Effect of number of marker-loci

In order to determine the effect of increasing number of marker-loci on *structure*, *clustering*, *assignment*, and *genotyping*, we performed analyses with a variable number of loci. In each case marker-loci were randomly chosen and success-rate of their performance was obtained by comparing results to those achieved with a full set of 766 markers. In general, the success-rate of analyses increased with the growing number of marker-loci.

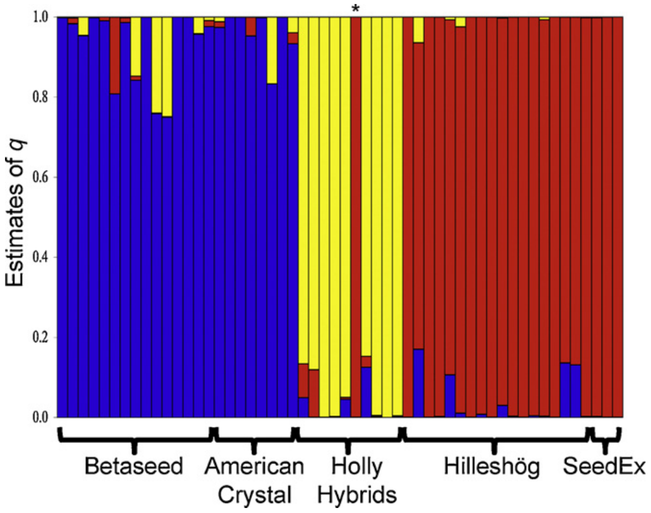

Fig. 4. Bar plot of population structure estimates for 54 sugar beet varieties from five seed companies. Population structure was assessed with combined 766 DArT, SSR, and SNP marker-loci. Each accession is represented by a single vertical bar broken into three colored segments, with lengths proportional to *q* of the three inferred populations (*K* = 3). The sum of *q* values for each bar is 1. Origin of the material is shown at the bottom. Asterisk on the top indicates a single accession (HH06) that does not group into the same population as all other varieties from the same seed company.

Table 1
Estimated number of loci needed to achieve 100% success-rate.

Markers	Genotyping	Assignment	Structure	Clustering
SSR	17	31*	30	34*
SNP	23	58*	69*	103*
DArT-HP	50	125*	157*	247*
DArT	83	250	400	400

* The value for analyses that did not reach 100% success-rate with the available number of loci was estimated by optimizing the $y = a \times \left[1 - \left(1 + \left(\frac{x}{b}\right)^{c}\right)^{-d}\right]$ function.

Seventeen SSR markers were enough to distinguish all varieties in genotypic diversity analysis (*genotyping*), while 30 markers were needed to correctly assign all varieties into populations (*structure*) (Table 1 and Fig. 5a). *Assignment* and *clustering* of varieties into populations reached 99.4% and 99.1% success-rate, respectively, when all 30 SSR markers were used for analysis with GenoDive. Overall, the AUC value (5–30 markers) for *clustering* (2149) was significantly smaller than those for the other types of analyses. The largest AUC value was reached for *genotyping* (2450), but it was not significantly different from the AUC value for *assignment* (2381).

Twenty-three SNP markers were needed to achieve 100% success-rate of *genotyping* (Table 1 and Fig. 5b). When the full set of 34 SNPs were used for analyses, success-rates of *assignment* and *clustering* with GenoDive were 94.4% and 92.5% respectively, while success-rate of *structure* was 90.7%. The significantly highest AUC value (5–34 markers) was calculated for *genotyping* (2763), while the smallest one was reached for *clustering* with GenoDive (2089).

To achieve 100% success-rate with DArT markers, 83 loci were needed for *genotyping*, 250 loci were needed for *assignment* of varieties, and 400 loci were needed for *structure* and *clustering* (Table 1 and Fig. 5c). The AUC values (5–702 markers) for the four types of analyses were divided into two groups with statistically different values. Higher AUC values were detected for *genotyping* (68,834) and *assignment* (68,009), while lower values were calculated for *structure* (64,683) and *clustering* (65,196).

Fifty DArT-HP markers were needed to reach 100% success-rate of *genotyping* (Table 1 and Fig. 5d). When 100 DArT-HP markers were used for analyses, the success-rate of *assignment* was 99.1%, *structure* 97.2%, and *clustering* 87.0%. The highest AUC value (5–100 markers) was reached for *genotyping* (9194), while the smallest one was reached for *clustering* (6989).

3.4. Comparison of marker-systems

Because the number of marker-loci from each marker-system (including DArT-HP sub-system) that were used for genotyping were unequal, a direct comparison of AUC values was not possible. To compare performance of marker-systems in different types of analyses, we calculated AUC values only for the identical subsets using 5, 10, 15, 20, 25, and 30 marker-loci (Table 2). In *genotyping* analyses the highest AUC values were observed for SSR (2450) and SNP (2363) markers. In contrast, the significantly smallest AUC value was calculated for DArT markers (1854). In both *assignment* and *structure* analyses the significantly highest AUC values were achieved for SSR markers (2381 and 2303, respectively), while the significantly lowest values were observed for DArT markers (1760 and 1468, respectively). In *clustering* analyses the results were similar to those seen at *assignment* and *structure*, with SSR reaching the highest AUC value (2149) and DArT reaching the smallest AUC

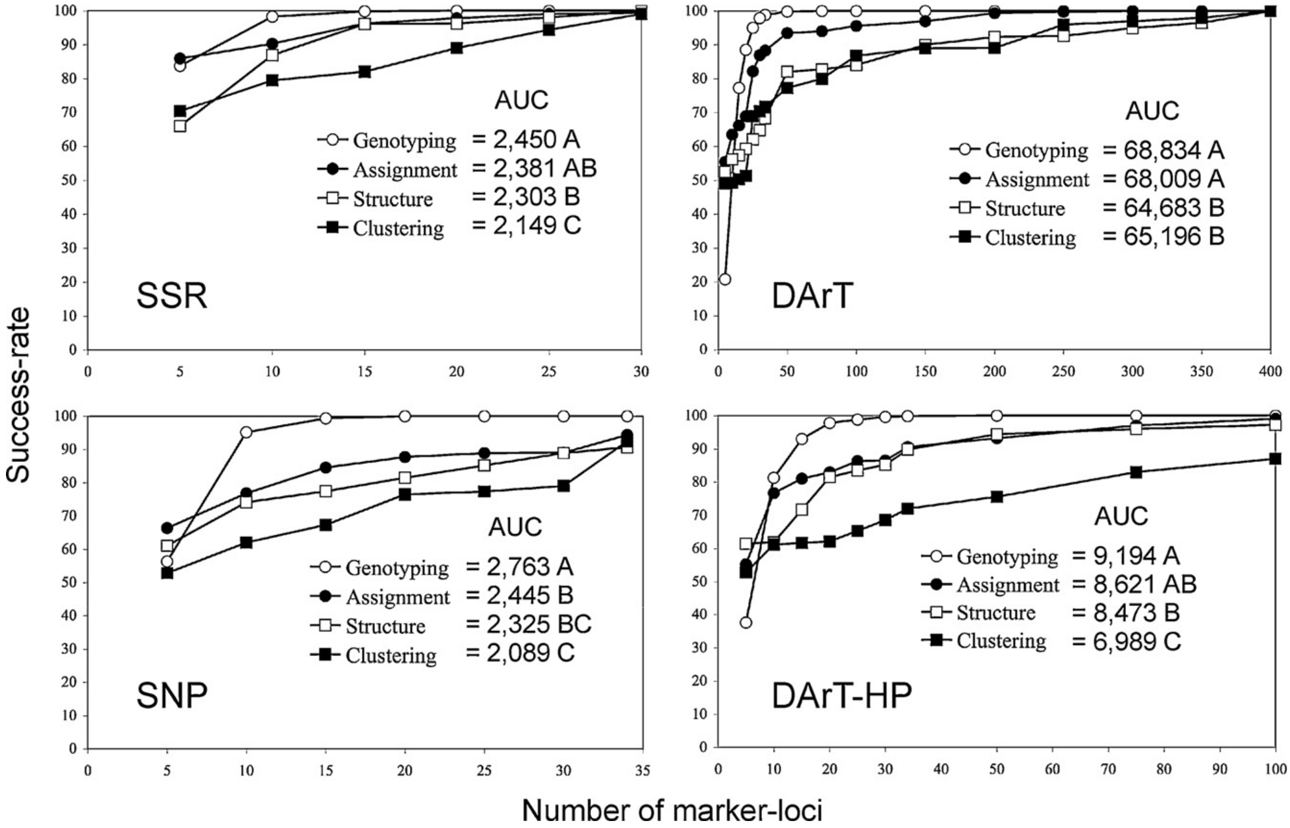

Fig. 5. Effect of the increasing number of marker-loci on the success-rate of *genotyping* (open circle), *assignment* (closed circle), *structure* (open square), and *clustering* (closed square). The AUC (Area Under the Curve) values are shown for SSR (a), SNP (b), DArT (c), and DArT-HP (d) marker-systems. AUC values within each type of marker-system followed by different letters are significantly different at $p \leq 0.05$.

Table 2
AUC values (5–30 markers) for four types of analyses when using different marker-systems.

Marker-system	Genotyping	Assignment	Structure	Clustering	Mean
SSR	2450 a	2381 a	2303 a	2149 a	2321 a
SNP	2363 a	2078 b	1966 b	1746 b	2038 b
DArT-HP	2197 b	1989 b	1859 b	1555 bc	1900 c
DArT	1854 c	1760 c	1468 c	1398 c	1620 d
Mean*	2216	2052	1899	1712	

Values within a column followed by different letters are significantly different at $p \leq 0.05$.
* All mean values within the row are significantly different at $p \leq 0.05$.

value (1398). However, AUC value for DArT markers was not statistically different from the one observed for DArT-HP markers (1555). When AUC values calculated for *genotyping*, *assignment*, *structure*, and *clustering* were combined; differences among the four marker-systems were significant. The mean AUC values decreased in the order SSR (2321), SNP (2038), DArT-HP (1900), and DArT (1620). Overall comparison of analyses revealed the highest AUC value for *genotyping* (2216), followed by *assignment* (2052) and *structure* (1899). The smallest value was observed for *clustering* (1712). All pair-wise differences were significant at *p*-value of at least 0.05.

4. Discussion

4.1. Marker polymorphism and data resolution

The level of marker polymorphism as estimated by average PIC value was the highest for SSR markers (0.59), followed by SNP markers (0.41), and the lowest for DArT markers (0.28). The results for SSR and DArT markers are similar to those observed on cassava, where PIC of 36 SSRs was 0.63 and PIC of 251 DArTs was 0.33 [7]. The higher PIC values for SSR markers than those for SNP or DArT markers are common, because multiallelic markers can reach higher PIC values. For example the maximum possible PIC for biallelic markers (such as SNP) is 0.5, while for tetraallelic markers it is 0.75.

All marker-systems show the expected shape of the data resolution curve with a relatively steep start previously observed in other datasets [12]. The starting point of the SSR curve (0.054) was higher than the starting points for the SNP (0.019) and DArT (0.013) curves (Fig. 2). This was expected because a multiallelic SSR contains more information than a single SNP or DArT marker-locus. Interestingly, the starting points of DR values for DArT-HP marker-system (0.064) were even higher than those for SSR. This may be due to the fact that the DR measures the correlation between similarity matrices based on two random halves of the dataset. Because correlation is largely determined by extremes and the present set of varieties is well structured, DArT-HP markers with high PIC value yield higher correlation than multiallelic SSRs. DArT markers with low PIC values describe mostly finer structure within main populations and contribute little to the DR, thus leading to low DR values for the complete set of 702 DArTs.

4.2. Population structure

A common problem in population genetics studies is assigning an accession to one of *K* populations on the basis of its genotype and information about distribution of the alleles in the *K* populations [3]. The knowledge of population structure, genetic relationship among varieties and identification of genotypes is useful for germplasm development, variety protection, population genetics and geneflow studies, and gene mapping. In this study we analyzed varieties originating from five sugar beet seed companies. All varieties (with a notable exception of HH06 from Holly Hybrids) originating from the same seed company always grouped together when analyzed with the full set of 766 markers. Interestingly, varieties from different companies were consistently placed into the same inferred populations, indicating their genetic similarity. For example, all material from Betaseed and American Crystal Hybrid Seeds grouped together, while all varieties from Hilleshög and SeedEx formed a different population. Holly Hybrids varieties (except accession HH06) were distinctly separated from other tested material. Grouping of material from different seed companies together is possibly due to intertwined history of many breeding programs. For instance, both Hilleshög and SeedEx trace their origin to the Great Western Sugar Company, while American Crystal Hybrid Seeds became a part of Betaseed (Lewellen and McGrath, Personal communication).

Information about parental material that was used to develop hybrid varieties is proprietary thus it is not possible to compare pedigree information with the population structure observed in this study. However, the co-dominant nature of the SSR and SNP markers allowed us to test the hypothesis of a common parent. This analysis indicates that none of the breeding companies used only a single common parent for developing all hybrid varieties included in our study, and that no hybrid variety from either Holly Hybrids (ten varieties) or SeedEx (three varieties) shares a common parent with other varieties from the same company (Supplementary file 4). However, the hypothesis of a common parent was not rejected for some pairs of hybrid varieties originating from Betaseed, American Crystal Hybrid Seeds, and Hilleshög companies. We also detected varieties from different breeding companies (but placed in the same population by STRUCTURE and GenoDive programs) that show a possibility of having a parent in common. Taking into consideration the observed distribution of genotypes in the SNP and SSR datasets, we calculated for a pair of varieties the chance of getting at least one allele in common at all loci is 2.8×10^{-5}. This chance increases to the maximum of 1.6×10^{-4} for the 0.92 cut-off. Considering 1431 pairwise comparisons among 54 varieties, we theoretically expect detecting a common parent for only 0.041–0.236 pairs of varieties by a chance alone. If the calculation of a chance is based on the frequency of alleles in the SNP and SSR datasets and assumption of the Hardy–Weinberg equilibrium, the number of pairs increases to the 0.536–1.131 range, still substantially less then 37 pairs of varieties that were observed in our analysis of a common parent (Supplementary file 4). However, it needs to be pointed out that our analysis was based only on a single, randomly selected plant per hybrid variety. Moreover, this analysis does not provide the evidence of a common parent *per se*, but rather it identifies pairs of hybrid varieties for which a possibility of having a common (or very closely related) parent cannot be rejected.

Previously Smulders et al. [23] genotyped 40 sugar beet varieties originating from 13 seed companies with 25 SSRs. They observed that nine (out of 10) varieties from a single company formed a separate branch, but there was no clear structure in the genetic relatedness of other varieties. However, these clustering results were based on a relatively small sample of varieties (nine companies each were represented by only one or two varieties). Including more material may provide a better resolution [24] and separate varieties from individual companies.

4.3. Effect of dominant markers

Molecular markers compared in this study belong to both dominant and co-dominant marker-systems. Despite some amount of genotypic uncertainty, dominant alleles provide correct estimates of population structure when they are handled properly [13]. Since co-dominant SNP markers and dominant DArT-HP makers used in the present study have approximately similar PIC (0.41 and 0.47, respectively), the effect of dominance could be roughly estimated (assuming that this is the only difference between the two marker-systems). There were approximately 2.2× more DArT-HP markers needed for *genotyping* and *assignment*, 2.3× more for *structure*, and 2.4× more for *clustering*. These empirical data are in good agreement with modeling results that showed that under supervised clustering about 1.7× more dominant than co-dominant markers are needed to attain the same success-rate [3].

De Riek et al. [25] and later Smulders et al. [23] genotyped sugar beet cultivars with SSR markers and tested a dominant scoring of alleles in which marker bands were recorded as either present or absent. Results of analyses performed with dominantly and co-dominantly scored markers were similar, indicating that both scoring methods can be used to distinguish genotyped material [23].

4.4. Comparison of marker-systems

Our primary criterion for evaluating the success-rate of *assignment*, *structure*, and *clustering* is based on assignment of varieties into the same populations as was detected with a combined set of all 766 markers. We assumed that an infinite number of markers describes the population structure perfectly and that any very large set of markers yields the same results [12]. The progress of success-rate curves indicates that this assumption is correct and a large number of markers of any type infer the same grouping of varieties as the combined set of all markers. Results from our empirical analyses are in agreement with modeling simulations which showed that the error rate of clustering rapidly decreases with a growing number of marker-loci [19].

It was previously proposed that to obtain comparable assessments of population structure and genetic diversity, around 7–11 times more SNPs are required than SSRs [26]. In our analyses only about 1.4 (*genotyping*) to 3 (*clustering*) times more SNP markers were needed to achieve 100% success-rate (Table 1). This substantial difference between two studies may be caused by use of different plant species (maize vs. sugar beet) with a dissimilar population structure. Another very likely possibility is that SNPs used in our study undergone selection when tested on various sets of sugar beet material, whereas selection performed on the genomic SSRs and the EST-SSRs was much more limited.

Selecting only highly polymorphic loci leads to a substantial reduction in number of markers that are needed to reach 100% success-rate. For example, using only the most polymorphic DArT loci (DArT-HP sub-system) decreases the number of marker-loci that are needed for analyses by approximately 38% for *clustering*, 40% for *genotyping*, 50% for *assignment*, and 61% for *structure* (Table 1). At the same time, data resolution of DArT-HP markers was substantially higher than data resolution of DArTs. However, the complete set of DArT markers possibly describes a relationship of varieties within each population more accurately then the more limited DArT-HP subset.

In addition to performance (success-rate), other important factors for selecting suitable marker-system are quality and reproducibility of data. In our datasets only 3.1% SSRs, 0.4% DArTs, and 0.2% SNPs data were missing (data not shown). A lower percentage of missing data in SNP- than in SSR-based system was previously observed also in maize [27], though the percentages of missing data were higher then in our datasets (13.8% for SSRs and 2.1–3.1% for SNPs). Also, repeatability (repeated analysis of the same material) was better for SNPs (98.1–99.3%) than for SSRs (91.7%) [27]. We did not test repeatability of these two marker-systems, but DArTs used in our study have repeatability of 99.5% (data not shown), which is comparable with results from SNP-Invader assay [27].

4.5. Comparison of analyses

Regardless of the marker-system, substantially fewer markers were needed for *genotyping*. Of the other three remaining analyses, *assignment* needed the least number of markers, while *clustering* required the highest number of marker-loci. Although the differences in the overall AUC values among these three types of analyses are relatively small, they are significant at *p*-value of 0.05 (Table 2). Higher AUC values for *assignment* than for *clustering* or *structure* are not surprising, as this type of analysis assigns only a single accession into one of the known populations already containing all other varieties. On the other hand, *clustering* and *structure* assign all varieties into predefined (supervised clustering problem) number of populations or clusters. It is interesting to note that from the two approaches used for detection of population structure, equal number or fewer loci is needed for *structure* (carried out by STRUCTURE) than for *clustering* (carried by GenoDive). The difference was the most pronounced for DArT-HP markers, where approximately 36% less loci were needed to reach the same (100%) level of success-rate. This information is important for selecting an appropriate analytical approach when only a limited number of markers are available for detection of population structure. However, our observations are based only on a single set of hybrid varieties with a highly structured population. More analyses on populations with both similar and different structure are needed to confirm differences between STRUCTURE and GenoDive observed in our study.

4.6. Conclusions

We studied three types of marker-systems for their use in detecting genetic diversity and population structure in cultivated sugar beet and for assigning a variety into a population. The set of 54 varieties from five seed companies was genotyped with 766 markers (SSR, SNP, and DArT) and analyzed for population structure. Each of the three inferred populations contained varieties from only one or two seed companies (with a single exception). These populations were well separated from each other as indicated by consistent and identical results achieved both by STRUCTURE (Figs. 3 and 4) and GenoDive.

The number of marker-loci that were needed to reach 100% success-rate depended upon the marker-system; but in general, more of them were required for detection of population structure than for detection of genotypic diversity. From all analyses the fewest number of marker-loci were needed for *genotyping* with SSRs, where no further gain was observed past 17 marker-loci. On the contrary, the most marker-loci (400) were needed for detection of population structure (*clustering* and *structure*) with DArT markers (Table 1). Overall, per locus success-rate decreased from SSR, followed by SNP, DArT-HP, and DArT. These results correspond to the level of marker polymorphism and co-dominant/dominant nature of markers. Though more DArTs than SSRs or SNPs are needed to reach 100% success-rate, using only highly polymorphic DArTs substantially decreases the number of marker-loci that are needed for analyses and improves consistency of the dataset (higher DR value).

The results from this study provide a premise to selecting the marker-system(s) and a number of marker-loci needed for a particular analysis of sugar beet genetic diversity. However, in our comparisons we did not consider aspects related to marker development and use, such as frequency of markers in the genome, reliability and reproducibility of data, amenability of markers to high-throughput screening, or the economy of genotyping. All these factors together with success-rate need to be taken into consideration by individual laboratories when selecting the best type of molecular markers for genotyping (fingerprinting), detection of population structure, and assignment of varieties into populations.

Acknowledgements

The authors thank Amy Atallah for technical assistance with data analysis and manuscript editing; Patrick Meirmans and Jonathan Pritchard for advise relating to data analyses with GenoDive and STRUCTURE programs, respectively, and Robert Lewellen and Rob van Treuren for valuable discussion.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.plantsci.2011.12.009.

References

[1] I. Simko, K.G. Haynes, R.W. Jones, Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers, Genetics 173 (2006) 2237–2245.
[2] S. Mariette, V. Le Corre, F. Austerlitz, A. Kremer, Sampling within the genome for measuring within-population diversity: trade-offs between markers, Molecular Ecology 11 (2002) 1145–1156.
[3] G. Guillot, A. Carpentier-Skandalis, On the informativeness of dominant and co-dominant genetic markers for Bayesian supervised clustering, The Open Statistics & Probability Journal 3 (2011) 7–12.
[4] D. Tautz, M. Renz, Simple sequences are ubiquitous repetitive components of eukaryotic genomes, Nucleic Acids Research 12 (1984) 4127–4138.
[5] I. Simko, Development of EST-SSR markers for the study of population structure in lettuce (*Lactuca sativa* L.), Journal of Heredity 100 (2009) 256–262.
[6] P. Wenzl, J. Carling, D. Kudrna, D. Jaccoud, E. Huttner, A. Kleinhofs, A. Kilian, Diversity Arrays Technology (DArT) for whole-genome profiling of barley, Proceedings of the National Academy of Sciences of the United States of America 101 (2004) 9915–9920.
[7] P. Hurtado, K.M. Olsen, C. Buitrago, C. Ospina, J. Marin, M. Duque, C. De Vicente, P. Wongtiem, P. Wenzel, A. Killian, M. Adeleke, M. Fregene, Comparison of simple sequence repeat (SSR) and diversity array technology (DArT) markers for assessing genetic diversity in cassava (*Manihot esculenta* Crantz), Plant Genetic Resources: Characterisation and Utilisation 6 (2008) 208–214.
[8] V. Laurent, P. Devaux, T. Thiel, F. Viard, S. Mielordt, P. Touzet, M.C. Quillet, Comparative effectiveness of sugar beet microsatellite markers isolated from genomic libraries and GenBank ESTs to map the sugar beet genome, Theoretical and Applied Genetics 115 (2007) 793–805.
[9] K. Schneider, D. Kulosa, T.R. Soerensen, S. Möhring, M. Heine, G. Durstewitz, A. Polley, E. Weber, Jamsari, J. Lein, U. Hohmann, E. Tahiro, B. Weisshaar, B. Schulz, G. Koch, C. Jung, M. Ganal, Analysis of DNA polymorphisms in sugar beet (*Beta vulgaris* L.) and development of an SNP-based map of expressed genes, Theoretical and Applied Genetics 115 (2007) 601–615.
[10] D. Jaccoud, K. Peng, D. Feinstein, A. Kilian, Diversity arrays: a solid state technology for sequence information independent genotyping, Nucleic Acids Research 29 (2001) e25.
[11] D. Botstein, R.L. White, M. Skolnick, R.W. Davis, Construction of a genetic linkage map in man using restriction fragment length polymorphisms, American Journal of Human Genetics 32 (1980) 314–331.
[12] T.J.L. van Hintum, Data resolution: a jackknife procedure for determining the consistency of molecular marker datasets, Theoretical and Applied Genetics 115 (2007) 343–349.
[13] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: dominant markers and null alleles, Molecular Ecology Notes 7 (2007) 574–578.
[14] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[15] P.G. Meirmans, P.H. Van Tienderen, GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms, Molecular Ecology Notes 4 (2004) 792–794.
[16] G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Molecular Ecology 14 (2005) 2611–2620.
[17] R.B. Calinski, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics - Theory and Methods 3 (1974) 1–27.
[18] B. Guinand, A. Topchy, K.S. Page, M.K. Burnham-Curtis, W.F. Punch, K.T. Scribner, Comparisons of likelihood and machine learning methods of individual classification, Journal of Heredity 93 (2002) 260–269.
[19] G. Guillot, F. Santos, Using AFLP markers and the Geneland program for the inference of population genetic structure, Molecular Ecology Resources 10 (2010) 1082–1084.
[20] M.J. Hubisz, D. Falush, M. Stephens, J.K. Pritchard, Inferring weak population structure with the assistance of sample group information, Molecular Ecology Resources 9 (2009) 1322–1332.
[21] D. Paetkau, W. Calvert, I. Stirling, C. Strobeck, Microsatellite analysis of population structure in Canadian polar bears, Molecular Ecology 4 (1995) 347–354.
[22] P.M. Agapow, A. Burt, Indices of multilocus linkage disequilibrium, Molecular Ecology Notes 1 (2001) 101–102.

[23] M.J.M. Smulders, G.D. Esselink, I. Everaert, J. De Riek, B. Vosman, Characterisation of sugar beet (*Beta vulgaris* L. ssp. vulgaris) varieties using microsatellite markers, BMC Genetics 11 (2010) 41.

[24] N.A. Rosenberg, T. Burke, K. Elo, M.W. Feldman, P.J. Freidlin, M.A.M. Groenen, J. Hillel, A. Mäki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, S. Weigend, Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds, Genetics 159 (2001) 699–713.

[25] J. De Riek, I. Everaert, D. Esselink, E. Calsyn, M.J.M. Smulders, B. Vosman, Assignment tests for variety identification compared to genetic similarity-based methods using experimental datasets from different marker systems in sugar beet, Crop Science 47 (2007) 1964–1974.

[26] D. Van Inghelandt, A.E. Melchinger, C. Lebreton, B. Stich, Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers, Theoretical and Applied Genetics 120 (2010) 1289–1299.

[27] E.S. Jones, H. Sullivan, D. Bhattramakki, J.S.C. Smith, A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L.), Theoretical and Applied Genetics 115 (2007) 361–371.