Theor Appl Genet (2007) 115:361–371
DOI 10.1007/s00122-007-0570-9

ORIGINAL PAPER

A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L.)

E. S. Jones · H. Sullivan · D. Bhattramakki · J. S. C. Smith

Received: 4 December 2006 / Accepted: 25 April 2007 / Published online: 22 May 2007

Abstract We report on the comparative utilities of simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers for characterizing maize germplasm in terms of their informativeness, levels of missing data, repeatability and the ability to detect expected alleles in hybrids and DNA pools. Two different SNP chemistries were compared; single-base extension detected by Sequenom MassARRAY®, and invasive cleavage detected by Invader® chemistry with PCR. A total of 58 maize inbreds and four hybrids were genotyped with 80 SSR markers, 69 Invader SNP markers and 118 MassARRAY SNP markers, with 64 SNP loci being common to the two SNP marker chemistries. Average expected heterozygosity values were 0.62 for SSRs, 0.43 for SNPs (pre-selected for their high level of polymorphism) and 0.63 for the underlying sequence haplotypes. All individual SNP markers within the same set of sequences had an average expected heterozygosity value of 0.26. SNP marker data had more than a fourfold lower level of missing data (2.1–3.1%) compared with SSRs (13.8%). Data repeatability was higher for SNPs (98.1% for MassARRAY SNPs and 99.3% for Invader) than for SSRs (91.7%). Parental alleles were observed in hybrid genotypes in 97.0% of the cases for MassARRAY SNPs, 95.5% for Invader SNPs and 81.9% for SSRs. In pooled samples with mixtures of alleles, SSRs, MassARRAY SNPs and Invader SNPs were equally capable of detecting alleles at mid to high frequencies. However, at low frequencies, alleles were least likely to be detected using Invader SNP markers, and this technology had the highest level of missing data. Collectively, these results showed that SNP technologies can provide increased marker data quality and quantity compared with SSRs. The relative loss in polymorphism compared with SSRs can be compensated by increasing SNP numbers and by using SNP haplotypes. Determining the most appropriate SNP chemistry will be dependent upon matching the technical features of the method within the context of application, particularly in consideration of whether genotypic samples will be pooled or assayed individually.

Communicated by A. Graner.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-007-0570-9) contains supplementary material, which is available to authorized users.

E. S. Jones (✉) · H. Sullivan · D. Bhattramakki · J. S. C. Smith
Pioneer Hi-Bred International Inc. (DuPont Agriculture and Nutrition), 7300 NW 62nd Avenue, P.O. Box 1004, Johnston, IA 51031-1004, USA
e-mail: liz.jones@pioneer.com

Introduction

Recent advances in marker technologies have enabled high-throughput, low cost markers to routinely be used to characterize germplasm and to select for favorable alleles in plant breeding programs. The ideal marker system is highly polymorphic, codominant, accurate, reproducible, high-throughput and low cost (both in terms of capital investment and cost per assay). Simple sequence repeats (SSRs) are currently the marker of choice for most crops. However, operationally, there have been problems in their use caused by; challenges in accurately sizing SSR alleles due to PCR and electrophoresis artifacts (Hatcher et al. 1993; Jones et al. 1997; Bovo et al. 1998; Fernando et al. 2001; Heckenberger et al. 2002; Davison and Chilba 2003); PCR competition effects that can cause unequal allele

amplification, resulting in the inability to observe heterozygotes or multiple alleles within a pool; null alleles arising from mutations in the primer region flanking the SSR (Isibashi et al. 1996; Fernando et al. 2001; Batley et al. 2003a), and size homoplasy, whereby alleles of the same size may not necessarily be identical in sequence (Estoup et al. 1995).

The generation of sequence information for many crop species has paved the way for a new marker system (Rafalski 2002). Single nucleotide polymorphisms (SNPs) offer the promise of even higher levels of throughput compared with SSRs due to simpler and quicker processes that collectively facilitate automation and sample-tracking. SNPs are the most abundant class of sequence variability in the genome and thus have the potential to provide the highest map resolution (Bhattramakki et al. 2002). Compared with the genomes of other cultivated plant species, SNP frequency in maize has been found to be high (Vroh Bi et al. 2006) with one SNP being found every 28–124 bp (Tenaillon et al. 2001; Ching et al. 2002; Vroh Bi et al. 2006). This compares with a frequency in maize of one SSR approximately every 8 kb (Wang et al. 1994). SNP markers are usually biallelic and so are less polymorphic than SSRs on an individual marker basis. However, this limitation can be compensated for by their abundance and by the ability to utilize SNP haplotypes (Gupta et al. 2001; Ching et al. 2002; Rafalski 2002). A database and resource for SNP discovery and trait dissection has been established for maize in which genotype, phenotype and polymorphism data can be accessed for a wide range of maize inbreds and populations (Zhao et al. 2006: http://www.panzea.org).

Several studies have addressed the relative capabilities of isozymes, RFLPs, RAPDs, AFLPs and SSRs to characterize inbred lines and hybrids in maize (Smith et al. 1997; Dubreuil and Charcosset 1998; Pejic et al. 1998; Bernardo et al. 2000; Lübberstedt et al. 2000; Heckenberger et al. 2003; Garcia et al. 2004), but none has yet compared SSRs with SNPs. The objective of this study is to investigate the utility of SSR and SNP markers for characterizing maize germplasm with respect to data quality measured with individual inbred and hybrid samples, as well as DNA pools.

Materials and methods

Genotypes

A total of 58 maize inbreds were selected consisting of 52 diverse public inbreds and six inbreds proprietary to Pioneer Hi-Bred International Inc. (Table 1). The inbreds are categorized (numbers of inbreds per class in parentheses) according to pedigree background as stiff stalk (22), non-stiff stalk (21), flint (11), or miscellaneous (4). Four hybrids were studied; one public hybrid (B73 × Mo17) and three hybrids that are proprietary to Pioneer Hi-Bred International Inc. All parental inbreds of the hybrids were included in the set of inbreds.

DNA extraction

For each inbred and hybrid, plant material was lyophilized and eight 1 cm leaf disks placed in each well of a 96-well plate. DNA was extracted using a modified CTAB method (Saghai-Maroof et al. 1984). The same plate of DNA was used for both SSR and SNP studies.

Generation of SSR data

Eighty publicly available SSR markers were utilized (Supplementary Table 1). SSRs were selected on the basis of their ability to discriminate among maize germplasm, their low tendency to stutter and cause scoring problems, and that collectively they allowed polymorphisms on each chromosome arm to be assayed. Map distribution was; chr1 (6 markers), 2 (8), 3 (5), 4 (13), 5 (5), 6 (13), 7 (12), 8 (3), 9 (4), and 10 (11). Primer sequences for all SSR markers are available at http://www.maizegdb.org/. SSR data were collected as described by Berry et al. (2002) apart from for electrophoresis which was at 7.5 kV for 90 min using the Applied Biosystems ABI PRISM® 3700 with GENESCAN v. 3.7 and POP6 polymer.

SNP marker design and generation of SNP data

For the SNP discovery process, primer design, sequencing, SNP calls and sequence haplotype calls (a combination of SNPs within a single sequence) were carried out by Myriad Genetics, Inc, Salt Lake City. Loci were selected for sequencing from the public unigene set (Cone et al. 2002; Gardiner et al. 2004). SNP markers were designed using the consensus sequence assembled from sequence data for 61 public inbred genotypes. SNPs were selected for marker design when they met the following criteria; (1) a high level of polymorphism, and (2) amenability to marker design based on having 100 readable bases upstream and downstream from the target SNP, 25 bp around the target SNP with no polymorphism and a GC content of 40–60%. Individual and consensus sequence data are available at http://www.panzea.org. 123 SNP loci contained within 117 sequences were targeted for this study and these collectively mapped to each of the ten maize chromosomes; chr1 (24 markers), 2 (21), 3 (14), 4 (10), 5 (13), 6 (9), 7 (4), 8 (16), 9 (7), 10 (7) (Supplementary Table 2). Two different SNP detection technologies were investigated for marker validation; single base primer extension using the Sequenom

Table 1 Maize inbreds

Maize inbred	Public/Proprietary	Heterotic group
38–11	Public	SSS
A188	Public	MISC
A509	Public	NSS
A556	Public	NSS
A619	Public	NSS
A632	Public	SSS
B	Public	NSS
B14	Public	SSS
B37	Public	SSS
B42	Public	NSS
B64	Public	SSS
B73	Public	SSS
B84	Public	SSS
B89	Public	SSS
B94	Public	SSS
C103	Public	NSS
C106	Public	FLINT
CI66	Public	MISC
CM49	Public	FLINT
CO109	Public	FLINT
D02	Public	FLINT
F2	Public	FLINT
F252	Public	MISC
F257	Public	FLINT
F283	Public	FLINT
F7	Public	FLINT
H84	Public	SSS
H99	Public	NSS
HATO4	Public	FLINT
HY	Public	SSS
Indiana H60	Public	NSS
K187–11217	Public	SSS
L1546	Public	NSS
L317	Public	NSS
Minn49	Public	NSS
MO17	Public	NSS
MP305	Public	MISC
N28	Public	SSS
OH07	Public	NSS
OH40B	Public	NSS
OH43	Public	NSS
OH45	Public	NSS
OS420	Public	SSS
OS426	Public	SSS
PA91	Public	NSS
R159	Public	SSS
SD105	Public	SSS
SRS303	Public	NSS
TR9-I 461	Public	NSS
V3	Public	FLINT
W153R	Public	SSS
WF9	Public	SSS
PH_I1	Proprietary	NSS
PH_I2	Proprietary	SSS
PH_I3	Proprietary	FLINT
PH_I4	Proprietary	SSS
PH_I5	Proprietary	SSS
PH_I6	Proprietary	NSS

MassARRAY® system and invasive cleavage using Invader® with PCR and fluorescent resonance energy transfer (FRET). For the Sequenom MassARRAY system, SNP markers were designed and validated by Genaissance Pharmaceuticals (New Haven, CT, USA). For the Invader assays, SNP markers were designed and validated by Third Wave Molecular Diagnostics (Madison, WI, USA). A subset of 118 markers was tested using MassARRAY chemistry and an additional sub-set of 69 was tested with Invader chemistry, with 64 markers being common to the two chemistry types.

Experimental design

Lyophilized plant material for the 58 inbreds and four hybrids was arrayed in a randomized format within a 96 well plate. The eight inbred parents of hybrids were each represented twice within the plate. All other inbreds were represented once. Each of the four hybrids was represented three times. This plate format was then duplicated in a second plate, with the same leaf material source being used for both plates. The two plates were treated as two separate projects, with the DNA being extracted and genotyped at different times within the laboratory. However, many of the same reagent lots and solutions would have been used for both plates, so that results from the two plates cannot be considered as completely independent. Sample identities were hidden to avoid any potential bias in scoring. Allele calls for the same genotypes were compared across replicates to study repeatability. For SSRs, the allele data were also compared to historical allele data collected through multiple iterations of screens that would have included several seed sources for each genotype. For SNPs, the allele data were compared to sequence data, which used a different seed source for many of the inbreds studied here. To assess the accuracy of allele detection in hybrids, the hybrid allele calls were compared to known parental

alleles. Partial mis-matches were declared where: (case 1) inbred parents were monomorphic, the expected allele was observed in the hybrid, but one or more additional alleles were also observed in the hybrid, or (case 2) inbred parents were polymorphic, but only one of the parental alleles was only observed in the hybrid. For cases where the inbred parent was heterozygous at a locus, either allele was considered as contributing to the hybrid.

Allele dosage

Leaf disks of inbreds and their hybrids were mixed in varying proportions to provide allele dosages ranging from 1/16 to 16/16. For example, to provide a 1/16 dose of B73 with 15/16 of Mo17, one leaf disk of the B73/Mo17 hybrid SX19 was combined with seven leaf disks of Mo17. These dosage levels were constructed for each of the four hybrids. Each dose series for each 'cross' was repeated twice, and dosages were randomized across the plate.

Data analysis

Four categories of polymorphisms were compared for their ability to discriminate among the 58 maize inbreds: (1) sequence haplotypes for 117 sequences (2) all individual SNP loci (435 loci) within the 117 sequences (3) a sub-set of SNP loci (123/435 loci) selected for their high polymorphism levels and marker designability, and used to assess repeatability in this study, and (4) the SSR set (80 loci) used to assess repeatability in this study.

Several measures of marker informativeness were assessed. First, expected heterozygosity (also called polymorphism information content; PIC) was calculated as follows:

$$H = 1 - \sum (p_i^2),$$

where p_i is the frequency of the ith allele. Expected heterozygosity is a measure of the number and frequency of alleles in a population and hence is the probability that two individuals taken at random from the population considered will have different alleles at a locus (Nei and Li 1987). A high value indicates that there are many alleles at approximately equal frequency.

Second, minor allele frequency (the proportion of the lowest frequency allele for each SNP) was calculated for the selected SNP set and all SNPs within the same set of sequences. Expected heterozygosity values and minor allele frequencies are correlated, so that increasing expected heterozygosity will increase minor allele frequency. Our interest in this additional measure of discrimination power was to determine how many 'rare alleles' (here defined as minor allele frequency <0.2) might be lost by selecting a SNP set for high polymorphism levels.

Third, pair-wise genetic distances were calculated among the 58 inbreds based on the proportion of common alleles across all markers tested (Nei and Li 1979) using SAS (SAS Institute Inc.) for the SSRs, selected SNPs and sequence haplotypes. As genetic distance using these analyses is based on the principal that shared alleles are identical by descent, this measure of discrimination power is usually only meaningful when most members of the germplasm set being analyzed share pedigree relatedness. We measured pedigree relatedness using Coefficients of Parentage (CP) calculated following Malécot (1948), Gizlice et al. (1994) and Cui et al. (2000).

Allele and genotype entry effects were analyzed using generalized linear modeling with PROC GLM in SAS. Correlations were analyzed using Spearman rank correlations with PROC CORR.

Results

Marker informativeness

For SSRs, the average allele number was 5.1 (range 2–11); 2.5 times higher than the number of alleles for SNPs (Table 2). Average expected heterozygosity for SSRs was 0.62, which is 50% higher than that for the selected SNP set (0.43) (Table 2). For sequence haplotypes, the number of alleles and average, range and distribution of expected heterozygosity values were comparable to SSRs (Table 2; Fig. 1). When all SNPs within the sequences utilized were examined, the average expected heterozygosity was 0.26; considerably lower than that for the selected SNP set (Table 2). When minor allele frequencies were compared, 55% of all SNP loci within the sequences studied had

Table 2 Comparison of allele number and expected heterozygosity values for SSRs, sequence haplotypes and SNP loci for 58 diverse corn inbreds

	Number of loci	Number of alleles		Expected heterozygosity	
		Average	Range	Average	Range
SSR loci	80	5.1	2–11	0.62	0.33–0.85
Sequence haplotypes	117	5.1	2–11	0.63	0.29–0.85
All SNP loci in sequences	435	2	2	0.26	0.02–0.5
SNP marker loci selected for high polymorphism levels	123	2	2	0.43	0.12–0.50

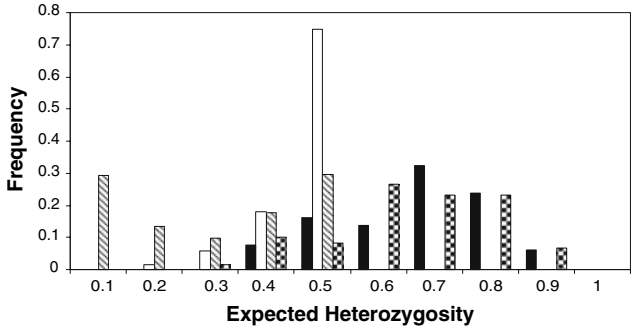

Fig. 1 Frequency distribution of expected heterozygosity values for 80 SSR loci (*black fill*), 117 sequence haplotypes (*checkered fill*), 435 SNPs within the 117 sequences (*diagonal lines*) and a sub-set of 123 SNPs selected for high levels of polymorphism (*white fill*)

minor allele frequencies <0.2, while only 10% of the SNP loci selected for high levels of polymorphism had minor allele frequencies in the same class (data not shown).

Coefficient of Parentage (CP) was used to assess the relatedness of the inbreds utilized in this study. The majority of inbreds were not related, at least according to known pedigree; 1,024/1,653 of the pair-wise inbred comparisons had CP values of 0, 593 had CP values >0 and ≤ 0.125, and only 36 had CP values >0.125.

Genetic distances between pairs of inbreds generated by SSRs, the selected SNP set and sequence haplotypes were compared. For all pair-wise inbred comparisons, correlations between the different marker types were not significant ($P < 0.05$; data not shown). Pairs of inbreds were then analyzed separately on the basis of their CP. For inbreds with CP values of zero or ≤ 0.125, correlations between the different marker types were also not significant (data not shown). For the 36 pair-wise comparisons with CP values >0.125, all correlations were significant at $P < 0.01$ with R^2 values 0.46–0.57 (statistical analyses data not shown; Fig. 2). Genetic distances using SSRs and sequence haplotypes were similar for this small inbred set, but genetic distances with SNPs were approximately three times lower. For example, inbreds with a genetic distance of 0.2 with SSRs had a genetic distance of 0.22 with sequence haplotypes, but only 0.07 with SNPs.

Missing data

The level of missing data was 4–5 times higher for SSRs than SNPs, with SSRs having an average level of 13.8% missing data, and SNPs having an average level of 2.1% missing data for SNP-MassARRAY and 3.1% for SNP-Invader (Table 3). The level of missing data for SNPs detected with either chemistries was significantly lower than SSRs ($P < 0.001$, data not shown). Sample genotype had a highly significant effect on the level of missing data for SSRs and for SNPs ($P < 0.001$, analyzed for each marker type separately, data not shown) demonstrating that the missing data effects were not due to random causes.

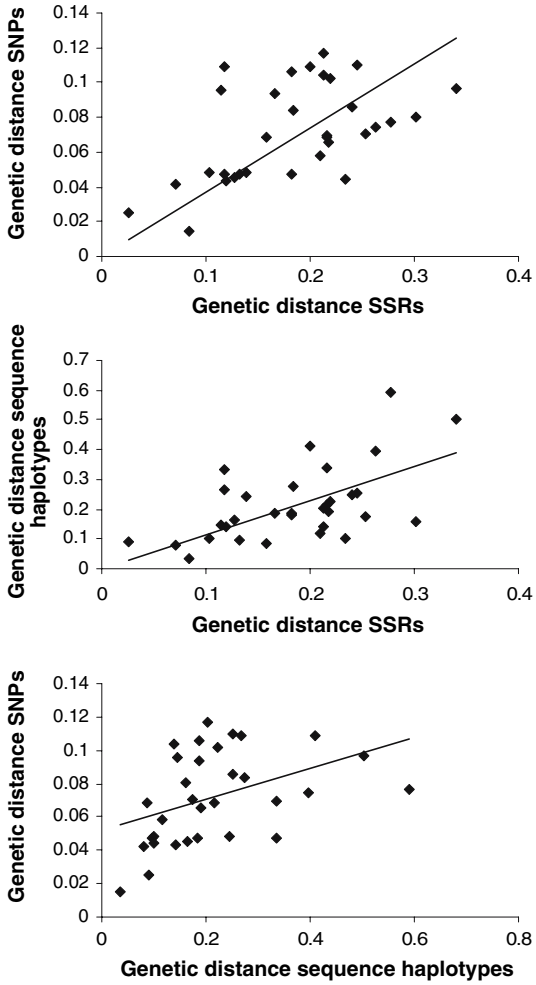

Fig. 2 Genetic distance correlations calculated between pairs of inbreds using 80 SSR loci, 123 SNP loci selected for their high levels of polymorphism and 117 sequence haplotypes. Only data for inbred pairs related by pedigree (CP > 0.125) are shown

Repeatability for inbreds

Repeatability of marker profiles generated from using SSR markers was assessed by comparing data collected for 69 samples across replicates 1 and 2 for each marker type, and then also comparing results from each project to historical data. Repeatability of SNP marker data was assessed by comparing data from replicate plates and also by reference to sequence data. SNPs had a significant ($P < 0.01$) and consistently higher level of repeatability than for SSRs (Table 4). Between replicates, alleles completely matched for 91.7% of the data for SSRs, 98.1% of the data for SNP-MassARRAY and 99.3% of the data for SNP-Invader. As the same leaf tissue was used for all of these projects, these values represent repeatability in the absence of any po-

Table 3 Missing data rates for SSRs and SNPs on maize inbreds and hybrids

		Records[a]	Missing data	Average % missing data ± standard deviation
SSRs	Replicate 1	5,520	652	11.8
	Replicate 2	5,520	868	15.7
	Average across replicates	5,520	760	13.8 ± 2.77
SNP-MassARRAY	Replicate 1	8,142	154	1.9
	Replicate 2	8,142	187	2.3
	Average across replicates	8,142	170.5	2.1 ± 0.28
SNP-Invader	Replicate 1	4,761	161	3.4
	Replicate 2	4,761	138	2.9
	Average across replicates	4,761	150	3.1 ± 0.34

[a] 69 samples, 80 markers for SSRs, 69 markers for SNP-MassARRAY and 118 markers for SNP–Invader

Table 4 Repeatability for SSRs and SNPs on inbred maize genotypes: Pair-wise comparisons of data from Project 1, Project 2, and historical allele data for SSRs, or sequence call data for SNPs

	Data comparison	Comparable records[a]	Match	Partial mismatch (1 allele in common)	Complete mismatch
SSRs	Project 1–Project 2	3,726	3418 (91.7%)	267 (7.2%)	41 (1.1%)
	Project 1–historical	4,124	3729 (90.4%)	278 (6.7%)	117 (2.8%)
	Project 2–historical	3,855	3504 (90.1%)	249 (6.5%)	102 (2.6%)
SNP-MassARRAY	Project 1–Project 2	6,936	6806 (98.1%)	123 (1.8%)	7 (0.1%)
	Project 1–sequence call	6,565	6343 (96.6%)	195 (3.0%)	27 (0.4%)
	Project 2–sequence call	6,552	6378 (97.3%)	146 (2.2%)	28 (0.4%)
SNP-Invader	Project 1–Project 2	4,085	4116 (99.3%)	29 (0.7%)	0 (0.0%)
	Project 1–sequence call	3,812	3756 (98.5%)	46 (1.2%)	10 (0.3%)
	Project 2–sequence call	3,814	3784 (99.2%)	18 (0.5%)	12 (0.3%)

[a] Allele calls that could be directly compared between data sets. Loss from the total number of records was due to missing data in either data set

tential seed source issues. When data were compared to historical data (for SSRs) or sequence data (for SNPs) the % match was approximately 1% lower (Table 4). These lower repeatabilities could be attributed to genotypic differences between seed sources, or for SSRs, changes in protocols and allele dictionaries over time. For SNPs, differences could also be due to differences in the accuracy of identifying SNP alleles with SNP detection platforms compared to allele calls obtained from sequence data. A significant correlation was found between % matched data per genotype across marker platforms (at $P < 0.001$ in all cases, data not shown), suggesting that certain genotypes had leaf sample or DNA quality issues that were affecting the accuracy of allele calls.

Ability to detect expected alleles in hybrids

Six replicates of each of the four hybrids were used to evaluate whether expected allele calls were generated compared to the parents (Table 5, Fig. 3). For loci that were not polymorphic between parents, the expected alleles were observed in 96.8% of cases for SSRs and 98.3% of cases for SNP data collected using either MassARRAY or Invader. However, where markers were polymorphic, SSRs revealed a much lower ability to detect parental alleles in the hybrids, with partial matches (only one of the parental alleles being detected) occurring in 26.4% of cases, compared with 4.6% for SNP-MassARRAY and 5.7% for SNP-Invader; a fivefold difference. Complete mismatches were low for SSRs at about 0.3%. This category of discrepancy cannot be measured for these SNP systems because only two specific alleles can be interrogated in each assay.

Ability to detect alleles in pooled samples

For each marker method, there was a rapid increase in the ability to detect the minor allele as allele dosage increased from 1/16 to 8/16. At the lower doses of 1/16 and 2/16 alleles, SSRs had allele detection rates that were significantly higher than SNP-MassARRAY, which in turn were significantly higher than for SNP-Invader (Table 6). At doses of 4/16–8/16 alleles, allele detection rates were highest for SNP-MassARRAY. At 8/16 alleles and above, detection rates leveled off to be at around 90% and each

Table 5 Evaluations of allele call accuracy in hybrids for SSRs and SNPs

Marker type	Polymorphism status of parents	Average % allele match to inbred parents ± S.D.[a]	Average % partial mismatch ± S.D.[a]	Average % complete mismatch ± S.D.[a]
SSRs	Monomorphic	96.8 ± 4.8	2.9 ± 4.6	0.3 ± 0.4
	Polymorphic	73.3 ± 1.6	26.4 ± 1.9	0.3 ± 0.4
	All markers	81.9 ± 1.4	17.8 ± 1.2	0.3 ± 0.3
SNP-MassARRAY	Monomorphic	98.3 ± 2.3	1.6 ± 2.3	–
	Polymorphic	95.4 ± 5.5	4.6 ± 5.5	–
	All markers	97.0 ± 3.8	3.0 ± 3.8	
SNP-Invader	Monomorphic	98.3 ± 1.6	1.8 ± 1.6	–
	Polymorphic	94.2 ± 6.3	5.8 ± 6.3	–
	All markers	95.5 ± 5.4	4.5 ± 5.4	–

[a] Average and S.D. (standard deviation) over four hybrids. There were six replicates per hybrid and data was collected for 80 SSR, 118 SNP-MassARRAY and 69 SNP-Invader markers

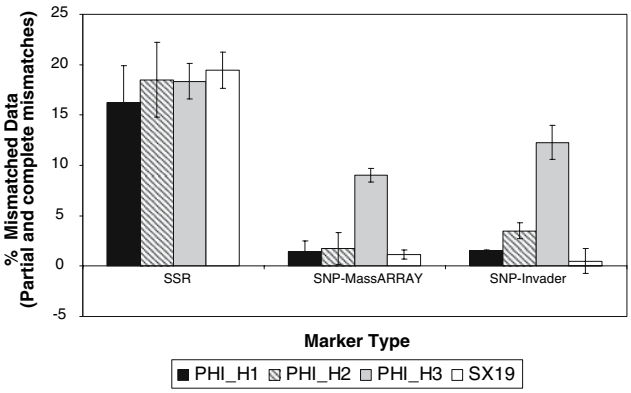

Fig. 3 Percent of hybrids with allele scores that did not match the expected allele scores for polymorphic and monomorphic SSR and SNP markers. Four hybrids were tested; three proprietary (PHI_H1, PHI_H2, PHI_H3) and one public (SX19)

marker method was roughly equivalent (Table 6, Fig. 4). Overall, across all samples, SNP-MassARRAY had significantly higher allele detection rates than for SSRs, which in turn had significantly higher allele detection rates than for SNP-Invader (Table 6).

In the allele dosage study, missing data rates for SNP-Invader were ten times higher than those found in the first experiment designed to determine missing data rates (Tables 3, 6; 34.8% in the allele dosage study compared with 3.1% in the first experiment) and 3.5 times higher than SSRs or SNP-MassARRAY. This result reflects scoring methodology. For SNP-Invader, data are grouped into clusters with the varying allele doses resulting in undefined clusters which are more likely to be scored as 'equivocal', or missing data.

Discussion

Marker informativeness

Marker informativeness was assessed using a number of different criteria. The number of alleles is the most basic criterion, where markers with a larger number of alleles are more likely to be polymorphic for any given germplasm set. Expected heterozygosity is a more accurate measure of polymorphism, as it further measures the distribution of those alleles across the germplasm being examined. Minor allele frequency is a measure often used to assess informativeness for SNP loci and is related to expected heterozygosity where the number of alleles is two, as is usually the case for SNPs. We also attempted to assess the relative informativeness of SSRs, SNPs and sequence haplotypes using genetic distance analysis which uses the proportion of shared alleles between pairs of inbreds across all markers within any marker set. Rather than giving a single value per marker based on a specific germplasm set (as for expected heterozygosity and minor allele frequency), genetic distance gives a single information value per marker type and inbred pair combination. Correlating genetic distance for inbred pairs for each marker type gives a relative assessment of distinguishing power. Such analysis is usually only meaningful when the germplasm being studied shows pedigree relatedness, so that common alleles are more likely to have identity by descent. For this reason we also assessed relatedness of the germplasm set using coefficient of parentage (CP).

The number of alleles per locus reported for SSRs ranged from 2 to 11 with an average of 5.1, and resulted in an average expected heterozygosity value of 0.62. Although the number of alleles and expected heterozygosity are dependent on the specific markers selected and the diversity of germplasm used, values were similar to those found in previous studies. The average number of SSR alleles reported has ranged from 4.4 to 6.8 (Pejic et al. 1998; Lu and Bernardo 2001; Heckenberger et al. 2002; Warburton et al. 2002) and expected heterozygosity values have ranged from 0.58 to 0.89 (Taramino and Tingey 1996; Smith et al. 1997; Pejic et al. 1998; Heckenberger et al. 2002; Garcia et al. 2004).

As individual SNPs generally only have two alleles, the maximum value for expected heterozygosity is 0.5. For all

Table 6 Allele dose–series experiments: % missing data and allele detection at different doses for polymorphic markers

	SSRs		SNP-MassARRAY		SNP-Invader	
Allele dosage	% Missing data	% Alleles detected*	% Missing data	% Alleles detected*	% Missing data	% Alleles detected*
1/16 (6.25%)	7.6 ± 1.9	32.1 ± 21.6[a]	15.7 ± 18.4	20.7 ± 11.6[b]	24.7 ± 14.9	7.0 ± 6.3[c]
2/16 (12.5%)	10.5 ± 3.5	51.6 ± 23.4[a]	15.1 ± 7.0	41.7 ± 21.4[b]	43.4 ± 17.6	14.8 ± 15.1[c]
4/16 (25%)	12.7 ± 6.1	68.3 ± 19.8[a]	11.1 ± 4.8	72.5 ± 22.7[a]	49.4 ± 9.2	52.8 ± 37.5[b]
6/16 (37.5%)	9.0 ± 5.2	79.6 ± 10.9[b]	6.5 ± 4.2	89.8 ± 8.4[a]	42.4 ± 9.8	82.3 ± 17.4[b]
8/16 (50%)	14.6 ± 11.5	85.6 ± 9.0[c]	4.3 ± 4.7	94.2 ± 4.5[a]	45.7 ± 9.3	89.0 ± 12.5[b]
10/16 (62.5%)	9.0 ± 5.2	88.3 ± 6.2[b]	6.5 ± 4.2	95.2 ± 4.0[a]	42.4 ± 9.8	94.2 ± 10.3[a]
12/16 (75%)	12.7 ± 6.1	89.8 ± 5.1[b]	11.1 ± 4.8	95.3 ± 4.9[a]	49.4 ± 9.2	95.2 ± 8.7[a]
14/16 (87.5%)	10.5 ± 3.5	91.1 ± 6.0[b]	15.1 ± 7.0	95.3 ± 4.4[a]	43.4 ± 17.6	94.3 ± 11.5[a]
15/16 (93.75%)	7.6 ± 1.9	93.3 ± 4.9[a]	15.7 ± 18.4	95.3 ± 4.3[a]	24.7 ± 14.9	95.9 ± 6.0[a]
16/16 (100%)	12.7 ± 6.4	97.8 ± 2.2[a]	8.4 ± 17.6	94.5 ± 4.3[a]	8.6 ± 6.2	97.0 ± 4.7[a]
Overall	10.9 ± 5.9	71.0 ± 31.0[b]	10.7 ± 11.7	73.0 ± 33.6[a]	34.8 ± 18.6	66.1 ± 40.0[c]

*Mean values that have the same letter within in a row (allele dose series) are not significantly different

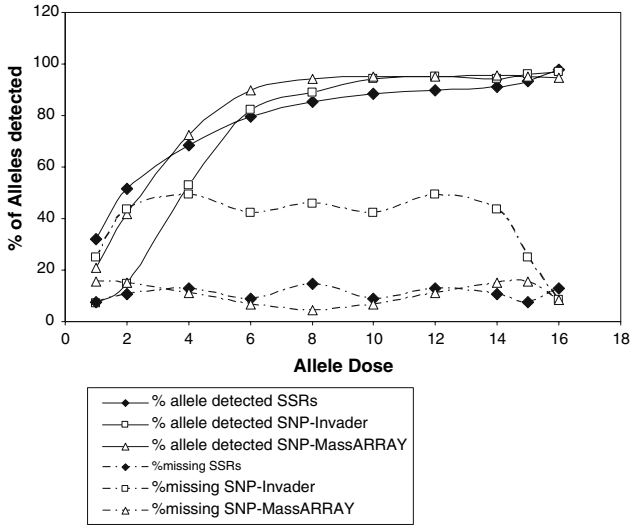

Fig. 4 The ability to detect alleles present at different doses in a pooled sample using SSR and SNP markers. Averages were taken for allele dose series for four different hybrids each with two replicate samples

SNPs within the sequences examined we found an average expected heterozygosity value of 0.26, which is identical to that found by Ching et al. (2002) in US maize inbreds. By selecting SNPs based on their high polymorphism levels, we increased the discriminating power of the SNP set to an average expected heterozygosity of 0.43. However this selection of SNPs also caused the frequency of the more rare SNPs (defined here as minor allele frequency <0.20) to be reduced from 55 to 10%. These more rare SNPs may be useful in germplasm outside of the set studied here i.e. an ascertainment bias may have been introduced by this selection.

By considering sequence haplotypes, Ching et al. (2002) found an average expected heterozygosity value of 0.56; a twofold increase over the value for individual SNPs. The sequence haplotypes in this study had slightly higher average expected heterozygosity values of 0.63; a greater than twofold increase over all SNPs within the sequences examined, and a 50% increase over the selected SNP set.

We found that SNP haplotypes exhibited allele numbers and expected heterozygosity values that compared favorably with SSRs, as has previously been observed (Ching et al. 2002; Vroh Bi et al. 2006). However, in practice, this increased resolving power may not always be possible to exploit using SNP markers, as the marker design criteria often lead to the exclusion of many of the SNPs that constitute the haplotype. This feature is particularly problematic in maize where the occurrence of SNPs and indels is frequent (Tenaillon et al. 2001; Bhattramakki et al. 2002; Ching et al. 2002; Batley et al. 2003b; Vroh Bi et al. 2006). Batley et al. (2003a) tested the suitability of SnuPE assay for SNP detection in maize and found that approximately half of the SNPs were unsuitable for marker design due to SNPs flanking the target SNP or insufficient space for primer design. A more successful approach might be to utilize haplotypes defined by a series of closely linked SNPs that are amenable to marker design and that are in linkage disequilibrium.

To assess relative informativeness using genetic distance analysis, we first assessed the relatedness of the germplasm being used using coefficient of parentage (CP). Most of the pair-wise inbred comparisons had CP values of zero, with only a small sub-set being related by known pedigree, defined here as CP > 0.125. Across all inbred comparisons, or where inbred pairs had CP values of 0 or <0.125, there was no significant correlation between the marker systems. Where CP values were >0.125, correlations between the different marker systems were significant. For these inbred pairs, genetic distances using SSRs

and SNP haplotypes were similar, while equivalent numbers of SNPs resulted in genetic distances approximately three times lower. However, only 123 SNP loci were used in comparison to 80 SSRs: a more appropriate comparison of technologies would be to test two to fourfold higher numbers of SNPs compared with SSRs.

Missing data, repeatability and Mendelian inheritance in hybrids

The level of missing data for SNPs was substantially lower than that for SSRs (2.1 and 3.1% compared with 13.8%), and repeatability for SNPs was substantially higher (98.1 and 99.3% compared with 91.7%). It has generally been observed that SNPs are more reliable that SSRs (Gupta et al. 2001). Our results confirm this observation with greater clarity and specificity using direct comparisons of data generated using the same DNA samples and with a large number of markers assayed. George et al. (2004) found a similar level of repeatability for maize SSRs tested across laboratories (>92%). Giancola et al. (2006) found similar levels of missing data (<3.3%) and repeatability (>96.8%) for three SNPs tested with TaqMan and Amplifor SNP chemistries in *Arabidopsis thaliana*. In human SNP studies, Lahermo et al. (2006) found accuracy to be >99% for 9 different SNP chemistries tested across multiple laboratories and Pati et al. (2004) found repeatability levels >98.4% with SNaPshot, Pyrosequencing and Biplex Invader SNPs chemistries. However, Giancola et al. (2006) also tested a third SNP chemistry (the GOOD assay detected using mass spectrometry) and found repeatability and missing data levels to be substantially lower. Therefore, not all SNP methods will produce high quality data and testing and optimization is required on a crop-specific basis. In this study we utilized validated markers and could not specifically compare marker design success rates, but these also can vary with different SNP technologies.

Few studies have examined the ability of marker systems to accurately detect Mendelian inheritance of alleles in hybrids. Smith et al. (1997) compared SSRs and RFLPs in maize and found SSRs to be superior, with 2.2% SSRs, compared with 3.6% RFLPs, segregating in a non-Mendelian fashion. In this study we obtained data that showed non-Mendelian inheritance in hybrids at a higher level of 18.1 % for SSR data but only 3–4.5% for SNP data. Thus these SNP methods are more reliable than SSRs for genotyping maize hybrids.

Error rates for SSRs may appear to be higher than for SNPs due to their inherently higher information content, which then results in sampling or cross-contamination errors being more readily detectable. We did not find any evidence for sampling or cross-contamination errors here, so that repeatability differences are most likely to be due to errors in processing and scoring SSRs themselves.

Ability to detect alleles at different frequencies in pooled samples

The ability to detect alleles at varying frequencies has applications in studying populations of individuals for germplasm analysis and maintenance, heterogeneous varieties bred in self-incompatible species, pre-screening populations for polymorphism, or extrapolating F_2 genotypes from F_3 families. We found that SNPs detected using the Invader chemistry did not perform as well as SNPs detected with MassARRAY or SSRs. Invader scoring utilizes clustering, where data are not scored as independent data points but rather within the context of all other data. However, it is possible that the scoring method for SNP-Invader could be custom-modified to use relative raw florescence values so that minor alleles could be more readily detected. For example, modifications have been made to mass spectrometry SNP chemistries to quantify alleles via peak sizes in humans (e.g. Werner et al. 2002). In plants, SNPs detected with Pyrosequencing have been used to quantify alleles in the complex polyloid and aneuploid crop sugarcane (Cordeiro et al. 2006).

Comparison of MassARRAY and Invader SNP chemistries

There are many SNP chemistries and detection methods available (see Gupta et al. 2001 and Jander et al. 2002 for reviews). Here, just two widely used chemistries were compared; MassARRAY and Invader. We found that pooled samples provided the greatest challenge to the generation of accurate data. Therefore, for any pooled sample experiments, the particular method of SNP detection should be a prime consideration and should be carefully evaluated.

MassARRAY and Invader are equivalent in their processing requirements that affect throughput and sample processing errors. Both technologies can be considered as being ‘medium-plexing technologies’. Such technologies are flexible in the number of markers that can be generated per sample at a low cost; a single plex can be run to generate a small number of markers per sample e.g. for marker-assisted selection, and multiple plexes can be run to generate larger marker numbers e.g. for genome-wide fingerprinting. Invader offers some advantages over MassARRAY in that plexing is at the PCR level, while the assay itself, which is the majority of the cost, is carried out in monoplex. Therefore, with Invader, only informative markers need to be assayed in any individual project, whereas for MassARRAY all markers within the fixed plex are assayed.

Overall, both MassARRAY and Invader technologies performed exceedingly well in terms of low levels of missing data, high repeatability and the ability to detect heterozygotes in hybrids. High levels of data quality make both methodologies attractive options for genotyping to support many breeding applications.

We have not attempted detailed cost comparisons of SSRs and the SNP technologies described here, largely because SNP assay costs continue to decrease and are also highly dependant on the sample volumes being assayed. We estimate SNP costs to be <$0.25 per marker/sample data point and five to ten times lower than for SSRs. The increased quality and reduced missing data for SNP data that we have demonstrated here further increases the value of SNP markers. The loss of resolution of SNPs compared with SSRs can be compensated by increasing SNP numbers (although the number needed to give equivalent discriminating power compared to SSRs was not resolved here), or by utilizing SNP haplotypes.

An additional significant advantage to SNP marker systems is that the nucleotide itself is interrogated, so that results across time and labs can be readily compared without the need for extensive checks and allele-binning synchronization that is required for comparative sizing of SSR alleles. Similarly, SNPs collected with different SNP chemistries and platforms can be readily compared, a practice not currently feasible with SSRs, where any differences in platform or protocol can effect allele designation. SNPs therefore not only offer lower cost, higher volume, repeatability and accuracy, they also offer increased opportunities to develop germplasm databases that are meaningful across different organizations and that, in contrast to the succession of marker technologies that has occurred over the past two decades, will not be redundant as further improved SNP technologies become available.

References

Batley J, Mogg R, Edwards D, O'Sullivan H, Edwards KJ (2003a) A high-throughput SNuPE assay for genotyping SNPs in flanking regions of *Zea mays* sequence tagged simple sequence repeats. Mol Breed 11:111–120

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003b) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91

Bernardo R, Romero-Severson J, Ziegle J, Hauser J, Joe L, Hookstra G, Doerge RW (2000) Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. Theor Appl Genet 100:552–556

Berry DA, Seltzer JD, Xie C, Wright DL, Smith JSC (2002) Assessing probability of ancestry using simple sequence repeat profiles: applications to maize hybrids and inbreds. Genetics 161:813–824

Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC, Tingey SV, Rafalski A (2002). Insertion–deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol Biol 48:539–547

Bovo D, Rugge M, Shiao YH (1998) Origin of spurious multiple bands in the amplification of microsatellite sequences. Mol Pathol 52:50–51

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Cone K, McMullen M, Vroh Bi I, Davis G, Yim YS, Gardiner J, Polacco M, Sanchez-Villeda H, Fang Z, Schroeder S, Havermann SA, Bowers JE, Paterson AH, Soderland CA, Engler FW, Wing RA, Coe EH (2002) Genetic, physical and informatic resources for maize: on the road to an integrated map. Plant Physiol 130:1598–1605

Cordeiro G, Eliott F, McIntyre CL, Casu RE, Henry RJ (2006) Characterization of single nucleotide polymorphisms in sugarcane ESTs. Theor Appl Genet 113:331–343

Cui Z, Carter TE Jr, Burton JW (2000) Genetic diversity patterns in Chinese soybean cultivars based on coefficient of parentage. Crop Sci 40:1780–1793

Davison A, Chilba S (2003) Laboratory temperature variation is a previously unrecognized source of genotyping error during capillary electrophoresis. Mol Ecol Notes 3:321–323

Dubreuil P, Charcosset A (1998) Genetic diversity within and among maize populations: a comparison between isozyme and nuclear RFLP loci. Theor Appl Genet 96:577–587

Estoup A, Tailliez C, Cornuet JM, Solignac M (1995) Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). Mol Biol Evol 12:1074–1084

Fernando P, Evans BJ, Morales JC, Melnick DJ (2001) Electrophoresis artifacts—a previously unrecognized cause of error in microsatellite analysis. Mol Ecol Notes 1:235–328

Garcia AAF, Benchimol LL, Barbosa AMM, Geraldi IO, Souza CL Jr, de Souza AP (2004) Comparison of RAPD, RFLP, AFLP and SSR markers for diversity studies in tropical maize inbred lines. Genet Mol Biol 27:579–588

Gardiner J, Schroeder SS, Polacco ML, Sanchez-Villeda H, Fang Z, Morgante M, Landewe T, Fengler K, Useche F, Hanafey M, Tingey S, Cou H, Wing R, Soderlund C, Coe EH Jr (2004) Anchoring 9371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. Plant Physiol 134:1317–1326

George MLC, Regalado E, Li W, Cao M, Dahlan M, Pabendon M, Warburton ML, Xianchun X, Hoisington D (2004) Molecular characterization of Asian maize inbred lines by multiple laboratories. Theor Appl Genet 109:80–91

Giancola S, McKhann HI, Berard A, Camilleri C, Durand S, Libeau P, Roux F, Rebound X, Gut IG, Brunel D (2006) Utilization of three high-throughput SNP genotyping methods, the GOOD assay, Amplifluor and Taqman, in diploid and polyploidy plants. Theor Appl Genet 112:1115–1124

Gizlice JAA, Carter TE, Burton J (1994) Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci 34:1143–1151

Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535

Hatcher SL, Lambert QT, Raymond LT, Carlson JR (1993) Heteroduplex formation: a potential source of errors from PCR products. Prenat Diagn 13:171–177

Heckenberger M, Bohn M, Ziegle JS, Joe LK, Hauser JD, Hutton M, Melchinger AE (2002) Variation of DNA fingerprints among

accessions within maize inbred lines and implications for identification of essentially derived varieties: I genetic and technical sources of variation in SSR data. Mol Breed 10:181–191

Heckenberger M, van der Voort JR, Melcinger AE, Peleman J, Bohn M (2003) Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties: II genetic and technical sources of variation in AFLP data and comparison with SSR data. Mol Breed 12:97–106

Isibashi Y, Saitoh T, Abe S, Yoshida MC (1996) Null microsatellite alleles due to nucleotide sequence variation in the grey-sided vole. Mol Ecol 5:589–590

Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL (2002) *Arabidopsis* map-based cloning in the post-genome era. Plant Physiol 129:440–450

Jones CJ, Edwards KJ, Castaglione S, Winfield MO, Sala F, van de Wiel C, Bredemeijer G, Vosman B, Matthes M, Daly A, Brettschneider R, Bettini P, Buiatti M, Maestri E, Malcevschi A, Marmiroli N, Aert R, Volckaert G, Rueda J, Linacero R, Vasquez A, Karp A (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol Breed 3:381–390

Lahermo P, Liljedahl U, Alnaes G, Axelsson T, Borrkes A, Ellonen P, Groop P, Halldén C, Holmberg D, Holmberg K, Keinanän M, Kepp K, Kere J, Kiviluoma P, Kristensen V, Lindgren C, Odeberg J, Osterman P, Parkkonen M, Saarela J, Sterner M, Strömqvist L, Talas U, Wessman M, Palotie A, Syvänen A (2006) A quality assessment survey of SNP genotyping laboratories. Hum Mutat 27:711–714

Lu H, Bernardo R (2001) Molecular marker diversity among current and historical maize inbreds. Theor Appl Genet 103:613–617

Lübberstedt T, Melchinger AE, Duble C, Vuylsteke M, Kuiper M (2000) Relationships among early European maize inbreds: IV genetic diversity revealed with AFLP markers and comparison with RFLP, RAPD, and pedigree data. Crop Sci 40:783–791

Malécot G (1948) Les mathématiques de l'hérédité. Masson and Cie, Paris. [English translation. The mathematics of heredity (1969). W.H. Freeman and Co., San Francisco]

Nei M, Li W (1979) Mathematical model for studying genetic variance in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Nei M, Li W (1987) Molecular evolutionary genetics. Columbia University Press, New York, pp 106–107

Pati N, Schwinsky V, Kokanovic O, Magnuson V, Ghosh S (2004) A comparison between SNaPshot, pyrosequencing and biplex invader SNP genotyping methods: accuracy, cost and throughput. J Biochem Biophys Methods 60:1–12

Pejic I, Ajmone-Marsan P, Morgante M, Kosumplick V, Castiglioni P, Taramino G, Motto M (1998) Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSR, and AFLPs. Theor Appl Genet 97:1248–1255

Rafalski JA (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci 162:329–333

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms I barley: Mendelian inheritance, chromosomal locations, and population dynamics. Proc Natl Acad Sci USA 81:8014–8018

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

Taramino G, Tingey S (1996) Simple sequence repeats for germplasm analysis and mapping in maize. Genome 39:277–287

Tenaillon MI, Sawkins MC, Long AD, aut RL, Doebley JF, gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L). PNAS 98:9161–9166

Vroh Bi I, McMullen MD, Villeda HS, Schroeder S, Gardiner J, Polacco M, Soderlund C, Wing R, Fang Z, Coe EH (2006) Single nucleotide polymorphisms and insertion-deletions for genetic markers and anchoring the maize fingerprint contig physical map. Crop Sci 46:12–21

Wang Z, Weber JL, Zhong G, Tanksley SD (1994) Survey of plant short tandem DNA repeats. Theor Appl Genet 88:1–6

Warburton ML, Xianchun X, Crossa J, Franco J, Melchinger AE, Frisch M, Bohn M, Hoisington D (2002) Genetic characterization of CIMMYT inbred maize lines and open pollinated populations using large scale fingerprinting methods. Crop Sci 42:1832–1840

Werner M, Sych M, Herbon N, Illig T, Konig IR, Wjst M (2002) Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry. Hum Mut 20:57–64

Zhao W, Canaran P, Jurkuta R, Fulton T, Glaubitz J, Buckler E, Doebley J, Gaut B, Goodman M, Holland J, Kresovich S, McMullan M, Stein L, Ware D (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res 34:752–757